DISCLAIMER:

This document does not meet the current format guidelines of the Graduate School at The University of Texas at Austin.

It has been published for informational use only.

The Dissertation Committee for Qian Feng
certifies that this is the approved version of the following dissertation:

Essays on Causal Inference with Endogeneity and Missing Data

Committee:

Stephen G. Donald, Supervisor

Jason Abrevaya

Haiqing Xu

Carlos M. Carvalho

Essays on Causal Inference with Endogeneity and Missing Data

by

Qian Feng, B.Eco.; M.S. Econ.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2017

Dedicated to my parents Liang Zhou and Zhenqi Feng, and to my husband Ryan Gu.

Acknowledgments

The first person I would like to thank is my supervisor, Stephen G. Donald, for his continuous support and guidance. Whenever I encounter difficulties and obstacles in doing research, he is always able to point out the right direction and offer extremely helpful advice. I can't forget the numerous fruitful and inspiring conversations I had with him. I wish I could grow into a successful scholar as him in the future and am so proud of being his student!

I would like to express my gratitude to Jason Abrevaya and Haiqing Xu. I not only learned great knowledge and receive help from them, but am also largely influenced by their working attitude and enthusiasm about research.

I want to thank everyone in our Econometrics Research Group at department of Economics. I thank Sukjin Han, Brendan Kline, Jessie Coe, Xinchen Gu, Sungwon Lee, Peter Toth and Kaixi Wang. I benefit a lot from participating in the study group, writing seminars as well as lunch seminars and this is exactly where I was shaped into an econometrician.

I am grateful to my parents and my husband, who have provided me through moral and emotional support in my life. I can't continue the journey towards a PhD without any of them. Special thanks to my husband Ryan Gu, who accompanies me getting over ups and downs and who is always ready to offer the unconditional love.

Essays on Causal Inference with Endogeneity and Missing Data

Publication No. ____________________

Qian Feng, Ph.D.
The University of Texas at Austin, 2017

Supervisor: Stephen G. Donald

This dissertation strives to devise novel yet easy-to-implement estimation and inference procedures for economists to solve complicated real world problems. It provides by far the most optimal solutions in situations when sample selection is entangled with missing data problems and when treatment effects are heterogenous but instruments only have limited variations.

In the first chapter, we investigate the problem of missing instruments and create the generated instrument approach to address it. Specifically, When the missingness of instruments is endogenous, dropping observations can cause biased estimation. This chapter proposes a methodology which uses all the data to do instrumental variables (IV) estimation. The methodology provides consistent estimation with endogenous missingness of instruments. It firstly forms a generated instrument for every observation in the data sample that: a) for observations without instruments, the new instrument is an imputation;

b) for observations with instruments, the new instrument is an inverse propensity score weighted combination of the original instrument and an imputation. The estimation then proceeds by using the generated instruments. Asymptotic theorems are established. The new estimator attains the semiparametric efficiency bound. It is also less biased compared to existing procedures in the simulations. As an illustrative example, we use the NLSYM data set in which IQ scores are partially missing, and demonstrate that by adopting the new methodology the return to education is larger and more precisely estimated compared to standard complete case methods.

In the second chapter, we provide Lasso-type of procedures for reduced form regression with many missing instruments. The methodology takes two steps. In the first step, we generate a rich instrument set from the many missing instruments and other observed data. In the second step, IV estimation is conduced based on the generated instrument set. Specifically, the (very) many generated instruments are used to approximate a "pseudo" optimal instrument in the reduced form regression. The approach has been shown to have efficiency gains compared to the generated instrument estimator developed in the first chapter. We also compare the finite sample behavior of the new estimator with other Lasso estimator and demonstrate the good performance of the proposed estimator in the Monte Carlo experiments.

The third chapter estimates individual treatment effects in a triangular model with binary–valued endogenous treatments. This chapter is based on the previous joint work with Quang Vuong and Haiqing Xu. Following the identification strategy established in (Vuong and Xu, forthcoming), we propose a two-stage estimation approach. First, we estimate the counterfactual outcome and hence the individual treatment effect (ITE) for every observational unit in the sample. Second, we estimate the density of individual treatment effects in the population. Our estimation method does not suffer from the ill-posed inverse problem associated with inverting a non–linear functional. Asymptotic properties of the proposed method are established. We study its finite sample properties in Monte Carlo experiments. We also illustrate our approach with an empirical application assessing the effects of 401(k) retirement programs on personal savings. Our results show that there exists a small but statistically significant proportion of individuals who experience negative effects, although the majority of ITEs is positive.

Table of Contents

List of Tables

List of Figures

Chapter 1

Instrumental Variable Estimation with Missing Instruments

1.1 Introduction

Missing instruments occur in instrumental variables(IV) estimation when an instrumental variable has potentially missing values and is only available to a subsample of observations. For example, in Acemoglu and Robinson (2001), the mortality rate faced by early European settlers is used as an instrument for a country's institutions, but the mortality rate is missing for about 56% of the sample[1]. However, the importance of missing instruments for empirical work has not been fully appreciated. Econometric literature has offered only a few solutions for addressing the problem. The most common procedure is simply dropping the observations with missing instruments in the IV estimation. When the missingness of instruments depends on the endogenous variable and/or other observed variables, existing solutions including dropping observations can result in biased and imprecise estimation.

In this chapter, I propose a methodology to deal with missing instruments. The methodology can provide consistent and less biased IV estimation

[1] 72 out of 163 countries with colonial origins have (estimated) mortality rates.

even when the missingness of instruments is endogenous. The main idea is to generate a new instrument, which is available to every observation in the data sample, to replace the original one with missing issues. I present a three-step estimation procedure. In the first step, a generated instrument is formed for every observation in the data sample. The generated instrument is an imputation for individuals with missing instruments. The imputation is a predicted value for the missing instrument using completely observed variables. The generated instrument has a weighted combination form for individuals with observed instruments. The weight is the inverse propensity score, which is the probability of instrument being missing for the individual, after controlling for other completely observed variables. The combination is between the observed instrument and a predicted value of the instrument. The second and third steps are analogous to a standard two-stage least square (2SLS) procedure. The second step is a reduced form regression of the endogenous variable on the generated instrument and other exogenous variables. The third step is a structural estimation of the dependent variable on the fitted value of the endogenous variable and other exogenous variables.

Under certain regularity conditions, I am able to show that the new estimators is $\sqrt{n}$-consistent and asymptotically normal. My estimator belongs to the class of semiparametric doubly robust estimators (SDREs)[2] in terms

[2] I follow the terminology of Rothe and Firpo (2013) and say that a semiparametric estimator based on doubly robust moment condition is SDRE if the moment condition depends on two unknown nuisance functions, but still identifies the parameter of interest if either one of these functions is replaced by some arbitrary value.

of that the generated IV is a Doubly Robust IV (DRIV). Even if one of the two nuisance parameters is misspecified, the DRIV remains valid. Another common feature shared by SDREs is that the required convergence rate for the nonparametric component is slower than $n^{-1/4}$. I present explicit rate results for sieve estimation of the nuisance functions and argue that compared to other two-step sieve estimation ((Chen, 2007), (Chen et al., 2008)), SDRE can have a richer choice of sieve spaces. In terms of efficiency, I first calculate semiparametric efficiency bounds under the model restrictions. Then I compare the asymptotic variances of my estimator to the efficiency bounds and prove that it attains the bounds under some conditions.

In the application, I revisit one empirical example of missing IQ scores from Card (1995). I apply the new methodology for evaluating the return to education in which IQ score is an instrument for ability, which is proxied by the "Knowledge of the World of Work" (KWW) test score. the data set is the young men cohort from the National Longitudinal Survey 1976 (NLSYM76), in which IQ scores are missing about 30% of the sample. Card (1995) simply omits the observations with missing data in the IV estimation. I adjust the original IQ scores to generated IQ scores. Results show that by using generated IQ scores, return to education is increased from 8.9% to 13.9%, while the standard error reduces by 13%.

1.1.1 Related Literature

The methodology proposed in this chapter integrates ideas from the doubly robust (DR) estimation literature and the generated regressors literature. My estimators are based on efficient moment conditions, which have the doubly robust property. This approach has precedent in literature on general missing data problems (e.g., (Robins et al., 1994), (Rotnitzky and Robins, 1995), (Scharfstein et al., 1999), (Van der Laan and Robins, 2003), (Bang and Robins, 2005), (Wooldridge, 2007), (Graham et al., 2012)). Recent studies focus on semiparametric versions of DR estimation. Rothe and Firpo (2013) provides theoretical results for estimates derived from DR moment conditions. Belloni et al. (forthcoming) considers a semiparametric DR estimation for treatment effects. They estimate the first stage nuisance parameters using high dimensional machine learning methods. This chapter contributes to this line of research by developing efficient estimation procedure only through adjusting the variable with missing values itself. In particular, I consider different treatments to observations with and without missing data, while maintaining the DR property of the estimator.

The new methodology is also related to semiparametric estimation using generated regressors. Newey (2009) proposes a two-step series estimation of sample selection models, where the generated regressors from the first step are used to approximate the correction term. Theoretical results that characterize the influence of the generation step on the final estimator appear in e.g. Escanciano et al. (2011), Mammen et al. (2012) and Hahn and Ridder (2013).

This chapter applies the methodology and techniques about generated regressors to the situation of generated instruments. I consider both parametric and nonparametric estimation in the second step using generated instruments formed in the first step.

This chapter differs from existing literature on missing instruments in two major aspects. First, I consider the endogenous missingness of instruments. Dahl and DellaVigna (2009) use a dummy variable approach to deal with missing value for an instrument. They enter a zero for the missing value and "compensates" by using dummies for "missingness". Abrevaya and Donald (forthcoming) adds the interaction of dummy for missingness of other exogenous variables to the instrument set of dummy variable approach and considers an efficient GMM estimator. However, these two approaches implicitly assume away the endogeneity issue of missing. They use a moment condition in which the error term is mean-independent of the instrument, within the subsample where the instrument is non-missing. Angrist et al. (2010) propose a full-sample instrument using a linear projection of the partially observed instrument on the covariates in the sub-sample with non-missing instrument. Mogstad and Wiswall (2012) consider another full-sample instrument using instead a nonlinear projection, which is a nonparametric approximation to a conditional expectation. Neither of these two papers allows the dependence of missingness on the dependent variable, or the endogenous variable.

Second, the starting point of the methodology in this chapter is the efficient influence function under missing instruments. This chapter is then the first to investigate efficient procedure to deal with missing instruments issue. Chaudhuri and Guilkey (2013) illustrate the good finite sample performance of an augmented inverse propensity score weighted estimator when there are two missing instruments in the simulations. The chapter provides theoretical foundations for estimators based on efficient influence function and proposes an easy-to-implement procedure for execution.

1.1.2 Examples of Missing Instruments

Here I list several empirical examples in which instrumental variables have missing values, besides the aforementioned missing mortality rates ((Acemoglu and Robinson, 2001)) and missing IQ scores ((Card, 1995)). The first example is the return to education. Some researchers use family background variables, e.g. parental education, as instrument for education ((Heckman and Li, 2004), (Flabbi et al., 2008) ,(Wang, 2013)). However, parental education is only available for individuals whose parents are present in the same household[3]. The IVs are missing for the rest of the population whose parents are living apart.

The second example is the quantity/quality model. The missing instruments issue occurs in Angrist et al. (2010) when they combine different instrument sets across partially-overlapping parity-specific subsamples.

The last example is Mendelian Randomization[4]. Recent studies in

[3] Due to typical survey designs, information on parental characteristics is asked only when they are present in the same household.

[4] See Burgess et al. (2015) for a comprehensive survey on the methodology of Mendelian

health economics and epidemiology examine the causal effect of a risk factor (e.g. obesity, early-childhood depression) on educational attainment. In Smith and Hemani (2014), von Hinke Kessler Scholder et al. (2011), Kang et al. (2016), they use genetic variation as instruments. In one of the datasets they use, the Wisconsin Longitudinal Study (WLS), only 47% of the original sample have complete genetic data.

This chapter is organized as follows. Section 1.2 presents the IV model with missing instruments, the observational equivalence of the model with moment equations. Section 1.3 specifies the three-step procedure as well as the Gen-IV estimator. Section 1.4 establishes asymptotic results of the estimator. It also states the semiparametric efficiency bounds and discusses efficiency of the Gen-IV estimator. Section 1.5 provides simulation evidence of finite sample behavior of the estimators. Section 1.6 details the empirical application of return to education. Section 1.7 concludes the paper.

1.2 Model, Identification and the Generated Instrument

Consider the following standard linear regression model

$$Y_i = X_i\alpha + V_i^{'}\beta + \epsilon_i, \qquad \mathbb{E}(V_i\epsilon_i) = 0, \quad \mathbb{E}(X_i\epsilon_i) \neq 0, \qquad i = 1, ..., n$$

Where X_i is an endogenous regressor and V_i is a d_v-vector of exogenous regressor. The first element of V_i is 1. There exists an instrument Z_i which can

Randomization.

be (partially) missing. Z_i satisfies

$$\mathbb{E}(Z_i \epsilon_i) = 0 \tag{1.1}$$

Let D_i denote the missing indicator for Z_i,

$$D_i = \begin{cases} 1, & \text{if } Z_i \text{ is missing} \\ 0, & \text{otherwise} \end{cases}$$

For notational convenience, I In the following, I make distinction among three cases of i.i.d. data samples. The *full data* sample consists of $(Y_i, X_i, V_i, Z_i, D_i)_{i=1}^n$, which is the data sample we would want to collect on all the individuals $i = 1, ..., n$. The *observed data* sample is $(Y_i, X_i, V_i, (1 - D_i)Z_i, D_i)_{i=1}^n$ which is the actually observed data sample. The *complete data* sample consists of $((1 - D_i)Y_i, (1 - D_i)X_i, (1 - D_i)V_i, (1 - D_i)Z_i)_{i=1}^n$, which is the subsample of data where the instrument Z_i is observed for every individual[5].

Let $W_i \equiv (Y_i, X_i, V_i^{'})'$, i.e. containing all the completely observed variables. The missing indicator D_i satisfies the "missing at random" (MAR) assumption,

Assumption 1 (MAR)**.** $D_i \perp Z_i | W_i$

In the example of missing IQ scores, MAR indicates that the missingness of IQ scores doesn't depend on the level of IQ scores, after conditioning on completely observed variables like wage, education and gender. In the missing

[5] One can instead write the complete data sample as $(Y_i, X_i, V_i, Z_i, D_i)_{i=1}^{n_c}$. n_c is the number of individuals with complete data. This writing won't change the estimation procedure proposed in this paper.

mortality rates example, MAR implies the propensity scores of missing mortality rates should be close for similar countries. Such kind of conditional independence assumption is also extensively used in econometrics and statistics to achieve identification with missing data. Examples include inference in models with attrition or nonresponse (e.g. (Little and Rubin, 2014), (Robins and Rotnitzky, 1995), (Rotnitzky and Robins, 1995), (Wooldridge, 2002),(Wooldridge, 2007)), the estimation of treatment effects (e.g. (Heckman and Vytlacil, 2007) and the references therein), the recovery of comparability over time of statistics calculated using data collected with different methodology.

Remark 1. MAR allows the dependence of missing indicator D_i on completely observed variable W_i. Consider the following example,

$$D_i = \mathbb{1}(\varrho_0 + \varrho_1 Y_i + \varrho_2 X_i + \varrho_3 V_i \leq u_i)$$

ϱ_0, ϱ_1, ϱ_2 and ϱ_3 are scaler constants, and u_i is an error distributed by the standard logistic distribution. D_i depends on Y_i, X_i and V_i as long as ϱ_1, ϱ_2, $\varrho_3 \neq 0$. But D_i satisfies MAR. I call the missingness endogenous if $\varrho_1^2 + \varrho_2^2 + \varrho_3^2 \neq 0$.

The propensity score of Z_i being missing is defined as the conditional probability of Z_i on the completely observed variable W_i,

$$p(W_i) \equiv \mathbb{P}(D_i = 1 | W_i)$$

The propensity score $p(W_i)$ is assumed to have overlap,

Assumption 2 (Overlap)**.** *$0 \leq p(W_i) < 1 - \nu$, for some $\nu > 0$.*

This assumption effectively guarantees that, for any given value $w \in \mathcal{W}$, where $\mathcal{W} \subset \mathbb{R}^{d_v+2}$ is the support of W_i, there is positive probability that the instrument is observed. For large enough sample size n, there will be enough individuals with instruments near any point w for local methods to work.

Given Assumption 1 and Assumption 2, the following lemma establishes an observational equivalence result between the IV model with single missing instrument and moment conditions. It is an extension to the equivalence result for a general missing data problem in Graham (2011).

Lemma 1 (Identification)**.** *Let $\widetilde{Z}_i \equiv (Z_i, V_i^{'})'$ be the instrument set. The single missing instrument problem under Assumption 1 and 2 is observationally equivalent to the following moment restrictions.*

$$\mathbb{E}\left(\frac{1-D_i}{1-p(W_i)}\widetilde{Z}_i\epsilon_i\right) = 0 \tag{1.2}$$

$$\mathbb{E}\left(\frac{p(W_i)-D_i}{1-p(W_i)}|W_i\right) = 0 \tag{1.3}$$

I follow terminology in Graham (2011) to call (1.2) the "identifying moment" and (1.3) the "auxiliary moment". The estimation method of inverse propensity score weighting (IPW) merely utilizes the identifying moment and regards the moment condition (1.3) as auxiliary since it only helps in estimating the nuisance parameter $p(W_i)$. See for example Hirano et al. (2003) for reference. Doubly robust methods, however, adopt a certain combination of (1.2) and (1.3) for estimation. It is well known that a conditional moment restriction is equivalent to infinite number of unconditional moment restrictions[6]. Thus conditional moment (1.3) is equivalent to unconditional moment

$$\mathbb{E}\left(\frac{p(W_i)-D_i}{1-p(W_i)}g(W_i)\right)=0 \tag{1.4}$$

for any measurable function $g(\cdot)\in L_2(\mathcal{W})$.

