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

Ficha catalográfica elaborada pela BMHS/FGV

Lins, Caio
Nonparametric Instrumental Variable Regression Through Kernel Methods and Stochastic Gradients/ Caio F. Lins Peixoto. – 2023.
17f.

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 Kernel Methods and Stochastic Gradients

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

---

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

# 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 $(\Omega, \mathcal{F}, \mathbb{P})$. By "random variable" we mean scalar or vector valued measurable functions with $\Omega$ 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 $d$-dimensional vector of explanatory variables, $Y$ is the scalar response, $\varepsilon$ 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 $\varepsilon$ 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):

Provide one example for each?

**Omitted Variables** This means $\varepsilon$ 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 $h^\star(X) + \mathbb{E}[g^\star(W) \mid X]$.

**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 $\varepsilon$ some measure of the difference between $X_k$ and $X_k'$. Depending on how these two variables are related, we may introduce endogeneity.

**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 $\varepsilon$ given $Z$ is almost surely null, i.e., $\mathbb{E}[\varepsilon \mid Z] = 0$.

Causal graph, discuss strength of $\mathbb{E}[\varepsilon \mid Z] = 0$, discuss $Z$ only influences $Y$ through $X$.

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). The examples ahead will clarify how instrumental variables may be chosen in practice.

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 $\varepsilon$. 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.

# 3 Conclusion

# References

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