In the context of single missing instrument, I choose $g(W_i)=\mathbb{E}(\widetilde{Z}_i|W_i)\epsilon_i$, where $\epsilon_i=\epsilon_i(W_i)=Y_i-X_i^{*'}\theta$, and consider the doubly robust, inverse propensity score weighted combination of identifying moment (1.2) and unconditional moment (1.4). Specifically, the moment condition I am going to use for estimation is

$$\begin{aligned}\mathbb{E}\left(\frac{1-D_i}{1-p(W_i)}\widetilde{Z}_i\epsilon_i-\frac{p(W_i)-D_i}{1-p(W_i)}\mathbb{E}(\widetilde{Z}_i|W_i)\epsilon_i\right) &= 0\\ \Longrightarrow \mathbb{E}\begin{pmatrix}\frac{1-D_i}{1-p(W_i)}Z_i-\frac{p(W_i)-D_i}{1-p(W_i)}\mathbb{E}(Z_i|W_i)\\ V_i\end{pmatrix}\epsilon_i &= 0\end{aligned}$$

Let $\mathcal{Z}_i\equiv\frac{1-D_i}{1-p(W_i)}Z_i-\frac{p(W_i)-D_i}{1-p(W_i)}\mathbb{E}(Z_i|W_i)$. I call $\mathcal{Z}_i$ the generated instrument. As a result, the actual instrument set used in estimation is $\widetilde{\mathcal{Z}}_i=(\mathcal{Z}_i,V_i')'$. The following lemma shows that the generated instrument $\mathcal{Z}_i$ is a valid IV.

Lemma 2 (Validity of Generated IV)**.** *The generated instrument $\mathcal{Z}_i$ is a valid, full data instrument in terms of that it satisfies the excluded restriction.*

$$\mathbb{E}(\mathcal{Z}_i\epsilon_i)=0 \tag{1.5}$$

where $\mathcal{Z}_i=\frac{1-D_i}{1-p(W_i)}Z_i-\frac{p(W_i)-D_i}{1-p(W_i)}\mathbb{E}(Z_i|W_i)$.

[6] Several examples pointing out this equivalence include Bierens (1982), Chamberlain (1987), and Donald et al. (2003).

Note that the generated IV $\mathcal{Z}_i$ contains two nuisance parameters $p(W_i)$ and $\mathbb{E}(Z_i|W_i)$. The following corollary summarizes the Doubly Robust (DR) property[7] of $\mathcal{Z}_i$.

Corollary 1 (Doubly Robust IV)**.** *The generated IV $\mathcal{Z}_i$ remains valid, i.e. $\mathbb{E}(\mathcal{Z}_i\epsilon_i) = 0$ if either $p(W_i)$ or $\mathbb{E}(Z_i|W_i)$ is misspecified.*

1.3 Estimation and the Gen-IV Estimator

Assume that the true value of the parameters of interest $\theta_0 = (\alpha_0, \beta_0)'$ lies in the interior of the compact parameter space $\Theta \in \mathbb{R}^{d_v+1}$. The estimation is based on the sample analog of moment condition:

$$\mathbb{E}(\widetilde{\mathcal{Z}}_i\epsilon_i) = 0 \tag{1.6}$$

where $\widetilde{\mathcal{Z}}_i$ is the generated instrument set. I propose a three-step semiparametric procedure for the estimation of θ based on (1.6). In the first step, the generated instrument $\mathcal{Z}_i$ is estimated nonparametrically, denoted as $\widehat{\mathcal{Z}}_i$. The second step is a reduced form regression of the single endogenous variable X_i on the generated instrument set $\widehat{\widetilde{\mathcal{Z}}}_i = (\widehat{\mathcal{Z}}_i, V_i')'$, which includes the estimated instrument $\widehat{\mathcal{Z}}_i$ and other exogenous variables V_i. The third step is a structural estimation of the dependent variable Y_i on exogenous variable V_i and the fitted value $\widehat{X}_i$ obtained in the second step. The following presents the estimation procedure in detail.

[7] A formal definition of DR is given in (Bang and Robins, 2005): An estimator is DR if it remains consistent when either (but not necessarily both) a model for the missingness mechanism or a model for the distribution of the complete data is correctly specified.

Step 1 Estimation of generated instrument

The generated instrument $\mathcal{Z}_i = \frac{1-D_i}{1-p(W_i)} Z_i - \frac{p(W_i)-D_i}{1-p(W_i)} \mathbb{E}(Z_i|W_i)$ contains two nuisance parameters which are the propensity score $p(W_i)$ and the conditional expectation $\mathbb{E}(Z_i|W_i)$. Let $h(W_i) \equiv \mathbb{E}(Z_i|W_i)$, hereafter. We first estimate $p(W_i)$ and $h(W_i)$ nonparametrically and then form the generated instrument by plugging in the nuisance estimates. In the following, I use subscript c and m to refer to observations belonging to the *complete data* sample and subsample with missing instruments,respectively. Individuals in the the *complete data* sample are indexed by $i = 1, ..., n_c$, while those in the missing instrument sample are indexed by $j = 1, ..., n_m$. It holds $n = n_c + n_m$.

Assumption 1 implies that the *full data* instrument Z_i is mean independent of the missing indicator D_i, controlling for completely observed variable W_i,

$$h(W_i) = \mathbb{E}(Z_i|W_i) = \mathbb{E}(Z_i|W_i, D_i = 0)$$

Hence we can use the *complete data* sample for the estimation of $h(\cdot)$. Let $\{q_l(w), l = 1, 2, ...\}$ be a sequence of known sieve basis functions, such as power series, splines, Fourier series, etc. Let $\mathcal{H}$ denote the sieve space spanned by $q_l(w)$

$$\mathcal{H} = \left\{ h(w) = q(w)^{'}\pi = \sum_{i=1}^{\infty} q_l(w)\pi_l \right\}$$

A sieve least square estimator for $h(w)$ is

$$\widehat{h}(w) = \sum_{i=1}^{n_c} Z_{ci} q^{k_h(n)}(W_{ci})(Q_h^{'} Q_h)^{-1} q^{k_h(n)}(w)$$

where

$$q^{k_h(n)}(w) = (q_1(w), ..., q_{k_h(n)}(w))'$$

and

$$Q_h = (q^{k_h(n)}(W_{c1}), ..., q^{k_h(n)}(W_{cn_c}))'$$

for some integer $k_h(n)$, with $k_h(n) \to \infty$ and $k_h(n)/n \to 0$ when $n \to \infty$.

A sieve least square estimator for the propensity score $p(w)$, proposed in Hahn (1998) is:

$$\widehat{p}(\cdot) = \arg \min_{p(\cdot) \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} (D_i - p(W_i))^2/2$$

$$\mathcal{S}_n = \left\{ s(w) = q^{k_p(n)}(w)'\pi = \sum_{j=1}^{k_p(n)} q_j(w)\pi_j \right\} \text{ for some known basis } (q_j)_{j=1}^{\infty}$$

where $\mathcal{S}_n$ is the sieve space spanned by the basis functions q_j, $j = 1, ..., k_p(n)$. The regularity conditions derived in this paper are based on the sieve estimation of the two nuisance functions. However, my estimation methodology doesn't restrict the choice for estimation techniques to sieve only. One can use other nonparametric estimation, like kernel methods in obtaining $\widehat{p}(\cdot)$ and $\widehat{h}(\cdot)$. In particular, machine learning methods, such as Lasso and random forests, are also suitable choices for estimating the two functions[8].

Furthermore, one can use parametric estimation for $\widehat{p}(\cdot)$ and $\widehat{h}(\cdot)$ as well, which is potentially more appealing in empirical studies. $h(W_i)$ can

[8] A valuable insight provided in Belloni et al. (forthcoming) is that, doubly robust moment conditions are key ingredients in deriving honest inference when machine learning methods are adopted for nuisance estimation

simply be a fitted value in a linear regression of instrument Z_i on W_i in the *complete data* sample:

$$\widehat{h}(W_i) = W_i'\widehat{\xi}$$

and $p(W_i)$ can be parametrized as

$$p(W_i) = p(W_i; \phi)$$

for a finite dimensional parameter ϕ. Estimates of ϕ can be obtained by maximizing

$$\Pi_{i=1}^{n}\{p(W_i;\phi)\}^{D_i}\{1 - p(W_i;\phi)\}^{1-D_i}$$

and $\widehat{p}(W_i) = p(W_i; \widehat{\phi})$. Limited dependent variable estimation like Probit and Logit can execute the estimation well.

As a result, the generated instrument $\mathcal{Z}_i$ is constructed by plugging in $\widehat{h}(W_i)$ and $\widehat{p}(W_i)$,

$$\widehat{\mathcal{Z}}_i = \frac{1 - D_i}{1 - \widehat{p}(W_i)} Z_i + \frac{D_i - \widehat{p}(W_i)}{1 - \widehat{p}(W_i)} \widehat{h}(W_i) \tag{1.7}$$

Note that in the subsample with missing instruments, the generated instrument is an impuation

$$\widehat{\mathcal{Z}}_{mi} = \widehat{h}(W_{mi})$$

On the other hand, in the *complete data* sample, the generated instrument is an inverse propensity score weighted combination of the original instrument Z_i and the estimated conditional expectation $\widehat{h}(W_i)$,

$$\widehat{\mathcal{Z}}_{ci} = \frac{1}{1 - \widehat{p}(W_{ci})} Z_{ci} - \frac{\widehat{p}(W_{ci})}{1 - \widehat{p}(W_{ci})} \widehat{h}(W_{ci})$$

Step 2 Reduced form estimation

The reduced form estimator $\widehat{\tau}$ is an OLS estimator,

$$\widehat{\tau} = \arg\min_{\tau \in B} \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \widehat{\widehat{\mathcal{Z}}}_i^{'} \tau \right)^2 = \left(\sum_{i=1}^{n} \widehat{\widehat{\mathcal{Z}}}_i \widehat{\widehat{\mathcal{Z}}}_i^{'} \right)^{-1} \sum_{i=1}^{n} \widehat{\widehat{\mathcal{Z}}}_i X_i$$

where $B \subset \mathbb{R}^{d_v+1}$ is a compact set, and $\widehat{\widehat{\mathcal{Z}}}_i = (\widehat{\mathcal{Z}}_i, V_i^{'})^{'}$.

Step 3 Structural estimation

I define the Gen-IV estimator $\widehat{\theta}_{GenIV}$ as follows

$$\begin{aligned} \widehat{\theta}_{GenIV} = (\widehat{\alpha}, \widehat{\beta})^{'} &= \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{X}_i \alpha - V_i^{'} \beta \right)^2 \\ &= \left(\sum_{i=1}^{n} \widehat{X}_i^{*} X_i^{*'} \right)^{-1} \sum_{i=1}^{n} \widehat{X}_i^{*} Y_i \\ &= \left(\sum_{i=1}^{n} \widehat{\widehat{\mathcal{Z}}}_i X_i^{*'} \right)^{-1} \sum_{i=1}^{n} \widehat{\widehat{\mathcal{Z}}}_i Y_i \end{aligned}$$

where $\widehat{X}_i = \widehat{\widehat{\mathcal{Z}}}_i^{'} \widehat{\tau}$ and $\widehat{X}_i^{*} = (\widehat{X}_i, V_i^{'})^{'}$.

I leave the large sample properties of $\widehat{\theta}_{GenIV}$ to Section 4.

1.4 Asymptotic Results

In this section, I present large sample properties of Gen-IV estimator.

1.4.1 Large Sample Properties of Gen-IV Estimator

For Gen-IV estimator, the asymptotic results established in this section is a generalization of the single missing instrument case. Namely, suppose now there is a fixed set of instruments $A_i = A(Z_i) \equiv (A_1(Z_i), A_2(Z_i), ..., A_t(Z_i))'$ satisfying unconditional independence assumption $\mathbb{E}(A_i\epsilon_i) = 0$. The number of instruments t doesn't increase with the sample size n. And t is relatively small compared to the sample size, $t << n$.

I also allow the "structural" equation to have a general separable nonlinear form $Y_i = g(X_i^*; \theta) + \epsilon_i$. $g(\cdot)$ is a known function, which can be nonlinear in X_i and/or V_i. In the single missing instrument case considered in Section 2, $g(X_i^*; \theta) = X_i^{*'}\theta$.

With fixed instrument set A_i, Step 2 and Step 3 in Section 2.2 are now altered by a standard GMM procedure[9] with estimated instrument set $\widehat{\mathcal{Z}}_i \equiv (\widehat{\mathcal{Z}}_{i1}, ..., \widehat{\mathcal{Z}}_{it})'$ and weighting matrix G. The GMM version of Gen-IV estimator with fixed instrument set A_i is

$$\widehat{\theta}_{GenIV} = \left(X^{*'}\widehat{\widehat{\mathcal{Z}}}\widehat{G}\widehat{\widehat{\mathcal{Z}}}' X^*\right)^{-1}\left(X^{*'}\widehat{\widehat{\mathcal{Z}}}\widehat{G}\widehat{\widehat{\mathcal{Z}}}' Y\right)$$

in matrix notation, where $\hat{G}$ is a consistent estimate of G, and $\widehat{\widehat{\mathcal{Z}}}_i = (\widehat{\mathcal{Z}}_i, V_i')'$.

[9] In the exact-identification case as in Section 2, the proposed three-step procedure coincides with a GMM procedure using generated instrument. The coincidence is analogous to that between a standard 2SLS and GMM.

1.4.1.1 Notation

Before listing regularity conditions, I first introduce some notations. Denote $\mathcal{W} \equiv \mathcal{X} \times \mathcal{Y} \times \mathcal{V} = \mathbb{R}^{d_v+2} = \mathbb{R}^{d_w}$ to be the support of the completely observed variables $W_i = (X_i, Y_i, V_i)$, and $\mathcal{W}$ is allowed to be unbounded. For any $1 \times d_w$ vector $\mathbf{a} = (a_1, ..., a_{d_w})$ of nonnegative integers, write $|\mathbf{a}| = \sum_{i=1}^{d_w} a_i$. Denote the $|\mathbf{a}|$th derivative of a function $l : \mathcal{W} \to \mathbb{R}$ as

$$\nabla^a l(w) = \frac{\partial^{|a|}}{\partial w_1^{a_1} \cdots \partial w_{d_w}^{a_{d_w}}} l(w)$$

The Hölder space $\Lambda^\gamma(\mathcal{W})$ is a space of functions with up to $[\gamma]$ continuous derivatives[10], and the highest (γth) derivatives are Hölder continuous with the Hölder exponent $\gamma - [\gamma] \in [0, 1)$. The Hölder space is endowed with the norm

$$||l||_{\Lambda^\gamma} = \sup_w |l(w)| + \max_{|a|=[\gamma]} \sup_{w \neq \bar{w}} \frac{|\nabla^{|a|} l(w) - \nabla^{|a|} l(\bar{w})|}{\sqrt{(w-\bar{w})'(w-\bar{w})}^{\gamma-[\gamma]}}$$

A Hölder ball $\Lambda_c^\gamma(\mathcal{W})$ with radius c is defined as

$$\Lambda_c^\gamma(\mathcal{W}) = \{l \in \Lambda^\gamma(\mathcal{W}) : ||l||_{\Lambda^\gamma} \leq c < \infty\}$$

Define a weighted sup-norm $||l||_{\infty\eta} \equiv \sup_{w \in \mathcal{W}} |l(w)(1 + ||w||^2)^{-\frac{\eta}{2}}|$ for some $\eta > 0$. The role $[1 + |w|^2]^{-\eta \cdot /2}$ plays is similar to that of a trimming procedure in Kernel methods, where smaller weight is imposed upon larger values of W. Denote $\Pi_{\infty n} l$ to be the projection of l onto the sieve space $\mathcal{S}_n$ under the norm $|| \cdot ||_{\infty\eta}$. A weighted Hölder ball $\Lambda_c^\gamma(\mathcal{W}, \eta)$ with radius c is then $\Lambda_c^\gamma(\mathcal{W}, \eta) \equiv \{l \in \Lambda^\gamma(\mathcal{W}) : ||l||_{\infty\eta} \leq c < \infty\}$.

[10] $[\cdot]$ is the largest integer less or equal than γ.

The two nuisance functions $p(\cdot)$ and $h(\cdot)$ belong to Hölder spaces $\Lambda^{\gamma_p}(\mathcal{W})$ and $\Lambda^{\gamma_h}(\mathcal{W})$. The weighted sup-norms for the two spaces are $||\cdot||_{\infty\eta_p}$ and $||\cdot||_{\infty\eta_h}$, respectively. To avoid tedious notation, I just let $\gamma = \gamma_p = \gamma_h$, and $\eta = \eta_p = \eta_h$.

1.4.1.2 Consistency

The regularity conditions in the following assumption are required for the consistency of Gen-IV estimator.

Assumption 3. *Let* $\widehat{G} - G = o_p(1)$ *for a positive semidefinite matrix* G, *and the Jacobian* $\mathfrak{J}_\theta \equiv -\frac{\partial}{\partial\theta}\mathbb{E}[\widetilde{Z}_i g(X_i^*;\theta)]$, *the following hold*

(3.1) $\mathfrak{J}_{\theta_0}$ *has full column rank equal to* $d_v + 1$.

(3.2) $p_0(\cdot)$ *belongs to Hölder ball* $\mathcal{S} = \{p(\cdot) \in \Lambda_c^\gamma(\mathcal{W}, \eta) : 0 < \underline{p} \le p(w) \le \bar{p} < 1, \forall w \in \mathcal{W}\}$, $h_0(\cdot)$ *is* $H(\gamma, \eta_1)$*-smooth for some* $\eta_1 \ge 0$.[11]

(3.3) $\mathbb{E}((1 + ||W_i||^2)^\eta) < \infty$ *for some* $\eta > \eta_1 \ge 0$.

(3.4) (i) $\mathbb{E}(||A(Z_i)\epsilon_i||^2) < \infty$, $\mathbb{E}(||h_0(W_i)\epsilon_i||^2) < \infty$, $\sigma_\epsilon^2 \equiv \mathbb{E}(\epsilon_i^2) < \infty$.
(ii) $\mathbb{E}(||A(Z_i)\epsilon_i||(1+||W_i||^2)^{\frac{\eta}{2}}) < \infty$, $\mathbb{E}(||A(Z_i)||^2) < \infty$, $\mathbb{E}(||h_0(W_i)||^2) < \infty$.

[11] Refer to Appendix for definition of $H(\cdot,\cdot)$-smooth.

(3.5) There is a function $b(\cdot)$ s.t. $b(\delta) \to 0$ as $\delta \to 0$ and

$$\begin{aligned} \mathbb{E}\left(\sup_{||\theta-\tilde{\theta}||<\delta} |g(X_i^*;\theta) - g(X_i^*;\tilde{\theta})|^2 \right) &\leq b^2(\delta) \\ \mathbb{E}\left(\sup_{||\theta-\tilde{\theta}||<\delta} |(Y_i - g(X_i^*;\tilde{\theta}))(1+||W_i||^2)^{\frac{\eta}{2}}| \right) &\leq \delta \end{aligned}$$

for a small positive value δ.

(3.6) For any function $l_1 \in \mathcal{S}$, and any $l_2 \in \Lambda_c^{\gamma}(\mathcal{W}.\eta)$, there are $\Pi_{\infty n} l_1$ and $\Pi_{\infty n} l_2$ in the sieve spaces $\mathcal{S}_n$ and $\mathcal{H}_n$ such that

$$\begin{aligned} ||l_1 - \Pi_{\infty n} l_1|| &= o_p(1) \\ ||l_2 - \Pi_{\infty n} l_2|| &= o_p(1). \end{aligned}$$

Also $\mathbb{E}[q^{k_{h(n)}}(W)' q^{k_{h(n)}}(W)]$ is non-singular uniformly in $k_{h(n)}$.

(3.1) is the usual rank condition for identification. (3.2)-(3.5) are standard conditions in the sieve literature. These conditions are tailored to accommodate the two nuisance functions. I extend results in Chen et al. (2003) Theorem 1, Pakes and Pollard (1989) Corrolary 3.2 and simultaneously control the influence of the two nuisance functions to the doubly robust moment conditions. (3.3) is similar to that in Chen et al. (2008) and the weighting η is needed since we allow the support of W_i to be unbounded. (3.6) requires that the two nuisance functions are well approximated by the sieve terms under the weighted sup-norm $||\cdot||_{\infty\eta}$. Note that the sieve spaces for estimating $p(\cdot)$ and $h(\cdot)$ can be very different.

Theorem 1. *Under Assumption 1, 2 and 3, if* $\frac{k_p(n)}{n} \to 0$, $\frac{k_h(n)}{n} \to 0$,$k_p(n) \to \infty$, $k_h(n) \to \infty$, *then the Gen-IV estimator* $\widehat{\theta}_{GenIV}$ *is consistent, i.e.* $\widehat{\theta}_{GenIV} - \theta_0 = o_p(1)$.

1.4.1.3 Asymptotic Normality

The next two assumptions are needed for the asymptotic normality of $\widehat{\theta}_{GenIV}$.

Assumption 4. *Let* $\theta_0 \in int(\Theta)$, $\mathbb{E}\left[\widetilde{\mathcal{Z}}_i\widetilde{\mathcal{Z}}_i'\epsilon_i^2\right]$ *be positive definite, the following hold*

(4.1) $\mathbb{E}\left[\widetilde{\mathcal{Z}}_{0i}\frac{\partial g(X_i^*;\theta)}{\partial\theta}\right]$ *exists for* $\theta \in \Theta_\delta \equiv \{\theta \in \Theta : ||\theta - \theta_0|| \leq \delta\}$ *and is continuous at* $\theta = \theta_0$, *where* $\widetilde{\mathcal{Z}}_{0i} \equiv \left(\frac{1-D_i}{1-p_0(W_i)}A(Z_i)' + \frac{D_i - p_0(W_i)}{1-p_0(W_i)}h_0(W_i)', V_i'\right)'$.

(4.2) $\mathbb{E}\left[\widetilde{\mathcal{Z}}_{0i}\frac{\partial g(X_i^*;\theta)}{\partial\theta}\right]|_{\theta=\theta_0}$ *is of full (column) rank.*

(4.3) $\mathbb{E}((1 + ||W_i||^2))^{2\eta} < \infty$.

(4.4) $\mathbb{E}(||A(Z_i)\epsilon_i||^4) < \infty$, $\mathbb{E}(||h_0(W_i)\epsilon_i||^4) < \infty$.

(4.5) For $\theta \in \Theta_\delta$,

$$\begin{aligned}
\mathbb{E}\left\{\sup_{||\theta-\theta_0||\leq\delta}\left|(A(Z_i) - h_0(W_i))\,(1 + ||W_i||^2)^{\frac{\eta}{2}}\epsilon_i\right|\right\} &< \infty \\
\mathbb{E}\left\{\sup_{||\theta-\theta_0||\leq\delta}(1 + ||W_i||^2)^{\frac{\eta}{2}}|Y_i - g(X_i^*;\theta)|\right\} &< \infty \\
\mathbb{E}\left\{\sup_{||\theta-\theta_0||\leq\delta}\left|\left|\frac{\partial g(X_i^*;\theta)}{\partial\theta}\right|\right|(1 + ||W_i||^2)^{\eta}\right\} &< \infty
\end{aligned}$$

(4.6) There exist functions $a_1(\cdot)$, $a_2(\cdot)$, s.t. $a_1(\delta) \to 0$, $a_2(\delta) \to 0$, as $\delta \to 0$,

$$\mathbb{E}\left\{\sup_{||\theta-\tilde{\theta}||\leq\delta}\left|A(Z_i)(g(X_i^*;\theta)-g(X_i^*;\tilde{\theta}))\right|^2\right\} \leq a_1^2(\delta)$$

$$\mathbb{E}\left\{\sup_{||\theta-\tilde{\theta}||\leq\delta}\left|g(X_i^*;\theta)-g(X_i^*;\tilde{\theta})\right|^4\right\} \leq a_2^4(\delta)$$

Assumption 5. *(5.1) Assumptions 3.1 and 3.2 hold with $\gamma > d_w/2$, and $\eta > \eta_1 + \gamma$.*

(5.2) Either (a) the growing speed of sieve terms

$$k_{p(n)} = O(n^{\frac{d_w}{2\gamma+d_w}}) \tag{1.8}$$

$$k_{h(n)} = O(n^{\frac{d_w}{2\gamma+d_w}}) \tag{1.9}$$

or (b) the $L_3(\mathcal{W})$-norms of the two nuisance functions satisfy convergence rates

$$||p(\cdot)-p_0(\cdot)||_3 = o_p(n^{-\frac{1}{6}}), \quad ||h(\cdot)-h_0(\cdot)||_3 = o_p(n^{-\frac{1}{6}})$$

holds.

Remark 2. Similar to consistency, I extend Chen et al. (2003) Theorem 2 and Pakes and Pollard (1989) Theorem 3.3 to a higher order asymptotics. To be more specific, if the growing speed for sieve terms is set as in (1.8), the resulting convergence rates for the two nuisance parameters would be $||\widehat{p}(\cdot)-p_0(\cdot)||_2 = O_p(n^{-\frac{\gamma}{2\gamma+d_w}})$, $||\widehat{h}(\cdot)-h_0(\cdot)||_2 = O_p(n^{-\frac{\gamma}{2\gamma+d_w}})$. These are the optimal rates in the sense of Stone (1982). And these rates are achievable by a lot of sieve terms, including othogonal series, splines, wavelets and sigmoid neural network sieve.

However, if we can actually have a slower convergence rate requirement for the nuisance functions than the optimal rates. (5.2) (b) is based on a second-order expansion for the remaining terms and sieve spaces like cosine neural network sieve, which is slower than the optimal rates (see other sieve space choices in Chen and Shen (1998)) will still be suitable choices.

Theorem 2. *Under Assumptions 1-4, the Gen-IV estimator* $\widehat{\theta}_{GenIV}$ *has* $\sqrt{n}(\widehat{\theta}_{GenIV} - \theta_0) \Rightarrow N(0, V_{GenIV})$*, with*

$$V_{GenIV} = (\mathfrak{J}_\theta' G \mathfrak{J}_\theta)^{-1} \mathfrak{J}_\theta' G \Omega_{GenIV} G \mathfrak{J}_\theta (\mathfrak{J}_\theta' G \mathfrak{J}_\theta)^{-1}$$

Furthermore, if $G = \Omega_{GenIV}^{-1}$*, then* $\sqrt{n}(\widehat{\theta}_{GenIV} - \theta_0) \Rightarrow N(0, V_0)$*, with*

$$V_0 = (\mathfrak{J}_\theta' \Omega_{GenIV}^{-1} \mathfrak{J}_\theta)^{-1}$$

where

$$\Omega_{GenIV} = \mathbb{E}\left(\frac{1}{1-p(W_i)} \mathbb{E}(\widetilde{Z}_i \widetilde{Z}_i' | W_i) \epsilon_i^2 - \frac{p(W_i)}{1-p(W_i)} \mathbb{E}(\widetilde{Z}_i | W_i) \mathbb{E}(\widetilde{Z}_i | W_i)' \epsilon_i^2 \right)$$

In practice, the standard errors are calculated according to the sample analog of the asymptotic variance. For example, Ω_{GenIV} can be estimated as

$$\begin{aligned} &\widehat{\Omega}_{GenIV} \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{1-\widehat{p}(W_i)} \mathbb{E}(\widehat{\widetilde{Z}_i \widetilde{Z}_i' | W_i}) \widehat{\epsilon}_i^2 - \frac{\widehat{p}(W_i)}{1-\widehat{p}(W_i)} \mathbb{E}(\widehat{\widetilde{Z}_i | W_i}) \mathbb{E}(\widehat{\widetilde{Z}_i | W_i})' \widehat{\epsilon}_i^2 \right) \end{aligned}$$

Note that $\widehat{p}(W_i)$ and $\mathbb{E}(\widehat{\widetilde{Z}_i | W_i})$ can be estimated using procedures proposed in Section 1.3. $\widehat{\epsilon}$ is a consistent estimate of the error term. At last, $\mathbb{E}(\widehat{\widetilde{Z}_i \widetilde{Z}_i' | W_i})$ can be estimated within the complete data sample since

$$\mathbb{E}(\widetilde{Z}_i \widetilde{Z}_i' | W_i) = \mathbb{E}(\widetilde{Z}_i \widetilde{Z}_i' | W_i, D_i = 0)$$

according to MAR.

1.4.2 Efficiency

In this section, I first state the semiparametric efficiency bounds for all the estimators derived through (2.1) under Assumption 1 and Assumption 2. I then compare the efficiency bound with the asymptotic variance of Gen-IV and shows that the estimator attains the bound if using optimal weighting matrix in the GMM estimation.

1.4.2.1 Semiparametric Efficiency Bounds

I consider efficiency bound for regular and asymptotically linear(RAL) estimators[12].

Theorem 3. *Let θ be defined by moment condition (2.1). Under Assumption 1 and Assumption 2, the asymptotic variance lower bound for all RAL estimator of θ is*

$$\left(Q'\Omega_{eff}^{-1}Q\right)^{-1}$$

where $Q = \mathbb{E}(\widetilde{Z}_i X_i^{*'})$ and

$$\Omega_{eff} = \mathbb{E}\left(\frac{1}{1-p(W_i)}\mathbb{E}(\widetilde{Z}_i\widetilde{Z}_i'|W_i)\epsilon_i^2 - \frac{p(W_i)}{1-p(W_i)}\mathbb{E}(\widetilde{Z}_i|W_i)\mathbb{E}(\widetilde{Z}_i|W_i)'\epsilon_i^2\right) \tag{1.10}$$

[12] Although most reasonable estimators are RAL, regular estimators do exist that are not asymptotically linear. However, as a consequence of Hájek (1970) representation theorem, it can be shown that the most efficient regular estimator is asymptotically linear; hence, it is reasonable to restrict attention to RAL estimators.

If we compare the efficiency bound in Theorem 3 to the *full data* efficiency bound when there is no missing instrument, we will find out that the only difference between the two bounds lie in the Ω term. In the *full data* efficiency bound, $\Omega_{full} = \mathbb{E}(\widetilde{Z}_i \widetilde{Z}_i' \epsilon_i^2)$. The difference between the two Ω terms is

$$\Delta_{loss} \equiv \Omega_{eff} - \Omega_{full} = \mathbb{E}\left(\frac{p(W_i)}{1 - p(W_i)} Var(\widetilde{Z}_i \epsilon_i | W_i) \right) \geq 0$$

where $Var(\cdot|W_i)$ is the conditional variance. The term Δ_{loss} quantifies the information loss due to the missing instrument issue.

1.4.2.2 Efficiency of Gen-IV Estimator

The following Corollary states the efficiency of the Gen-IV estimator.

Corollary 2. *The Gen-IV estimator attains the semiparametric efficiency bound specified in Theorem 3.*

A by-product in calculating the efficiency bound is the efficient influence function. In this case, the influence function used to derive θ_{GenIV} coincides with the efficient influence function.

1.5 Monte Carlo Experiments

The previous sections' results suggest that using Gen-IV estimator to deal with missing instruments should result in good estimation and inference properties. In this section, I provide simulation evidence regarding these properties. The simulation design incorporates both exogenous and endogenous missing mechanisms. I compare the Gen-IV estimator with five existing estimators and present the good performance of the new one.

1.5.1 Data Generating Process (DGP)

In this design, the simulations are based on a simple instrumental variables model data generating process (DGP) with single missing instrument Z_i ($D_i = 0$ if observed):

$$\begin{aligned}
Y_i &= \beta_0 + \alpha_0 X_i + \beta_1 V_i + \epsilon_i \\
X_i &= \gamma_0 + \gamma_1 Z_i + \gamma_2 V_i + \upsilon_i \\
D_i &= \mathbb{1}(\sin(\varrho_0 Y_i + \varrho_1 X_i + \varrho_2 V_i) + u_i \leq p) \\
(\epsilon_i, \upsilon_i) &\sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_{\epsilon\upsilon} \\ \rho_{\epsilon\upsilon} & 1 \end{pmatrix}\right) \\
(Z_i, V_i) &\sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_{zv} \\ \rho_{zv} & 1 \end{pmatrix}\right) \\
u_i &\sim Unif[0, 1]
\end{aligned}$$

where $\alpha_0 = \beta_0 = \beta_1 = 1$ are the parameters of interest. In all simulations, $\gamma_0 = \gamma_2 = 1$, $\rho_{\epsilon\upsilon} = 0.3$, $\rho_{zv} = 0.4$.

For other parameters, I consider various settings. I use two different specifications for the missing indicator D_i. In the first specification (DGP1), missingness is endogenous and depends on the dependent variable Y_i, the endogenous variable X_i, as well as the exogenous variable V_i. I set $\varrho_0 = -0.25$, $\varrho_1 = 0.5$, $\varrho_2 = 0.25$. In the second specification (DGP2), the missingness doesn't depend on any observed variable, $\varrho_0 = \varrho_1 = \varrho_2 = 0$. This missing pattern is referred to as the Missing-Completely-at-Random (MCAR) case in the statistics literature. For both DGP1 and DGP2, experiments are conducted with the specifications for

$$n \in \{250, 500\}, \quad p \in \{0.25, 0.5\}, \quad \gamma_1 \in \{1, 0.3\}$$

For each setting of the simulation parameter values, I report results from six different estimators. The estimators could have a unified representation as

$$\begin{aligned} \widehat{\theta} &= \left(X^{*'} P_{\widetilde{Z}} X^{*} \right)^{-1} X^{*'} P_{\widetilde{Z}} Y \\ P_{\widetilde{Z}} &= \widetilde{Z} \left(\widetilde{Z}' \widetilde{Z} \right)^{-1} \widetilde{Z}' \end{aligned}$$

The full data estimator is the 2SLS estimator using the *full data* sample, in which $\widetilde{Z}_i = (Z_i, V_i^{'})'$. The complete case estimator is the 2SLS estimator using the *complete data* subsample, in which $\widetilde{Z}_i = ((1 - D_i)Z_i, (1 - D_i)V_i^{'})'$. The GMM-Dummy estimator uses instrument set $\widetilde{Z}_i = ((1 - D_i)Z_i, D_i, V_i^{'})'$, which is proposed in Dahl and DellaVigna (2009). The GMM-AD estimator, proposed in Abrevaya and Donald (forthcoming), considers another GMM estimator in which the iteraction of missing indicator and exogenous variables $(1 - D_i)V_i$ is added to the instrument set, and $\widetilde{Z}_i = ((1 - D_i)Z_i, D_i, V_i^{'}, (1 - D_i)V_i^{'})'$. The IPW-IV estimator is an IV version of inverse propensity score weighted estimator, the instruments set of which is $\widetilde{Z}_i = (\frac{1-D_i}{1-p(W_i)} Z_i, V_i^{'})'$.

1.5.2 Results

For each estimator, three summary statistics are reported: median bias (MB), median absolute deviation (MAD), and root mean squared error (RMSE). Table 1.1 and Table 1.2 contain the summary statistics for the estimators for each of the experiments. The case I consider most relevant for applications is in Table 1.1.

There is clear evidence that estimates of three estimators, the Complete Case estimator, GMM-Dummy estimator, and GMM-AD estimator are very biased. Both MB and MAD are much larger for these three estimators than those of the rest. The estimates are even more biased when the instrument is relatively weak, i.e. $\gamma_1 = 0.3$. One possible reason for the biasedness is that all the three estimators are based on the moment condition $\mathbb{E}((1 - D_i)Z_i\epsilon_i) = 0$. When the missing indicator D_i depends on W_i , it is plausible that the moment condition $\mathbb{E}((1 - D_i)Z_i\epsilon_i) \neq 0$. To see this more explicitly, note that under Assumption 1 and using Law of iterated expectations,

$$\begin{aligned}\mathbb{E}((1 - D_i)Z_i\epsilon_i) &= \mathbb{E}\left(\mathbb{E}\left((1 - D_i)Z_i\epsilon_i|W_i, Z_i\right)\right)\\ &= \mathbb{E}(\mathbb{E}((1 - D_i)|W_i)Z_i\epsilon_i)\\ &= \mathbb{E}((1 - p(W_i))Z_i\epsilon_i)\end{aligned}$$

The *full data* moment function $Z_i\epsilon_i$ is multiplied by the propensity score $1 - p(W_i)$, which could results in the inconsistency of estimation based on this moment equation. This also explains the necessity for the IPW-IV and Gen-IV estimators to adjust the *observed data* moment function $(1 - D_i)Z_i\epsilon_i$ by the inverse of propensity score $\frac{1}{1-p(W_i)}$.

Since the error term ϵ_i can be consistently estimated via Gen-IV, one can directly test the validity of using such moment condition based on the following null and alternative hypothesis:

$$H_0 :\mathbb{E}((1 - D_i)Z_i\epsilon_i) = 0$$

$$H_1 :\mathbb{E}((1 - D_i)Z_i\epsilon_i) \neq 0$$

Overall, the Gen-IV estimator dominates the IPW-IV estimator in the perspective of every summary statistic. The performance of the Gen-IV estimator is the closest to that of the infeasible Full data estimator among all the estimators. Even when there is nearly one quarter of missing instruments, the summary statistics for the full data estimator and the Gen-IV estimator are still quantitatively similar.

To further explore how the missing proportion will influence the behavior of estimators, I conduct a Wald-type test with "$H_0 : \alpha_0 = \beta_0 = \beta_1 = 1$" for a wide range of values of p, $p \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. I report rejection frequencies of 5% level tests for each of the six estimators under DGP1 in Table 1.3. The rejection frequencies for the three biased estimators, Complete Case, GMM-Dummy and GMM-AD are quite high, even if the missing proportion p is as low as 5%. When the instrument is relatively weak, $\gamma_1 = 0.3$, there are more than half of the iterations in which Complete Case, GMM-Dummy, and GMM-AD estimators reject the null. However, there is no clear conclusion about the patterns of Gen-IV estimator and IPW-IV estimator. Both of them have similar rejection frequencies. And these frequencies are quantitatively similar compared to the Full data estimator. One interesting finding in Table 1.3 is that the rejection frequencies for Gen-IV estimator and

IPW-IV estimator won't change much as the missing proportion increases.

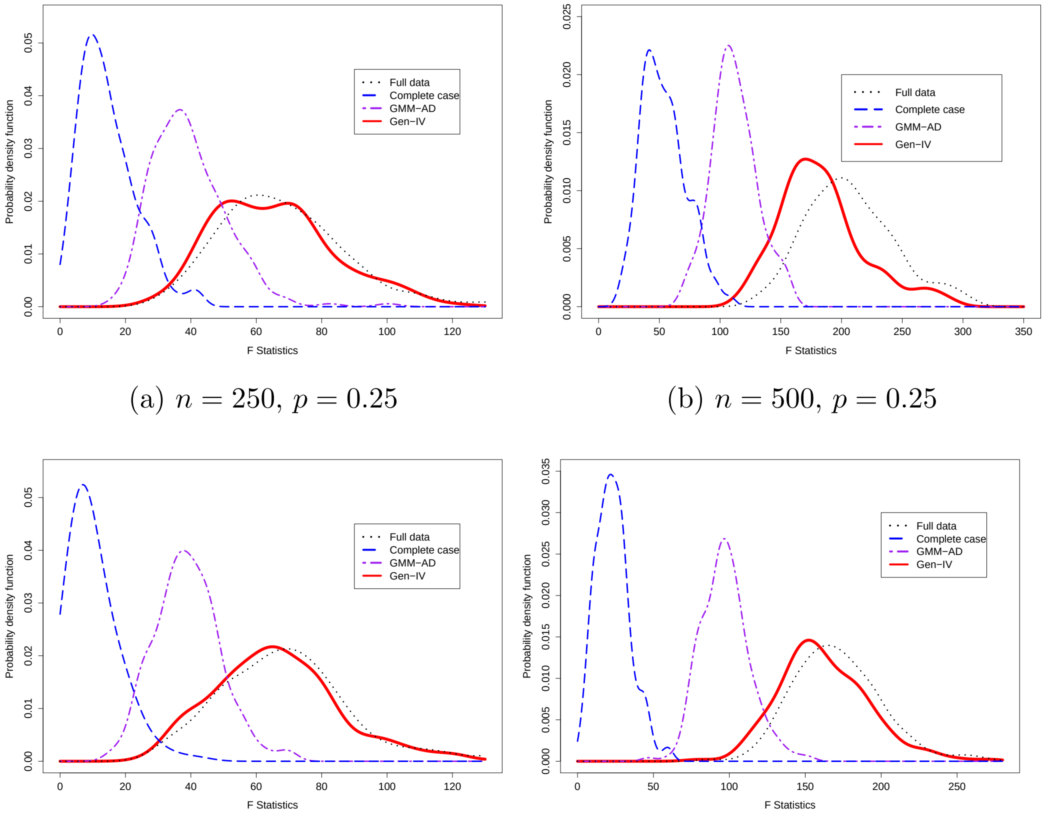
Table 1.2 summarizes the behavior of the six estimators under DGP2. Since D_i is completely exogenous, it holds that

$$\mathbb{E}((1-D_i)Z_i\epsilon_i) = \mathbb{E}(1-D_i)\underbrace{\mathbb{E}(Z_i\epsilon_i)}_{0} = 0$$

In this case, all the estimators are consistent. There does not seem to exist a best estimator in terms of MB and MAD. The Gen-IV estimator has slightly smaller RMSE than the others. In particular, the advantage of the Gen-IV estimator is more obvious when the missing proportion is higher, $p = 0.5$, or when the instrument is stronger, $\gamma_1 = 1$.

To compare the identifying power of instrument sets from different estimators, I draw distributions of the F statistic from reduced form regression in Figure 1. These figures present density estimates of the F statistic when instrument Z_i is relatively weak $\gamma_1 = 0.3$ under DGP1. Results show that the identifying power of generated instruments is very close to the full data instruments. On the other hand, restricting estimation within the *complete data* sample will severely contaminate the identifying power of IV. Sometimes when missing proportion is high (e.g., (c) $n = 250$, $p = 0.5$), researchers might get wrong conclusion about the strength of the instrument, with suspicion of weak instrument.

Figure 1.1: Distribution of the first-stage F statistic: DGP1



(a) $n = 250$, $p = 0.25$

(b) $n = 500$, $p = 0.25$

(c) $n = 250$, $p = 0.5$

(d) $n = 500$, $p = 0.5$

Table 1.1: Summary Statistics for DGP1

	α_0			β_0			β_1		
	MB	MAD	RMSE	MB	MAD	RMSE	MB	MAD	RMSE
Sample A: n=250, $p = 0.25$,$\gamma_1 = 1$									
Full data	-0.0016	0.0479	0.0726	0.0080	0.0652	0.0959	-0.0070	0.0874	0.1275
Complete Case	0.0961	0.1048	0.1333	-0.3174	0.3174	0.3501	-0.0037	0.0922	0.1525
GMM-Dummy	-0.1353	0.1353	0.1700	0.1361	0.1436	0.1863	0.1867	0.1867	0.2552
GMM-AD	-0.1328	0.1328	0.1637	0.1212	0.1245	0.1811	0.1797	0.1797	0.2454
IPW-IV	0.0516	0.0721	0.1058	-0.0375	0.0755	0.1179	-0.0657	0.1182	0.1713
Gen-IV	0.0004	0.0553	0.0839	0.0018	0.0728	0.1029	-0.0064	0.1016	0.1451
Sample B: n=250, $p = 0.25$,$\gamma_1 = 0.3$									
Full data	-0.0069	0.0997	0.1569	-0.0033	0.0967	0.1680	0.0071	0.1298	0.1958
Complete Case	0.0815	0.1168	0.1853	-0.2952	0.2952	0.3686	0.0269	0.1178	0.2079
GMM-Dummy	-0.3654	0.3654	0.4214	0.3635	0.3635	0.4225	0.4120	0.4120	0.5023
GMM-AD	-0.3345	0.3345	0.4053	0.3452	0.3452	0.4059	0.3847	0.3847	0.4844
IPW-IV	0.0579	0.1246	0.1869	-0.0580	0.1443	0.1990	-0.0713	0.1482	0.2322
Gen-IV	-0.0205	0.1065	0.1744	0.0231	0.1184	0.1847	0.0198	0.1385	0.2171
Sample C: n=250, $p = 0.5$,$\gamma_1 = 1$									
Full data	-0.0093	0.0482	0.0713	0.0072	0.0608	0.0965	0.0061	0.0816	0.1201
Complete Case	0.1231	0.1271	0.1530	-0.4731	0.4731	0.5205	0.0166	0.1022	0.1485
GMM-Dummy	-0.2032	0.2032	0.2265	0.2174	0.2174	0.2415	0.2996	0.2996	0.3374
GMM-AD	-0.1933	0.1933	0.2181	0.2048	0.2048	0.2329	0.2853	0.2853	0.3266
IPW-IV	-0.0022	0.0658	0.0966	0.0232	0.0761	0.1176	0.0149	0.1002	0.1555
Gen-IV	-0.0169	0.0695	0.0988	0.0155	0.0724	0.1044	0.0355	0.1124	0.1577
Sample D: n=500, $p = 0.25$,$\gamma_1 = 1$									
Full data	0.0025	0.0335	0.0482	0.0031	0.0464	0.0671	-0.0015	0.0601	0.0827
Complete Case	0.0953	0.0953	0.1139	-0.3249	0.3249	0.3417	-0.0019	0.0558	0.0926
GMM-Dummy	-0.1380	0.1380	0.1540	0.1439	0.1439	0.1663	0.1996	0.1996	0.2243
GMM-AD	-0.1362	0.1362	0.1498	0.1389	0.1389	0.1620	0.1921	0.1921	0.2189
IPW-IV	0.0530	0.0567	0.0829	-0.0467	0.0646	0.0944	-0.0729	0.0852	0.1220
Gen-IV	-0.0005	0.0355	0.0586	0.0064	0.0508	0.0753	0.0078	0.0583	0.0944
Sample E: n=500, $p = 0.25$,$\gamma_1 = 0.3$									
Full data	-0.0019	0.0688	0.1009	0.0095	0.0769	0.1096	0.0018	0.0851	0.1298
Complete Case	0.0870	0.1067	0.1559	-0.3149	0.3149	0.3516	0.0087	0.0976	0.1424
GMM-Dummy	-0.3562	0.3562	0.3802	0.3671	0.3671	0.3884	0.4385	0.4385	0.4580
GMM-AD	-0.3468	0.3468	0.3690	0.3585	0.3585	0.3770	0.4203	0.4203	0.4452
IPW-IV	0.0755	0.1071	0.1509	-0.0679	0.1031	0.1560	-0.1046	0.1401	0.1861
Gen-IV	-0.0098	0.0805	0.1295	0.0107	0.0870	0.1350	0.0009	0.1027	0.1606
Sample F: n=500, $p = 0.5$,$\gamma_1 = 1$									
Full data	0.0024	0.0367	0.0518	-0.0058	0.0490	0.0703	0.0027	0.0639	0.0844
Complete Case	0.1260	0.1260	0.1519	-0.5170	0.5170	0.5276	0.0009	0.0765	0.1143
GMM-Dummy	-0.1895	0.1895	0.2033	0.1892	0.1892	0.2137	0.2718	0.2718	0.2923
GMM-AD	-0.1853	0.1853	0.1963	0.1847	0.1847	0.5070	0.2578	0.2578	0.2830
IPW-IV	0.0021	0.0538	0.0750	-0.0107	0.0650	0.090	-0.0016	0.0840	0.1172
Gen-IV	0.0053	0.0485	0.0716	-0.0091	0.0573	0.0860	-0.0058	0.0765	0.1100

Table 1.2: Summary Statistics for DGP2

	α_0			β_0			β_1		
	MB	MAD	RMSE	MB	MAD	RMSE	MB	MAD	RMSE
Sample A: n=250, $p = 0.25$,$\gamma_1 = 1$									
Full data	-0.0050	0.0519	0.0719	-0.0099	0.0673	0.1011	0.0050	0.0818	0.1236
Complete Case	0.0036	0.0509	0.0814	-0.0139	0.0777	0.1186	0.0070	0.0846	0.1326
GMM-Dummy	0.0040	0.0549	0.0826	-0.0201	0.0727	0.1102	0.0067	0.0881	0.1325
GMM-AD	0.0076	0.0531	0.0789	-0.0144	0.0718	0.1075	-0.0035	0.0922	0.1289
IPW-IV	-0.0008	0.0553	0.0828	-0.0089	0.0755	0.1103	0.0123	0.0913	0.1333
Gen-IV	-0.0017	0.0483	0.0764	-0.0077	0.0692	0.1031	0.0077	0.0873	0.1270
Sample B: n=250, $p = 0.25$,$\gamma_1 = 0.3$									
Full data	-0.0025	0.0933	0.1521	0.0007	0.1041	0.1720	0.0034	0.1268	0.2004
Complete Case	-0.0212	0.1062	0.1988	0.0201	0.1111	0.2163	0.0119	0.1460	0.2566
GMM-Dummy	-0.0076	0.1145	0.1688	0.0032	0.1107	0.1819	0.0185	0.1296	0.2196
GMM-AD	-0.0176	0.1038	0.1584	0.0016	0.1107	0.1722	0.0070	0.1262	0.2070
IPW-IV	-0.0183	0.1256	0.1987	0.0166	0.1178	0.2188	0.0241	0.1532	0.2512
Gen-IV	-0.0293	0.1104	0.1915	0.0145	0.1181	0.2120	0.0175	0.1482	0.2437
Sample C: n=250, $p = 0.5$,$\gamma_1 = 1$									
Full data	0.0120	0.0519	0.0814	-0.0028	0.0668	0.0976	-0.0123	0.0855	0.1326
Complete Case	-0.0010	0.0672	0.1070	-0.0049	0.0824	0.1363	-0.0242	0.1074	0.1736
GMM-Dummy	0.0071	0.0634	0.1113	0.0007	0.0829	0.1220	-0.0099	0.1104	0.1743
GMM-AD	0.0055	0.0654	0.1086	-0.0127	0.0815	0.1206	-0.0146	0.1003	0.1673
IPW-IV	-0.0025	0.0644	0.1184	-0.0005	0.0817	0.1304	0.0051	0.1060	0.1823
Gen-IV	0.0103	0.0670	0.0968	-0.0009	0.0698	0.1082	0.0047	0.0968	0.1543
Sample D: n=500, $p = 0.25$,$\gamma_1 = 1$									
Full data	0.0007	0.0310	0.0504	0.0033	0.0420	0.0665	-0.0052	0.0598	0.0865
Complete Case	-0.0029	0.0376	0.0593	-0.0023	0.0459	0.0785	-0.0006	0.0714	0.1010
GMM-Dummy	-0.0053	0.0405	0.0604	0.0027	0.0474	0.0761	0.0039	0.0733	0.0988
GMM-AD	-0.0012	0.0399	0.0600	-0.0018	0.0474	0.0760	-0.0050	0.0693	0.0983
IPW-IV	-0.0057	0.0387	0.0605	0.0006	0.0466	0.0761	0.0056	0.0728	0.0991
Gen-IV	0.0006	0.0338	0.0548	0.0026	0.0459	0.0718	-0.0006	0.0667	0.0913
Sample E: n=500, $p = 0.25$,$\gamma_1 = 0.3$									
Full data	0.0039	0.0752	0.1037	-0.0026	0.0786	0.1121	0.0089	0.0967	0.1270
Complete Case	-0.0111	0.0803	0.1186	0.0006	0.0828	0.1286	0.0157	0.1000	0.1464
GMM-Dummy	-0.0065	0.0868	0.1213	-0.0059	0.0888	0.1298	0.0078	0.1003	0.1494
GMM-AD	-0.0037	0.0828	0.1154	-0.0002	0.0830	0.1247	0.0031	0.1017	0.1424
IPW-IV	-0.0088	0.0843	0.1225	0.0291	0.0861	0.1307	0.0200	0.1019	0.1499
Gen-IV	-0.0115	0.0820	0.1177	-0.0046	0.0838	0.1261	0.0236	0.0988	0.1441
Sample F: n=500, $p = 0.5$,$\gamma_1 = 1$									
Full data	-0.0020	0.0360	0.0512	0.0006	0.0397	0.0680	0.0027	0.0545	0.0789
Complete Case	0.0040	0.0454	0.0706	-0.0032	0.0552	0.0909	-0.0024	0.0743	0.1113
GMM-Dummy	0.0132	0.0487	0.0757	-0.0147	0.0529	0.0875	-0.0126	0.0705	0.1109
GMM-AD	0.0043	0.0474	0.0711	-0.0118	0.0519	0.0843	-0.0109	0.0619	0.1056
IPW-IV	0.0083	0.0485	0.0754	-0.0144	0.0540	0.0872	-0.0147	0.0670	0.1108
Gen-IV	0.0027	0.0418	0.0639	-0.0118	0.0507	0.0806	-0.0047	0.0636	0.0943

Table 1.3: Rejection Rates: DGP2

p	Full data	Complete Case	GMM-D	GMM-AD	IPW-IV	Gen-IV
Sample A: $n = 500$, $\gamma_1 = 1$						
5%	0.06	0.79	0.35	0.345	0.065	0.06
10%	0.05	0.77	0.415	0.41	0.075	0.065
20%	0.065	0.89	0.57	0.535	0.08	0.08
30%	0.02	0.96	0.6	0.59	0.04	0.045
40%	0.065	0.97	0.705	0.68	0.13	0.08
50%	0.025	0.995	0.69	0.665	0.115	0.04
60%	0.075	0.98	0.775	0.775	0.09	0.055
70%	0.08	0.995	0.79	0.785	0.185	0.11
Sample B: $n = 500$, $\gamma_1 = 0.3$						
5%	0.045	0.69	0.88	0.87	0.055	0.065
10%	0.035	0.755	0.945	0.89	0.035	0.03
20%	0.035	0.905	0.965	0.95	0.035	0.03
30%	0.03	0.97	0.985	0.985	0.065	0.05
40%	0.08	0.985	0.99	0.99	0.075	0.065
50%	0.06	0.975	0.99	0.995	0.07	0.065
60%	0.01	0.995	0.99	0.995	0.04	0.04
70%	0.045	0.99	0.995	0.995	0.035	0.065

1.6 Application

In this section, I apply the new estimation methodology to study the causal effect of education in labor market outcomes. It is well understood in the literature that education is endogenous. One of the famous candidate instruments for education is the college proximity. Card (1995) uses an indicator for the presence of an accredited 4-year college in the local labor market as an instrument for education.

Other factors like ability affect both education and wage at the same time. Ability then enters into the wage regression as an important confounder. And there has been a long tradition in the literature to use "Knowledge of the World of Work" (KWW) test score[13] as a measure of "ability", which can date back to Griliches (1976), Griliches (1977). A potential criticism about KWW is that it is treated as an error-free measure of "'ability". To address this criticism, IQ score is used to instrument for the KWW score.

I use the following specification in this section:

$$lwage_i = \beta_0 + \beta_1 \underbrace{Education_i}_{IV: college\ proximity} + \alpha \underbrace{KWW_i}_{IV: IQ\ score} + other\ controls^{'}\gamma + \epsilon_i$$

where education and KWW score are instrumented by college proximity and IQ score, respectively. The dependent variable is the log of weekly wage.

[13] In the NLSYM76 dataset, the KWW test items were questions on the job activities of 10 specific occupations, the education requirements for these 10 occupations, and the relative earnings of 8 different pairs of occupations, with a total of 28 items.

1.6.1 Data and Missing Instruments

The data sample consists of 2,963 observations on male workers from the National Longitudinal Survey of Young Men (NLSYM) in 1976. Other control variables include years of experience (and its square), an SMSA indicator (=1 if living in an SMSA in 1976), a South indicator (=1 if living in the south in 1976), and a black-race indicator. Regions dummies include 9 region indicators and family background consists of 14 variables representing mother's and father's education, indicators for missing father's or mother's education, interactions of mother's and father's education, and dummies for family structure at age 14.

The IQ data are missing for 923 observations. The *complete data* sample, where IQ scores are observed, has 2,040 observations.

Simulation studies in the previous section suggest that Complete Data method only works when the instrument is MCAR. Table 1.4 checks the dependence of the missing indicator on completely observed variables by running Logit and Probit regression of D_i on W_i. Significant coefficients are found for KWW, education, experience (and its square) and the South indicator. And results are robust if we include family background as well as the region dummies. There is clear evidence that IQ is not MCAR, which implies that the results under Complete Data method might not be reliable.

Table 1.4: Checking for MCAR of IQ

	Missingness of IQ Score, (D)					
	Logit-I	Probit-I	Logit-II	Probit-II	Logit-III	Probit-III
	(1)	(2)	(3)	(4)	(5)	(6)
Wage	-0.1685	-0.0927	-0.1633	-0.0890	-0.1618	-0.0886
	(0.1221)	(0.0708)	(0.1227)	(0.0711)	(0.1243)	(0.0718)
Ability(KWW)	-0.0307***	-0.0181***	-0.0285***	-0.0167***	-0.0275***	-0.0160***
	(0.0072)	(0.0042)	(0.0073)	(0.0043)	(0.0074)	(0.0043)
Education	-0.2529***	-0.1385***	-0.2285***	-0.1253***	-0.2338***	-0.1286***
	(0.0223)	(0.0125)	(0.0235)	(0.0132)	(0.0237)	(0.0133)
Experience	-3.1546***	-1.8111***	-3.2040***	-1.8415***	-3.2598***	-1.8744***
	(0.3019)	(0.1739)	(0.3033)	(0.1745)	(0.3050)	(0.1754)
Experience squared	5.3667***	3.0849***	5.4426***	3.1322***	5.5355***	3.1863***
	(0.5236)	(0.3016)	(0.5258)	(0.3025)	(0.5286)	(0.3040)
Black	0.8172***	0.5018***	0.7046***	0.4317***	0.7160***	0.4411***
	(0.1188)	(0.0708)	(0.1239)	(0.0737)	(0.1251)	(0.0744)
SMSA	-0.0258	-0.0110	-0.0138	-0.0058	-0.0008	0.0015
	(0.1037)	(0.0605)	(0.1045)	(0.0609)	(0.1077)	(0.0626)
South	0.4307***	0.2461***	0.3959***	0.2287***	0.2604	0.1450
	(0.1020)	(0.0597)	(0.1032)	(0.0602)	(0.2082)	(0.1207)
Constant	49.8121***	28.4567***	51.1235***	29.2531***	52.1515***	29.8705***
	(4.3385)	(2.4941)	(4.3757)	(2.5105)	(4.4036)	(2.5246)
Family background	N	N	Y	Y	Y	Y
Region dummies	N	N	N	N	Y	Y
N	2,963	2,963	2,963	2,963	2,963	2,963

1.6.2 Results

Table 1.5 reports the IV estimation results. Column (1) treats KWW as exogenous, as one of the specifications considered in Card (1995). It is served as a benchmark for the other two IV results. Column (2) is the procedure adopted in Card (1995) where IV estimation is conducted only to the complete data sample. Column (3) is based on the generated IQ scores. Results show that the return to education is insignificant using Complete Data method. Instead, the proposed method in this paper will restore the significance of return to education. Meanwhile, the standard errors of most of the coefficients are smaller compared to Column (2).

Table 1.5: Instrumental Variables Estimation of Return to Education

	Weekly log(Wage)(Dependent variable)		
	KWW Exogenous 2SLS (1)	KWW Endogenous Complete Case (2)	KWW Endogenous Gen-IV (3)
Education	0.136**	0.089	0.131**
	(0.078)	(0.085)	(0.076)
KWW	-0.014	0.037	-0.142
	(0.089)	(0.249)	(0.220)
Experience	0.063***	0.059**	0.096***
	(0.019)	(0.035)	(0.032)
Experience Square	-0.117	-0.112	-0.220**
	(0.108)	(0.105)	(0.092)
Black	-0.172***	-0.138	-0.256**
	(0.036)	(0.119)	(0.116)
SMSA	0.091**	0.117***	0.123***
	(0.044)	(0.025)	(0.023)
South	-0.145***	-0.102***	-0.148***
	(0.028)	(0.032)	(0.028)
Constant	4.040***	4.278***	4.180***
	(0.657)	(0.654)	(0.539)
Family Background	Y	Y	Y
Region Dummies	Y	Y	Y
N	2,963	2,040	2,963

1.7 Conclusion

I study consistent and efficient IV estimation when instruments are missing endogenously. Under a conditional version of “missing at random” assumption, I am able to generate new instruments for every observation in the original data sample. With a generated instrument set, the identifying power of the infeasible *full data* instrument can be largely restored, making valid inference possible.

Empirical researchers need to be more cautious when facing missing instruments in the data set. If endogenous missingness exists, simply ignoring the observations with missing instruments may result in insignificant coefficients of interest and very large standard errors. Furthermore, the diagnosis about the strength of IV could also be wrong.

There are several directions for future research. First, it is interesting to investigate other missing data problems in IV estimation. Examples include missing endogenous variables and missing dependent variables. Inference methods under other more complicated situations e.g. IV estimation entailing missing instruments and weak instruments would also worth a try. Second, one can develop formal tests on the consistency of Complete Data methods based on certain moment conditions. Also, the theoretical results in this paper has focused on monotone missing patterns of instruments. The idea of generated instruments can be extended to multiple/non-monotone missingness of instruments.

Chapter 2

Methods for Optimal Instruments with Many Missing Instruments

2.1 Introduction

In this chapter, I study another case of missing instruments, i.e. many missing instruments. It occurs in empirical studies when one has a rich instrument set but each instrument can have missing values. The methodology developed in this chapter is an extension to the generated instrument approach in Chapter 1. I also propose a three-step estimation procedure. In the first step, many generated instruments are formed and estimated. These generated instruments are used to improve the efficiency of IV estimation or approximate the infeasible optimal instruments in the spirit of Amemiya (1974), Chamberlain (1987), and Newey (1990). Although the improvement in efficiency is attractive, there are two potential problems with generating many instruments. The first problem is the well-known "many-instrument" problem where the IV estimators based on many instruments may have poor properties.[1] The second problem is the possible large estimation error in the formation of the many generated instruments. Keeping these two problems in mind, in the

[1] These poor properties include inaccurate inference and large standard errors. See Bekker (1994), Hansen et al. (2012), Chao et al. (2012) for discussions on these problems.

second step, I develop a shrinkage-based method for estimating the reduced form regression of the endogenous variable on the many generated instruments. My method can accomplish selection among many generated instruments and parameter estimates within one step. At the same time, it controls for the estimation bias brought by first-step estimation. I extend the methods of Belloni et al. (2012) to a pseudo-approximation of the optimal instruments. The IV estimation proceeds in the third step by regressing the dependent variable on the pseudo-approximation.

The approach is new and easy for implementation. It recovers the full data instrument set from the original many missing instruments. The new instrument set does not suffer from missing data issues again. I also allow a flexible instrument set in which the number of instruments can be increasing with the sample size and can even exceed the sample size. At the same time, every instrument in the instrument set could have missing values. Both Muris (2011) and Chaudhuri and Guilkey (2013) consider multiple missing data problems including missing instruments. But the instrument set is fixed in their settings. To my knowledge, this is the first paper in the literature to study many missing instruments. In particular, I am able to show that under a "pseudo" sparsity condition and several regularity conditions, the parameters of interest are estimated at the parametric rate.

This paper also makes several theoretical contributions. First, I calculate semiparametric efficiency bounds under conditional moment equality when the conditioning variable has missing data. Hristache and Patilea (2014) characterize the semiparametric efficiency bounds for conditional moment restriction models with different conditioning variables as a decreasing sequence of unconditional moment restriction models. An iterative procedure for approximating the efficient score when this is not explicit is provided. My paper complements this strand of literature by analytically characterizing the optimal instruments and I also provide a direct approximation of the optimal instruments under certain conditions.

This chapter is organized as follows. Section 2.2 presents the IV model with many missing instruments. Section 2.3 proposes the three-step estimation approach as well as the Pen-Gen-IV estimator. Section 2.4 establishes $\sqrt{n}$-consistency of the proposed estimator and states the semiparametric efficiency bounds. It also discusses the efficiency of the proposed estimator and characterizes the "pseudo" optimal instruments. Section 2.5 provides simulation evidence of finite sample behavior of the Pen-Gen-IV estimator. Section 2.6 concludes the chapter.

2.2 The Many Missing Instruments Case

In the previous chapter, I consider the single missing instrument case, in which the *full data* moment condition is

$$\mathbb{E}(\widetilde{Z}_i \epsilon_i) = 0 \tag{2.1}$$

The instrument set $\widetilde{Z}_i$ satisfies the unconditional independence assumption and the parameter of interest θ is exactly identified. In this section, I proceed by investigating an over identification case when the instrument set $\widetilde{Z}_i$ satisfies the conditional independence assumption,

$$\mathbb{E}(\epsilon_i|\widetilde{Z}_i) = 0 \tag{2.2}$$

Estimation based on conditional moment restriction like (2.2) has been rigorously studied in the literature, e.g. Amemiya (1974), Newey (1990), Blundell and Powell (2003), Chen and Pouzo (2015), Chernozhukov et al. (2015), to name a few. The choice of instrument set is more flexible under (2.2) than (2.1), since every measurable function $A(\widetilde{Z}_i)$ (assuming expectation exists) will qualify as an instrument in the sense that $\mathbb{E}(A(\widetilde{Z}_i)\epsilon_i) = 0$. It is also known that setting $A(\widetilde{Z}_i) = \mathcal{D}(\widetilde{Z}_i) \equiv \mathbb{E}(X_i^*|\widetilde{Z}_i)$ will minimize the asymptotic variance of θ under *full data* moment condition $\mathbb{E}(\epsilon_i|\widetilde{Z}_i) = 0$. $\mathcal{D}(\widetilde{Z}_i)$ is called the optimal instrument in the literature. In our case, let $\mathcal{D}_i \equiv \mathcal{D}(\widetilde{Z}_i) = (\mathbb{E}(X_i|\widetilde{Z}_i), V_i^{'})^{'}$. For notational convenience, I assume that there are no other exogenous variables and V_i only contains a constant term ($V_i = 1$). With some abuse of notation, let $Z_i \equiv (Z_i, 1)^{'}$. Then $\mathcal{D}_i = \mathcal{D}(Z_i) = \mathbb{E}(X_i|Z_i)$.

The many missing instruments case hence refers to the situation in which (i) a rich instrument set $A(Z_i) = (A_1(Z_i), ..., A_t(Z_i))^{'}$ is used to approximate the *full data* optimal instrument $\mathbb{E}(X_i|Z_i)$; (ii) every instrument $A_k(Z_i)$ in the instrument set has missing values, $k = 1, ..., t$. For example, in the missing IQ case, empirical researchers may want to add (a) higher order polynomianls $(IQ, IQ^2, IQ^3, ...)$ and/or (b) interaction terms $(IQ \times Fathereduc, IQ \times Mothereduc, IQ \times Family\ Background, IQ \times Region\ Dummies, ...)$ to the instrument set[2]. Each instrument in (a) or (b) can be viewed as a function of IQ scores, and is thus a missing instrument.

Similar to Lemma 1, there is an observational equivalence between the IV model with many missing instruments and conditional moment restrictions,

Lemma 3 (Identification(2))**.** *The many missing instruments problem with conditional independence assumption (2.2) under Assumption 1 and 2 is observationally equivalent to the following conditional moment restrictions.*

$$\mathbb{E}\left(\frac{1-D_i}{1-p(W_i)}\epsilon_i|Z_i\right) = 0 \tag{2.3}$$

$$\mathbb{E}\left(\frac{p(W_i)-D_i}{1-p(W_i)}|W_i\right) = 0 \tag{2.4}$$

Conditional moment restriction (2.3) is equivalent to infinite number of unconditional moment restrictions. Any measurable function $A(\cdot) \in L_2(\mathcal{Z})$, where $\mathcal{Z}$ is the support of Z, satisfies

$$\mathbb{E}\left(\frac{1-D_i}{1-p(W_i)}A(Z_i)\epsilon_i\right) = 0 \tag{2.5}$$

Also recall the equivalence of (2.4) with unconditional moment

$$\mathbb{E}\left(\frac{p(W_i)-D_i}{1-p(W_i)}g(W_i)\right) = 0 \tag{2.6}$$

The moment condition I am going to use for estimation is then based upon the inverse propensity score weighted combination of (2.6) and (2.5):

$$\mathbb{E}\left(\frac{1-D_i}{1-p(W_i)}\underbrace{\mathcal{D}(Z_i)}_{A(Z_i)}\epsilon_i + \frac{D_i-p(W_i)}{1-p(W_i)}\underbrace{\mathbb{E}\left(\mathcal{D}(Z_i)|W_i\right)\epsilon_i}_{g(W_i)}\right) = 0 \tag{2.7}$$

[2] Other family background variables may include family structures and various interactions of parental education and family structures.

where $\mathcal{D}(Z_i) = \mathbb{E}(X_i|Z_i)$ is the *full data* optimal instrument.

Remark 3. The moment condition used in the single missing instrument case

$$\mathbb{E}\left(\frac{1-D_i}{1-p(W_i)}Z_i\epsilon_i + \frac{D_i - p(W_i)}{1-p(W_i)}\mathbb{E}(Z_i|W_i)\epsilon_i\right) = 0$$

can be viewed as a special case of moment condition (2.7) in the sense that only one *full data* instrument Z_i is used to approximate $\mathbb{E}(X_i|Z_i)$.

Notice that without imposing further restrictions, the *full data* reduced form regression equation can be written as

$$X_i = \mathcal{D}(Z_i) + \nu_i, \qquad \mathbb{E}(\nu_i|Z_i) = 0.$$

When Z_i has missing values, a direct approximation of $\mathcal{D}(Z_i)$ through instrument set $A(Z_i) = (A_1(Z_i), ..., A_t(Z_i))'$ is impossible. Moment condition (2.7) suggests an indirect, pseudo approximation of the *full data* optimal instrument by generated instruments. Namely, consider a reduced form regression equation

$$X_i = \mathcal{A}_i\left(\mathcal{D}(Z_i), Z_i, W_i, D_i; p, h\right) + \nu_i, \qquad \mathbb{E}\left(\nu_i|Z_i\right) = 0 \tag{2.8}$$

where $\mathcal{A}_i\left(\mathcal{D}(Z_i), Z_i, W_i, D_i; p, h\right) = \frac{1-D_i}{1-p(W_i)}\mathcal{D}(Z_i) + \frac{D_i-p(W_i)}{1-p(W_i)}\mathbb{E}(\mathcal{D}(Z_i)|W_i)$. I call $\mathcal{A}_i(\cdot; p, h)$ the *observed data* optimal instrument to distinguish from the *full data* optimal instrument $\mathcal{D}(Z_i)$. To see why we can have such reduced form regression, recall that by Assumption 1, $\mathbb{E}(\mathcal{A}_i(\cdot; p, h)|W_i) = \mathbb{E}(\mathcal{D}(Z_i)|W_i)$, and $\mathbb{E}(\mathcal{D}(Z_i)|W_i) + \mathbb{E}(\nu_i|W_i) = \mathbb{E}(X_i|W_i) = X_i$, since W_i includes X_i.

2.3 Estimation

According to moment condition (2.7), I now propose a three-step procedure for the estimation of θ. This procedure differs from the previous three-step estimation with single missing instrument in two ways. First, in the first step, there is now a (very) large list of generated instruments to be estimated. Second, I now consider a nonparametric reduced form regression instead of the parametric specification in the single missing instrument case. Specifically, I extend Belloni et al. (2012)'s Lasso-based methods on estimating the *full data* optimal instrument, to accommodate the (very) many generated instrument setting and to approximate the *observed data* optimal instrument. The following details the three-step estimation.

Step 1 Estimation of many generated instruments

Suppose a very large list of instruments,

$$A_i \equiv A(Z_i) \equiv (A_1(Z_i), ..., A_t(Z_i))'$$

is chosen to estimate the *full data* optimal instrument $\mathcal{D}(Z_i)$. The number of instruments t is possibly much larger than the sample size n ($t >> n$). t can also be increasing with the sample size n. As mentioned in Belloni et al. (2012), by allowing t to be much larger than the sample size, we are able to consider many more instruments than in Newey (1990) and Hahn (2002). The purpose of this is to get a better approximation of the optimal instrument to improve estimation efficiency.

Following **Step 1** in Section 2.2, we know how to form a generated instrument given a single instrument $A_k(Z_i)$. With t missing instruments, the generated instrument set $\mathcal{Z}_i$ forms as

$$\begin{aligned}
\mathcal{Z}_i &\equiv (\mathcal{Z}_{i1}, ..., \mathcal{Z}_{it})' \\
\mathcal{Z}_{ik} &= \frac{1-D_i}{1-p(W_i)} A_k(Z_i) + \frac{D_i - p(W_i)}{1-p(W_i)} \mathbb{E}(A_k(Z_i)|W_i), \quad k = 1, ..., t.
\end{aligned}$$

The nuisance parameters $p(W_i)$ and $h(W_i) \equiv \mathbb{E}\left((A_1(Z_i), ..., A_t(Z_i))'|W_i\right)$ can be estimated either parametrically or nonparametrically. For example, we can still conduct sieve least square estimation. The estimates for the nuisance functions are denoted by $\widehat{p}(W_i)$ and $\widehat{h}(W_i)$, respectively. The estimated generated instrument set $\widehat{\mathcal{Z}}_i$ is obtained by plugging in the nuisance parameter estimates:

$$\widehat{\mathcal{Z}}_i = \frac{1-D_i}{1-\widehat{p}(W_i)} A(Z_i) + \frac{D_i - \widehat{p}(W_i)}{1-\widehat{p}(W_i)} \widehat{h}(W_i) \tag{2.9}$$

Note that the number of nuisance parameters is $t+1$, including the propensity score $p(W_i)$, and t conditional expectations $\mathbb{E}(A_{ik}|W_i)$, $k = 1, ..., t$.

Step 2 Penalized Reduced Form Estimation

Armed with the estimated generated instrument set $\widehat{\mathcal{Z}}_i$, we are able to implement the reduced form regression specified in (2.8). The penalized reduced form estimator $\widehat{\tau}_{pen}$ is defined as the minimizer of the optimization program

$$\widehat{\tau}_{pen} = \arg\min_{\tau \in \mathbb{R}^t} \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \widehat{\mathcal{Z}}_i' \tau\right)^2 + \frac{\lambda}{n} \sum_{j=1}^{t} |\zeta_j \tau_j| \tag{2.10}$$

where λ is the penalty term, and $\zeta = (\zeta_1, ..., \zeta_t)'$ is a vector specifying penalty loadings. Suppose $\{\phi_n\}$ is a sequence satisfying $\phi_n = o_p(1)$ and $\log(1/\phi_n) = O_p(\log(p \vee n))$. Then

$$\lambda \equiv c\sqrt{n}\Phi^{-1}(1 - \phi_n/(2t)) \tag{2.11}$$

$$\zeta_j \equiv \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathcal{Z}_{ij}^2 \nu_i^2)} \quad j = 1, ..., t. \tag{2.12}$$

In practice, we use estimates of the penalty loadings $\widehat{\zeta}$. Refer to the appendix for algorithm obtaining $\widehat{\zeta}$. Sometimes, when the true value of τ_0 is close to zero, the optimization procedure in (2.10) may fail to choose any generated instrument. There are two reasons for this. First, the regularization by the l_1-norm employed in (2.10) shrinks the estimated coefficients toward zero. This is the so called "shrinkage bias" in the Lasso literature. Second, there is another source of estimation bias brought by the estimated generated instruments. I call this bias the "generated bias", which is a common drawback shared by estimation based on generated regressors, see e.g. Mammen et al. (2012), Hahn and Ridder (2013).

Aiming to remove some of the bias, I propose the following procedure for the estimation of τ and regard (2.10) merely as a model selecting procedure. I denote this estimator by $\widehat{\tau}$. Define $\widehat{T}_{pen}$ to be the support of $\widehat{\tau}_{pen}$,

$$\widehat{T}_{pen} = \text{support}(\widehat{\tau}_{pen}) = \{j \in (1, ..., t) : |\widehat{\tau}_{pen\ j}| > 0\}$$

The support for $\widehat{\tau}$, denoted by $\widehat{T}$, on the other hand, can contain other variables of interest. For example, even if some generated instruments are not chosen, one can still add them back to $\widehat{T}$ out of empirical relavance. One can also add other non-missing instruments to $\widehat{T}$. As a result, the reduced form estimator $\widehat{\tau}$ is defined as the minimizer of the optimization program

$$\widehat{\tau} = \arg \min_{\tau \in \mathbb{R}^t : \tau_{\widehat{T}^c} = 0} \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \widehat{\mathcal{Z}}_i' \tau \right)^2$$

The advantage of this penalized reduced form estimation is that it allows the identities of the relevant *full data* instruments to be unknown. And it builds on the fact that if a single missing instrument $A_k(Z_i)$ is a valid *full data* instrument, then $\mathcal{Z}_{ik}$ would be a valid generated instrument. By selecting among generated instruments, we can then get right choice of the full data instruments. In other words, we are indirectly approximating the full data optimal instrument $\mathcal{D}(Z_i)$ through the pseudo-approximation of $\mathcal{A}_i(\cdot; p, h)$.

A key condition that guarantees the effective use of this large generated instrument set is sparsity, which says that the observed data optimal instrument $\mathcal{A}_i(\cdot; p, h)$ can be well approximated by a small number of unknown generated instruments. The sparsity is summarized in the following condition.

Condition 1. *(Sparsity) The infeasible $\mathcal{A}_i \equiv \mathcal{A}(\mathcal{D}(Z_i), Z_i, W_i, D_i; p.h)$ is well approximated by a function of unknown $s \geq 1$ generated instruments:*

$$\mathcal{A}_i = \mathcal{Z}_i' \tau + a_i, \qquad \sum_{j=1}^{t} 1\{\tau_j \neq 0\} \leq s = o(n)$$

and $\sqrt{\frac{1}{n} \sum_{i=1}^{n} a_i^2} = O_p(\sqrt{s/n})$.

Condition 1 requires that there are at most s generated instruments to approximate $\mathcal{A}_i(\cdot; p, h)$ up to approximation error a_i, chosen to be no larger than the conjectured size $\sqrt{s/n}$ of the estimation error of the infeasible estimator that knows the identity of the true generated instruments, the "oracle estimator".

However, the true generated instrument set $\mathcal{Z}_i$ is infeasible, and we use estimated generated instrument set $\widehat{\mathcal{Z}}_i$ in the penalized reduced form regression. The procedure implicitly requires that the nuisance parameters are well estimated to get a good approximation of $\mathcal{Z}_i$. Denote the fitted value obtained in this step as

$$\widehat{\mathcal{A}}_i \equiv \widehat{\mathcal{A}}_i(\cdot; \hat{p}, \hat{h}) = \widehat{\mathcal{Z}}_i' \widehat{\tau}$$

Step 3 Structural estimation

The structural estimation is similar to that in the single missing instrument case. I define the Pen-Gen-IV estimator $\widehat{\theta}_{PenGenIV}$ as follows

$$\widehat{\theta}_{PenGenIV} = \left(\sum_{i=1}^{n} \widehat{\mathcal{A}}_i X_i^{*'} \right)^{-1} \sum_{i=1}^{n} \widehat{\mathcal{A}}_i Y_i$$

I leave the consistency and asymptotic normality results of $\widehat{\theta}_{PenGenIV}$ to the next section.

2.4 Asymptotic Results

In this section, I present large sample properties of Pen-Gen-IV estimator.

2.4.1 Large Sample Properties of Pen-Gen-IV Estimator

The following assumptions are needed for the asymptotic normality of the Pen-Gen-IV estimator.

Assumption 6 (Condition Sparse Eigenvalues)**.** *Let $\widehat{M}$ be the estimated empirical Gram matrix, $\widehat{M} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{Z}}_i \widehat{\mathcal{Z}}_i'$. We define the minimal and maximal m-sparse eigenvalues of $\widehat{M}$ as follows:*

$$\phi_{min}(m)(\widehat{M}) = \min_{\xi \in \Psi(m)} \xi^{'} \widehat{M} \xi \quad and \quad \phi_{max}(m)(\widehat{M}) = \max_{\xi \in \Psi(m)} \xi^{'} \widehat{M} \xi$$

where $\Psi(m) = \{\xi \in \mathbb{R}^t : \sum_{i=1}^{t} \mathbb{1}\{\xi_i \neq 0\} \leq m, ||\xi||_2 = 1\}$.

For any constant C, there exist $0 < \kappa_1 < \kappa_2 < \infty$, κ_1 and κ_2 do not depend on n but may depend on C, s.t. with probability approaching 1, as $n \to \infty$,

$$\kappa_1 \leq \phi_{min}(Cs)(\widehat{M}) \leq \phi_{max}(Cs)(\widehat{M}) \leq \kappa_2$$

Assumption 6 imposes restrictions on the estimated Gram matrix. It requires that there exists some $m \times m$ submatrix of the large $t \times t$ matrix is bounded. The following assumption is about several moment restrictions of the reduced error ν_i and generated instrument $\mathcal{Z}_i$.

Assumption 7. *Recall the reduced form $X_i = \mathcal{A}_i + \nu_i$, the following hold*

(7.1) $\mathbb{E}(\nu_i^4) < \infty$, $||\mathcal{Z}_i||_\infty \leq K_n$ *a.s., where* $K_n^2 s \log^2(n) \log^2(s \log n) log(t \vee n) = o(n)$.

(7.2) $\max_{j \leq t} \mathbb{E}|\mathcal{Z}_{ij}^3 \nu_i^3| = O(K_n^{\dagger})$, *where* ${K_n^{\dagger}}^2 \log^3(t \vee n) = o(n)$.

Before presenting the next assumption, we need to introduce some new notation. Let $\widehat{\tau}^{\star}$ be the estimator obtained through using the true generated instrument $\mathcal{Z}_i$ instead of the estimated one $\widehat{\mathcal{Z}}_i$, i.e.

$$\widehat{\tau}^{\star} = \arg\min_{\tau \in \mathbb{R}^t} \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \mathcal{Z}_i' \tau \right)^2 + \frac{\lambda}{n} \sum_{j=1}^{t} |\zeta_j \tau_j|$$

and $\widehat{\mathcal{A}}_i^{\star} \equiv \mathcal{Z}_i' \widehat{\tau}^{\star}$.

Assumption 8. *The following hold*

(8.1) $\sqrt{\frac{1}{n} \sum_{i=1}^{n} (\widehat{\mathcal{A}}_i^{\star} - \widehat{\mathcal{A}}_i)^2} = O_p(\sqrt{\frac{s \log(t \vee n)}{n}})$

(8.2) $||\widehat{\tau}^{\star} - \widehat{\tau}||_2 = o_p(\sqrt{\frac{s^2 \log(t \vee n)}{n}})$.

Assumption 9. *Let Ξ_0 be a diagonal matrix of ideal penalty loadings. $\Xi_0 \equiv diag(\zeta_1 ..., \zeta_t)$. Let $\widehat{\Xi}$ be the empirical loadings. $\widehat{\Xi}$ is asymptotically valid in the sense that*

$$l \Xi_0 \leq \widehat{\Xi} \leq u \Xi_0$$

for $u > l \geq 0$.

Note that the penalty loadings we use for estimation will satisfy Assumption 9. We also impose the following moment restrictions on the *full data* instruments Z_i, the structural error ϵ_i, the generated instrument $\mathcal{Z}_i$ and the size of the instrument set t.

Assumption 10. *The following hold*

(10.1) *The eigenvalues of* $\tilde{Q} = \mathbb{E}(\mathcal{D}(Z_i)\mathcal{D}(Z_i)')$ *are bounded uniformly from above and away from zero, uniformly in* n*.*

(10.2) *The conditional variance* $\mathbb{E}(\epsilon_i^2|Z_i)$ *is bounded uniformly from above and away from zero, uniformly in* i *and* n*.*

(10.3) *Normalize the instruments so that* $\mathbb{E}(\widehat{\mathcal{Z}}_i^2\epsilon_i^2) = 1$ *for each* $1 \leq j \leq t$ *and for all* n*, for some* $p_1 > 2$*,* $p_2 > 2$*, uniformly in* n*,*

$$\max_{1\leq j\leq t} \mathbb{E}(|\widehat{\mathcal{Z}}_{ij}\epsilon_i|^3)+\mathbb{E}(||\mathcal{D}_i||_2^{p_1}|\epsilon_i|^{2p_1})+\mathbb{E}(||\mathcal{D}_i||_2^{p_1})+\mathbb{E}(|\epsilon_i|^{p_2})+\mathbb{E}(||X_i||_2^{p_1}) = O_p(1)$$

(10.4) *The following growth conditions hold: (a)* $\frac{s\log(t\vee n)}{n}n^{\frac{2}{p_2}} \to 0$*, (b)* $\frac{s^2\log^2(t\vee n)}{n} \to 0$*, (c)* $\max_{1\leq j\leq t}\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathcal{Z}}_i^2\epsilon_i^2) = o_p(1)$*, (d)* $\log^3 t = o(n)$*.*

Assumption (10.1) is a generalization of Assumption (3.1). It ensures that the identification is strong. Assumption (10.2) requires that the structural errors are bounded heteroscedastically. Assumption (10.3) imposes several moment restrictions. Assumption (10.4) guarantees that the impact of the estimation of generated instrument on the IV estimator is asymptotically negligible.

Theorem 4. *Under Assumptions 1-10, the Pen-Gen-IV estimator* $\sqrt{n}(\widehat{\theta}_{PenGenIV}-\theta_0) \Rightarrow N(0, \tilde{V}_0)$*, with*

$$\tilde{V}_0 = (\tilde{Q}'\Omega^{-1}_{PenGenIV}\tilde{Q})^{-1}$$

where $\tilde{Q} = \mathbb{E}(\mathcal{D}(Z_i)\mathcal{D}(Z_i)')$ *and*

$$\Omega_{PenGenIV}$$

$$= \mathbb{E}\left(\frac{1}{1-p(W_i)}\mathbb{E}(\mathcal{D}(Z_i)\mathcal{D}(Z_i)'|W_i)\epsilon_i^2 - \frac{p(W_i)}{1-p(W_i)}\mathbb{E}(\mathcal{D}(Z_i)|W_i)\mathbb{E}(\mathcal{D}(Z_i)|W_i)'\epsilon_i^2\right)$$

2.4.2 Efficiency

In this section, I establish two sets of efficiency results. The first set of results contains semiparametric efficiency bounds for the estimation of θ defined by moment condition (2.2), under Assumption 1 and Assumption 2. I also provide an analytical solution to the optimal instruments and discuss its relationship to the *full data* optimal instruments under (2.2). The second set of results include comparison between the asymptotic variance of the estimator and the semiparametric efficiency bounds. I state the conditions under which the Pen-Gen-IV estimator attains the efficiency bounds.

2.4.2.1 Efficiency Bounds and Efficiency of Pen-Gen-IV Estimator

The semiparametric efficiency bounds for the estimation of θ under the conditional independence assumption is provided in the following theorem.

Theorem 5. *Let θ be defined by moment condition (2.2). Under Assumption 1 and Assumption 2, the asymptotic variance lower bound for all RAL estimator of θ is*

$$\left(\mathbb{E}(A^*(Z_i)\mathcal{D}(Z_i)')' \left(\mathbb{E}(S_{eff}S_{eff}^{'})\right)' \mathbb{E}(A^*(Z_i)\mathcal{D}(Z_i)')\right)^{-1}$$

where $\mathcal{D}(Z_i) = \mathbb{E}(X_i|Z_i)$, *and*

$$S_{\textit{eff}} = \frac{1-D_i}{1-p(W_i)}A^*(Z_i)\epsilon_i + \frac{D_i - p(W_i)}{1-p(W_i)}\mathbb{E}(A^*(Z_i)|W_i)\epsilon_i \tag{2.13}$$

where $S_{\textit{eff}}$ *is the efficient score vector.* $A^*(Z_i)$ *is the unique solution to*

$$\mathbb{E}\left[\frac{1}{1-p(W_i)}A(Z_i)\epsilon_i^2 - \frac{p(W_i)}{1-p(W_i)}\mathbb{E}(A(Z_i)|W_i)\epsilon_i^2|Z_i\right] = \mathcal{D}(Z_i) \tag{2.14}$$

I call $A^*(Z_i)$ the "star-optimal" instrument. Note that $A^*(Z_i)$ is a full data instrument. It might be different from the *full data* optimal instrument $\mathcal{D}(Z_i)$. A special case when $A^*(Z_i)$ coincides with $\mathcal{D}(Z_i)$ is provided in the following remark.

Remark 4. If there is zero possibility of missing, i.e. $p(w) = 0$ for $\forall w \in \mathcal{W}$, and assuming conditional homoskedasticity,

$$\mathbb{E}\left[\epsilon_i^2|Z_i\right] = \Psi \tag{2.15}$$

for a constant, positive number Ψ, then (2.14) shrinks to

$$A(Z_i)\mathbb{E}\left[\epsilon_i^2|Z_i\right] = \mathcal{D}(Z_i)$$

The star-optimal instrument $A^*(Z_i)$ equals to the full data optimal instrument $\mathcal{D}(Z_i)$ up to a constant Ψ.

$$A^*(Z_i) = \mathcal{D}(Z_i)\Psi^{-1} \tag{2.16}$$

The existence and uniqueness of $A^*(Z_i)$ has implication about the efficiency of the Pen-Gen-IV estimator $\widehat{\theta}_{PenGenIV}$. Compare the efficient score vector in (2.13) with the moment condition (2.7), we will see that whether the Pen-Gen-IV estimator is efficient or not will depend on the equivalence between $A^*(Z_i)$ and $\mathcal{D}(Z_i)$. The following Corollary provides a sufficient condition under which the equivalence holds.

Corollary 3. *The Pen-Gen-IV estimator is efficient if*

$$\mathbb{E}\left(\frac{p(W_i)}{1-p(W_i)}(\mathbb{E}(\mathbb{E}(X_i|Z_i)|W_i) - X_i)\epsilon_i^2|Z_i\right) = 0 \tag{2.17}$$

then

$$A^*(Z_i) = \mathcal{D}(Z_i).$$

The efficiency of the Pen-Gen-IV estimator won't depend on the conditional homoskedasticity assumption as in (2.15). However, a direct test on the conditional moment (2.17) is infeasible, since the conditioning variable Z_i has missing values. One can seek a pseudo-approximation of the conditional expectation in the spirit of $\mathcal{A}_i$ to $\mathcal{D}_i$. But this is beyond the scope of the current paper.

2.4.2.2 The Star-optimal Instrument $A^*(Z_i)$

In this subsection, we seek to find an explicit expression for the optimal instruments. This has been done in two ways. Firstly, we can solve for an analytical solution to Equation (2.14), which is possible by utilizing theory of integral equations. The results are provided in Lemma 4. Secondly, if the conditional expectations in (2.14) are approximated by some series, e.g.

power series, splines, etc. with mild restrictions, we are able to obtain an approximation for the optimal instruments, see Lemma 5.

Assume that after adjusted by the inverse of propensity score, it still holds conditional homoskedasticity,

$$\mathbb{E}\left[\frac{\epsilon^2}{1-p(W)}|Z\right] = \widetilde{\Omega}$$

then (2.14) can be rewritten into

$$A(z) = h(z) + \widetilde{\Omega}^{-1}\int K(z,t)A(t)dt \tag{2.18}$$

where

$$h(z) = \widetilde{\Omega}^{-1}\mathcal{D}(z)', \text{ and } K(z,t) = \int \frac{p(w)}{1-p(w)}\epsilon^2 f_{W|Z}(w|z)f_{Z|W}(t|w)dw \text{ for } \forall z,t.$$

Notice that (2.18) is a Fredholm integral equation of the second kind, see e.g. Zemyan (2012). The following lemma applies the Fredholm Theorem and presents the optimal instruments explicitly.

Lemma 4 (Analytical)**.** *Suppose that the assumptions in Theorem 5 hold with the adjusted conditional homoskedasticity, then we have*

$$A^*(z) = h(z) + \widetilde{\Omega}^{-1}\int R(z,t;\widetilde{\Omega}^{-1})h(t)dt \tag{2.19}$$

The resolvent kernel

$$R(z,t;\widetilde{\Omega}^{-1}) = \sum_{m=1}^{\infty}\widetilde{\Omega}^{1-m}K_m(z,t)$$

where $K_1(z,t) = K(z,t)$, *and*

$$K_m(z,t) = \int Q(s_m) f_{W|Z}(s_1|z) f_{Z|W}(t|s_m) \prod_{i=1}^{m-1} Q(s_i) \mathbb{E}\left[f_{W|Z}(s_{i+1}|Z)|W = s_i\right] ds^m$$

when $m \geq 2$, *with* $Q(s_i) = \frac{p(s_i)}{1-p(s_i)}\epsilon^2$.

Proof. See Appendix B.0.3. □

Lemma 5 (Approximation)**.** *Suppose that all the assumptions in Lemma 4 hold. In addition, for a large* N, $K(z,t)$ *can be approximated by* $\sum_{i=1}^{N} a_i(z)b_i(t)$, *where each function* $a_i(z)$ *and* $b_i(t)$ *is continuous and the sets* $\{a_i(z)\}$, $\{b_i(t)\}$ *are linearly independent, then the optimal instruments*

$$A^*(z) = h(z) + \sum_{i=1}^{N} c_i a_i(z) \tag{2.20}$$

where $c_i = \widetilde{\Omega}^{-1} \int b_i(t)A(t)dt$.

2.5 Monte Carlo Experiments

In this section, I present the small sample behavior of the Pen-Gen-IV estimator. I consider three cases of missing many instruments. I compare the performance of the new estimator to other Lasso-type of estimators as well as to the Gen-IV estimator developed in Chapter 1.

2.5.1 Data Generating Process (DGP) with Many Missing Instruments

In this design, the simulations are based on an IV model with many missing instruments. In particular, the number of instruments can exceed the sample size.

$$\begin{aligned}
Y_i &= \alpha_0 X_i + \epsilon_i \\
X_i &= Z_i^{'}\Gamma + \upsilon_i \\
D_i &= \mathbb{1}(\sin(\varrho_0 Y_i + \varrho_1 X_i) + u_i \leq p) \\
(\epsilon_i, \upsilon_i) &\sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_{\epsilon\upsilon} \\ \rho_{\epsilon\upsilon} & 1 \end{pmatrix}\right) \\
Z_i &= (Z_{i1}, ..., Z_{i100})^{'} \sim N(\mathbf{0}, \Sigma_Z) \\
u_i &\sim Unif[0,1]
\end{aligned}$$

where $\alpha_0 = 1$ is the parameter of interest. In all the simulations, $\varrho_0 = -1$, $\varrho_1 = 1$, $\rho_{\epsilon\upsilon} = 0.6$. The variance-covariance matrix[3] Σ_Z is specified as $\mathbb{E}(Z_{ih}^2) = 1$, and $Corr(Z_{ih}, Z_{ij}) = 0.5^{|j-h|}$, for $\forall h, j \in \{1, 2, ..., 100\}$, $j \neq h$.

I use three different settings for the reduced form coefficients, Γ. I follow Belloni et al. (2012) to name the three settings "exponential", "Cut5" and "Cut20". In the first exponential design, $\Gamma = \gamma_1(1, 0.8^2, ..., 0.8^{99})^{'}$, in which the coefficients on instrument decreases exponentially. In the Cut5 and Cut20 designs, $\Gamma = \gamma_1(\iota_s, \mathbf{0}_{n-s})^{'}$, where ι_s is a $1 \times s$ vector of ones and $\mathbf{0}_{n-s}$ is a $1 \times (n-s)$ vector of zeros, for $s = 5$ and $s = 20$, respectively. All these three designs satisfy the Sparsity condition. Experiments are conducted with the specifications for

$$n \in \{100, 250\}, \quad p \in \{0.25, 0.5\}, \quad \gamma_1 \in \{1, 0.3\}$$

[3] Other variance-covariance structures are also considered. The simulation results are qualitatively similar, thus omitted. See Appendix C for reports on a polynomial-type instruments set, in which $Z_{1h} \sim N(0,1)$, and $Z_{ih} = Z_{i1}^h$, for $h \in \{2, 3, ...100\}$.

similar to the single missing instrument case.

2.5.2 Results

Table 2.1 compares performance of three estimators. The Full data Lasso and Complete Case Lasso both utilize methods in Belloni et al. (2012) for selecting among instruments. This table reveals that the Pen-Gen-IV estimator has significant smaller bias than Complete Case Lasso in most of the cases. When the instrument set is relatively strong $\gamma_1 = 1$, the summary statistics of Pen-Gen-IV are very similar to Full data Lasso.

Finally, I conduct a comparison between the two estimators developed in this paper in Table 2.2. The Gen-IV (Full) uses the full instrument set and forms a generated instrument set with 100 IVs. In contrast, Gen-IV (Single) adopts the first *full data* instrument Z_{i1}; Gen-IV(5) adopts $(Z_{i1}, ..., Z_{i5})'$; Gen-IV(5)-mis uses $(Z_{i16}, ..., Z_{i20})'$. Note that in Cut5 design, $(Z_{i1}, ..., Z_{i5})'$ is the true instrument set. And Gen-IV(5) is based on a correct specification on instruments. It is not surprising that Gen-IV(5) outperforms other estimators, though Pen-Gen-IV is still the second best. Similar fashion is found about Gen-IV (5) in Exponential and Cut20. This is because in these two cases, the first five *full data* instruments are still the most relevant ones. In the remaining cases, Pen-Gen-IV estimator is superior to the others, which is evidence that the shrinkage-based methods have good performance in selecting among generated instruments.

Table 2.1: Summary Statistics for Many Missing instruments

	Exponential			$S = 5$			$S = 20$		
	MB	MAD	RMSE	MB	MAD	RMSE	MB	MAD	RMSE
Sample A: n=100, $p = 0.25$,$\gamma_1 = 1$									
Full data Lasso	0.0105	0.0265	0.0380	0.0015	0.0186	0.0282	0.0017	0.0087	0.0139
Complete Case Lasso	0.0285	0.0381	0.0557	0.0115	0.0256	0.0391	0.0021	0.0145	0.0220
Pen-Gen-IV	0.0177	0.0275	0.0442	0.0076	0.0223	0.0325	0.0052	0.0097	0.0152
Sample B: n=100, $p = 0.25$,$\gamma_1 = 0.3$									
Full data Lasso	0.0783	0.0892	0.3373	0.0389	0.0599	0.0996	0.0303	0.0354	0.0522
Complete Case Lasso	0.4260	0.4260	1.1891	0.2516	0.2516	0.2984	0.0870	0.0870	0.0953
Pen-Gen-IV	0.3859	0.3859	0.4544	0.2570	0.2570	0.2832	0.0827	0.0827	0.1205
Sample C: n=100, $p = 0.5$,$\gamma_1 = 1$									
Full data Lasso	0.0192	0.0222	0.0319	0.0175	0.0227	0.0268	0.0055	0.0122	0.0138
Complete Case Lasso	0.0673	0.0673	0.0914	0.0329	0.0329	0.0697	0.0087	0.0120	0.0182
Pen-Gen-IV	0.0513	0.0513	0.0529	0.0182	0.0265	0.0346	0.0121	0.0153	0.0191
Sample D: n=100, $p = 0.5$,$\gamma_1 = 0.3$									
Full data Lasso	0.1728	0.1952	0.2412	0.1489	0.1489	0.1511	0.0644	0.0645	0.0739
Complete Case Lasso	0.4989	0.4989	0.8074	0.4142	0.4145	0.3906	0.1430	0.2038	0.3340
Pen-Gen-IV	0.3093	0.4394	0.4469	0.3778	0.3778	0.3914	0.1305	0.1305	0.1353
Sample E: n=250, $p = 0.25$,$\gamma_1 = 1$									
Full data Lasso	0.0105	0.0192	0.0258	0.0018	0.0127	0.0186	-0.0014	0.0063	0.0084
Complete Case Lasso	0.0413	0.0413	0.0522	0.0265	0.0265	0.0346	0.0041	0.0073	0.0156
Pen-Gen-IV	0.0159	0.0226	0.0337	0.0084	0.0159	0.0224	0.0025	0.0062	0.0155
Sample F: n=250, $p = 0.25$,$\gamma_1 = 0.3$									
Full data Lasso	0.0416	0.0841	0.1017	0.0222	0.0571	0.0685	0.0088	0.0147	0.0334
Complete Case Lasso	0.4009	0.4009	0.4024	0.2277	0.2277	0.2415	0.0548	0.0548	0.0726
Pen-Gen-IV	0.2623	0.3564	0.9499	0.1074	0.1074	0.1301	0.0598	0.0598	0.0769
Sample G: n=250, $p = 0.5$,$\gamma_1 = 1$									
Full data Lasso	-0.0048	0.0128	0.0129	-0.0072	0.0119	0.0129	-0.0000	0.0038	0.0060
Complete Case Lasso	0.0595	0.0595	0.0601	0.0293	0.0293	0.0313	0.0115	0.0132	0.0123
Pen-Gen-IV	0.0232	0.0345	0.0398	0.0052	0.0120	0.0145	-0.0006	0.0052	0.0273
Sample H: n=250, $p = 0.5$,$\gamma_1 = 0.3$									
Full data Lasso	0.0397	0.0581	0.1080	0.0220	0.0318	0.0878	0.0260	0.0260	0.0294
Complete Case Lasso	0.4732	0.4732	0.4937	0.3177	0.3177	0.3003	0.0695	0.0695	0.0784
Pen-Gen-IV	0.3166	0.3166	0.3396	0.1632	0.1632	0.2156	0.0632	0.0632	0.0642

2.6 Conclusion

I provide Lasso-type of procedure for reduced form regression with many missing instruments. After generating a rich, full data instrument set, these many generated instruments are used to approximate the infeasible full data optimal instruments, via a so called “pseudo” approximation.

This chapter states clearly that if the *full data* instrument satisfies

Table 2.2: Comparison of Gen-IV and Pen-Gen-IV

	Exponential			$S = 5$			$S = 20$		
	MB	MAD	RMSE	MB	MAD	RMSE	MB	MAD	RMSE
Sample A: n=100, $p = 0.25$,$\gamma_1 = 1$									
Gen-IV (Full)	0.0804	0.0804	0.0842	0.0525	0.0525	0.0563	0.0077	0.0080	0.0124
Gen-IV (Single)	-0.0511	0.0609	0.0734	0.0223	0.0633	0.0971	0.0019	0.0420	0.1480
Gen-IV (5)	0.0197	0.0337	0.0403	0.0236	0.0297	0.0331	0.0098	0.0207	0.0339
Gen-IV (5)-mis	0.0979	0.1081	0.3403	0.0972	0.1399	0.1737	0.0177	0.0340	0.0452
Pen-Gen-IV	0.0598	0.0598	0.0628	0.0361	0.0361	0.0406	0.0061	0.0103	0.0193
Sample B: n=100, $p = 0.25$,$\gamma_1 = 0.3$									
Gen-IV (Full)	0.3700	0.3700	0.3777	0.3068	0.3068	0.3017	0.0915	0.0915	0.1081
Gen-IV (Single)	0.0842	0.2291	0.6007	-0.0366	0.2905	0.5472	0.0230	0.1980	0.5209
Gen-IV (5)	0.1329	0.1471	0.2220	0.0772	0.1243	0.1451	-0.0408	0.0541	0.0845
Gen-IV (5)-mis	0.5638	0.5841	0.6290	0.4360	0.5452	0.5837	0.0623	0.0786	0.1095
Pen-Gen-IV	0.3683	0.3959	0.4239	0.2706	0.2706	0.2882	0.0860	0.0860	0.0920
Sample C: n=100, $p = 0.5$,$\gamma_1 = 1$									
Gen-IV (Full)	0.0530	0.0530	0.0708	0.0295	0.0295	0.0450	0.0081	0.0127	0.0165
Gen-IV (Single)	0.0421	0.0847	0.1393	0.0229	0.1095	0.1508	0.0387	0.0903	0.1915
Gen-IV (5)	0.0101	0.0280	0.0488	-0.0048	0.0255	0.0357	0.0202	0.0460	0.0483
Gen-IV (5)-mis	0.0226	0.0982	0.2452	0.0746	0.1152	0.1569	-0.0019	0.0218	0.0276
Pen-Gen-IV	0.0364	0.0416	0.0611	0.0185	0.0261	0.0424	0.0068	0.0130	0.0170
Sample D: n=250, $p = 0.5$,$\gamma_1 = 1$									
Gen-IV (Full)	0.0599	0.0599	0.0684	0.0395	0.0395	0.0444	0.0058	0.0081	0.0112
Gen-IV (Single)	0.0017	0.0324	0.0556	0.0128	0.0376	0.0540	-0.0288	0.0550	0.1837
Gen-IV (5)	-0.0011	0.0256	0.0328	0.0009	0.0239	0.0263	0.0005	0.0172	0.0296
Gen-IV (5)-mis	0.1610	0.1733	0.3389	0.1293	0.1293	0.2865	-0.0076	0.0152	0.0253
Pen-Gen-IV	0.0230	0.0361	0.0391	0.0024	0.0219	0.0284	0.0010	0.0058	0.0255
Sample E: n=250, $p = 0.5$,$\gamma_1 = 0.3$									
Gen-IV (Full)	0.3543	0.3543	0.3647	0.2790	0.2790	0.2848	0.0903	0.0903	0.0945
Gen-IV (Single)	-0.0101	0.1210	0.3325	-0.0510	0.2175	0.3601	0.0298	0.1345	0.5228
Gen-IV (5)	0.1075	0.1215	0.1768	0.0565	0.0869	0.1077	0.0360	0.0699	0.1049
Gen-IV (mis)	0.6230	0.6230	0.6886	0.5866	0.5866	0.6387	0.0394	0.0777	0.1098
Pen-Gen-IV	0.3267	0.3267	0.4407	0.1244	0.1372	0.1843	0.0594	0.0593	0.0696

the more restricted conditional independence assumption, a more efficient IV estimation with generated instruments is available. Namely, we can then adopt a flexible generated instrument set, the number of instruments in which can be very large, and can be increasing with the sample size. The shrinkage-based methods I develop can select among (very) many generated instruments to avoid overfitting, and at the same time controlling for the "generated bias".

Chapter 3

Estimation of Heterogeneous Individual Treatment Effects with Endogenous Treatments

3.1 Introduction

Nonseparable triangular models have been studied extensively in the recent econometric literature, thereby allowing researchers to understand the nature of instrumental variables in the presence of endogeneity. See e.g. Chesher (2003, 2005) and Imbens and Newey (2009). One appealing feature of nonseparable models is that the non-additive error in the causal relationship implies that the *ceteris paribus* effects of covariates on the outcome variable "vary across individuals that, measured by covariates, are identical," Chesher (2003). Such heterogeneous causal effects are referred as *"individual treatment effects"*(ITE) in the literature. See e.g. Rubin (1974), Heckman et al. (1997) and Heckman and Vytlacil (2005).

Estimating ITE and its distribution is crucial for evaluating a social program, especially in view of the political issues associated with it (see Heckman et al., 1997). From an individual's perspective, however, her ITE is more helpful for evaluating her treatment participation decision than an average effect. While the "average person" may benefit from a particular treatment, some individuals may experience little benefit or even some loss from participating, in which case alternative treatment options may be preferred. Indeed, while the individual treatment effects of 401(k) retirement programs on personal savings are mostly positive in our sample, our empirical analysis indicates that there are individuals who experience negative benefits from participating to 401(k) retirement programs.

In this paper, we consider a triangular model with a binary endogenous regressor. Because of the self–selection issue, individuals who are treated are different from those who choose not to be treated. We address this issue with a binary valued instrumental variable (see e.g. Imbens and Angrist, 1994). Limited variations of instrumental variables have been emphasized in the recent treatment effect literature. Moreover, natural experiments (e.g. Angrist and Evans, 1998; Post et al., 2008) and eligibility for treatment participation (e.g. Angrist, 1990; Abadie, 2003) provide commonly used binary–valued instrumental variables.

The distribution of heterogeneous treatment effects has also been studied using quantiles. For instance, Abadie et al. (2002) and Froelich and Melly (2013) estimate the quantile treatment effects (QTE) for the complier group, a subpopulation defined by Imbens and Angrist (1994) under binary–valued instruments. For the population QTE, Chernozhukov and Hansen (2004) propose a GMM–type approach in a linear quantile specification. Subsequently, Chernozhukov and Hansen (2006, 2008) generalize Chernozhukov and Hansen (2004)'s estimation procedure by using quantile regression methods. In a fully nonparametric setting, Horowitz and Lee (2007) and Gagliardini and Scaillet (2012) modify Chernozhukov and Hansen (2004)'s moment conditions using the Tikhonov regularization to deal with the ill–posed inverse problem for deriving asymptotic properties of their estimators.

Our approach is novel and simple to implement. Instead of solving the moment conditions in Chernozhukov and Hansen (2005), we use the quantile invariance condition to match the realized outcome with its counterfactual outcome for every observational unit in the sample through a so-called counterfactual mapping. Specifically, our approach recovers the ITE for every individual in the sample and does not suffer from the ill–posed inverse problem associated with inverting a non–linear functional. In particular, we show that the ITEs are estimated uniformly at the parametric rate. Given the recovered ITEs, we estimate the density by kernel methods and establish its asymptotic properties. Though it might be possible to obtain a density estimate from QTE estimates, this would involve a more complicated two–stage procedure and a delicate trimming scheme (see e.g. Marmer and Shneyerov, 2012).

We apply our approach to study the effects of 401(k) retirement programs on personal savings. Introduced in the early 1980s, the 401(k) retirement programs aim to increase savings for retirement. Endogeneity arises as individuals with a higher preference for savings are more likely to participate and also have higher savings than those with lower preferences (see, e.g., Poterba et al., 1996). Following e.g. Abadie (2003) and Chernozhukov and Hansen (2004), we use 401(k) eligibility as an instrumental variable for 401(k) participation. We estimate the ITEs for every individual in the sample as well as its density. Our results show that there exists a small but statistically significant proportion (about 8.77%) of individuals who experience negative effects, although the majority of ITEs is positive. It has been argued in the literature that some individuals could suffer from the program due to the *Crowding Out Effect*. We offer a complementary explanation as individuals with negative ITEs are more likely to be younger, single, from smaller and lower income families but with higher family net financial assets than the rest of the sample.

The structure of the paper is organized as follows. In Section 2, we introduce the triangular model and discuss its identification and estimation. Section 3 provides Monte Carlo experiments to illustrate the performance of our proposed estimator. Section 4 derives its asymptotic properties. Section 5 applies our estimation method to assess the effects of 401(k) retirement programs on personal savings. Proofs of our results are collected in the Appendix.

3.2 Model, Identification and Estimation

3.2.1 The triangular model

Following Chesher (2005), we consider a nonseparable triangular model with an outcome equation and a selection equation:

$$Y = h(D, X, \epsilon), \tag{3.1}$$

$$D = \mathbb{1}\{\nu \leq m(X, Z)\}. \tag{3.2}$$

Here $Y \in \mathbb{R}$ is the outcome variable, $D \in \{0, 1\}$ is an endogenous dummy that indicates the treatment status, $X \in \mathbb{S}_X \subseteq \mathbb{R}^k$ is a vector of observed covariates (not necessary exogenous) and $Z \in \{0, 1\}$ is a binary instrumental variable for D, i.e., $Z \perp (\epsilon, \nu) | X$. The two latent random variables ϵ and ν are scalar valued disturbances. Moreover, the function h and m are unknown structural relationships. In particular, h is continuous and strictly increasing in ϵ.

The key feature in the above triangular model is the nonseparability of h in the error term ϵ. With a nonseparable h, the *ceteris paribus* effects on the outcome variable from covariates "vary across individuals that, measured by covariates, are identical," Chesher (2003). In the treatment effect literature, such heterogeneous causal effects are referred as *"individual treatment effects"*(ITE), i.e.,

$$\Delta \equiv h(1, X, \epsilon) - h(0, X, \epsilon).$$

See e.g. Rubin (1974) and Heckman et al. (1997). After controlling for X, the ITE Δ is still a random object since it depends on the latent variable ϵ. Our interest is to recover the ITE for each individual from her observables (Y, D, X), and to estimate the probability density function of ITE in the population. In particular, a decision-maker can use the former to evaluate an individual's participation choice, while the latter characterizes the distribution of treatment effects, which has been central in the program evaluation literature (see e.g. Heckman et al., 1997).

We now provide two examples to illustrate the nonseparability of the structural relationship h.

Example 2.1 (Additive error with generalized heteroscedasticity): Let

$$Y = h^*(D, X) + \sigma^*(D, X) \cdot \epsilon,$$

where h^* is a real-valued function, σ^* is a positive function that captures the heteroscedasticity in the disturbance, and $\epsilon \in \mathbb{R}$ has zero mean and unit variance, unconditionally. This model is a generalization of a nonparametric regression model with heteroskedastic errors studied by e.g. Andrews (1991). The difference is that the heteroscedasticity term σ^* depends on the endogenous binary variable D. In particular, when σ^* is a constant, the above specification becomes an additive nonparametric regression with some endogenous regressor as studied by e.g. Newey and Powell (2003) and Darolles et al. (2011).

Example 2.2 (Semiparametric transformation model): Consider

$$\Gamma(Y) = X'\beta + \gamma D + \epsilon,$$

where $(\beta', \gamma)' \in \mathbb{R}^{k+1}$ and $\Gamma : \mathbb{R} \to \mathbb{R}$ is an unknown monotone function. See Horowitz (1996) when (X, D) is exogenous. A parametric example of the monotone function Γ is the Box–Cox transformation when Y is positive:

$$\Gamma(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0, \end{cases}$$

where $\lambda \in \mathbb{R}$ is a model parameter. Such a transformation is useful when the dependent variable has a limited support. Indeed, the transformed dependent variable can have an unlimited support thereby ensuring a linear model specification with its usual assumptions. Various extensions of the Box–Cox transformation have been developed in the literature (see e.g. Sakia, 1992), where monotonicity is a common feature in all these transformations. Recently, Chiappori et al. (2015) have studied the case where some variables such as D is endogenous.

3.2.2 Identification

Vuong and Xu (forthcoming) establish identification of the triangular model (1)-(2) in a constructive way and show that it only requires binary variations of the instrumental variable Z. Given the monotonicity of h, the ITE can be written as a function of the observables (Y, D, X):

$$\Delta = D \times (Y - \phi_{0X}(Y)) + (1 - D) \times (\phi_{1X}(Y) - Y), \tag{3.3}$$

where $\phi_{dX}(\cdot)$ for $d = 0, 1$ are defined as the counterfactual mappings that depend on covariates X and the value of d, namely,[1]

$$\phi_{0X}(y) = h(0, X, h^{-1}(1, X, y)), \quad \forall\ y \in \mathscr{S}_{h(1,X,\epsilon)|X},$$

$$\phi_{1X}(y) = h(1, X, h^{-1}(0, X, y)), \quad \forall\ y \in \mathscr{S}_{h(0,X,\epsilon)|X}.$$

By definition, ϕ_{dX} are monotone functions mapping $\mathscr{S}_{h(d',X,\epsilon)|X}$ onto $\mathscr{S}_{h(d,X,\epsilon)|X}$, where $d' = 1 - d$, and we have $\phi_{0X} = \phi_{1X}^{-1}$.

To obtain the ITE for an individual with $(Y, D, X) = (y, d, x) \in \mathscr{S}_{YDX}$, it suffices to identify the counterfactual mapping $\phi_{d'x}(y)$, where $d' = 1 - d$.

[1] The function $h^{-1}(d, x, \cdot)$ denotes the inverse of $h(d, x, \cdot)$. Hereafter, for a generic random variable W with distribution F_W, we denote its support by $\mathscr{S}_W$, defined as the closure of the open set $\mathscr{S}_W^o \equiv \{w : F_W(w)$ is strictly increasing in a neighborhood of $w\}$.

Let $p(x,z) = \Pr(D = 1|X = x, Z = z)$ be the propensity score function. For expositional simplicity, suppose $\mathscr{S}_{XZ} = \mathscr{S}_X \times \{0,1\}$ and $p(x,0) \neq p(x,1)$ for all $x \in \mathscr{S}_X$. W.l.o.g., throughout we assume $p(x,0) < p(x,1)$. Moreover, for any $y \in \mathbb{R}$ and $d = 0,1$, let

$$C_{dx}(y) \equiv \frac{\Pr(Y \leq y; D = d|X = x, Z = 0) - \Pr(Y \leq y; D = d|X = x, Z = 1)}{\Pr(D = d|X = x, Z = 0) - \Pr(D = d|X = x, Z = 1)}. \tag{3.4}$$

Imbens and Rubin (1997) show that $C_{dx}(\cdot)$ is the conditional distribution function of $h(d, X, \epsilon)$ given the complier group, namely, $\{X = x, m(x,0) < \nu \leq m(x,1)\}$. Let $\mathscr{C}_{dx}$ be the support of $C_{dx}(\cdot)$. It is straightforward to see that $\mathscr{C}_{dx} \subseteq \mathscr{S}_{h(d,X,\epsilon)|X=x}$. Next, we present the identification of ϕ_{dx} established in Vuong and Xu (forthcoming).

Theorem 6. *(Vuong and Xu, forthcoming) In the triangular model (1)-(2), suppose (i) h is continuous and strictly increasing in ϵ; (ii) Z is conditionally independent of (ϵ, ν) given X, i.e., $Z \perp (\epsilon, \nu)|X$ with $p(x,0) \neq p(x,1)$ for all $x \in \mathscr{S}_X$; (iii) conditional on X, the joint c.d.f. $F_{\epsilon\nu|X}$ is continuous; (iv) $\mathscr{C}_{dx} = \mathscr{S}_{h(d,X,\epsilon)|X=x}$ for $d = 0,1$ and $x \in \mathscr{S}_X$. Then, $\mathscr{S}_{h(d,X,\epsilon)|X=x} = \mathscr{S}_{Y|D=d,X=x}$, and the counterfactual mapping ϕ_{dx} is identified by*

$$\phi_{dx}(y) = C_{dx}^{-1}\big(C_{d'x}(y)\big), \quad \forall\ y \in \mathscr{S}_{Y|D=d',X=x}$$

where $C_{dx}(\cdot)$ is continuous on $\mathbb{R}$ and strictly increasing on $\mathscr{C}^\circ_{dx} \equiv \mathscr{S}^\circ_{Y|D=d,X=x}$ for $d = 0,1$, and $d' = 1 - d$.

In Theorem 6, condition (i) – (iii) are standard in the triangular model literature. The support condition (iv) requires that, conditional on $X = x$, the subpopulation $m(x,0) < \nu \leq m(x,1)$, i.e., the complier group introduced in Imbens and Angrist (1994), contains the same information on individual treatment effects as the whole population. It is weak as it is satisfied as soon as (ϵ, ν) has a rectangular support given X. See Vuong and Xu (forthcoming). It is testable since C_{dx} is identified by (3.4). When (iv) fails to hold, the counterfactual mappings are partially identified on intervals. It is worth pointing our that (iv) is needed for identification of ITE even if one assumes the error term ϵ was observed in the data.

With ϕ_{dx} identified, we can use (3.3) to construct the counterfactual outcome for any individual in the population from her observables (Y, D, X). Moreover, the probability distribution of ITE is also identified under the conditions in Theorem 6.

3.2.3 Estimation

We now develop nonparametric estimators of the counterfactual mappings ϕ_{dx} for $d = 0, 1$ and the probability density function f_{Δ} of ITE. On one hand, ϕ_{dx} can be used to construct the ITE for any individual in the population from her observables (Y, D, X). On the other hand, the probability density function is a convenient way to characterize the distribution of the ITE when the ITE is continuously distributed.[2] Our estimation approach is

[2] Under Condition (i)–(iii), the ITE can have a mass point when ϕ_{dx} has slope one in some intervals contained in its support, i.e., $\phi_{1x}(y) = g(x) + y$ on some $[a, b] \subseteq \mathscr{S}_{h(0,x,\epsilon)|X=x}$. Then, conditional on $X = x$, ITEs take the same value $g(x)$ for all $\epsilon \in \{e : h(0, x, e) \in [a, b]\}$. Hence, ITE has a mass point at $g(x)$. Such a case, however, can be detected given the identification of ϕ_x.

fully nonparametric. To present the basic ideas, we assume that the covariates X are discrete random variables with a finite support. Our analysis can be extended using e.g. the kernel method to the case where X are continuous at the cost of exposition.

Let $\{(Y_i, D_i, X_i', Z_i)' : i = 1, \cdots, n\}$ be an i.i.d. sample generated from the underlying structure of the triangular model. Our proposed estimation procedure takes two steps: First, for a given value of $(y, d, x) \in \mathscr{S}_{YDX}$, we estimate the counterfactual mapping $\phi_{d'x}(y)$ by a simple estimator that minimizes a convex population objective function. In the second step, we construct a pseudo sample of the counterfactual outcomes for all individuals in the sample and then nonparametrically estimate the density function f_Δ using the kernel method. We introduce some notation. Fix $x \in \mathscr{S}_X$. For simplicity, we suppress the dependence on $X = x$ in the following discussion. For each $(y_0, y_1) \in \mathbb{R}^2$ and $z \in \{0, 1\}$, let

$$\begin{aligned}\rho_0(y_0, y_1; z) = \mathbb{E}\big[|Y - y_0|(1-D)\big|X = x, Z = z\big]\\ - \mathbb{E}\big[\text{sign}(Y - y_1) \cdot D\big|X = x, Z = z\big] \cdot y_0\end{aligned}$$

$$\begin{aligned}\rho_1(y_0, y_1; z) = \mathbb{E}\left[|Y - y_1|D\big|X = x, Z = z\right]\\ - \mathbb{E}\big[\text{sign}(Y - y_0) \cdot (1-D)\big|X = x, Z = z\big] \cdot y_1.\end{aligned}$$

where $\text{sign}(u) \equiv 2 \times \mathbb{1}(u > 0) - 1$.

For $d = 0, 1$, let

$$Q_d(y_0, y_1) = (-1)^d \times \big[\rho_d(y_0, y_1; 0) - \rho_d(y_0, y_1; 1)\big]$$

be the population objective function. Such an objective function is motivated by the quantile regression method in Koenker and Bassett (1978). To see this, note that the quantile invariant condition in Chernozhukov and Hansen (2005) implies that for $(y_0, y_1) \in \mathbb{R}^2$ satisfying $y_1 = \phi_{1x}(y_0)$ (equivalently, $y_0 = \phi_{0x}(y_1)$), we have

$$\begin{aligned} &\Pr(Y \le y_1; D = 1|X = x, Z = 0) + \Pr(Y \le y_0; D = 0|X = x, Z = 0) \\ &\quad = \Pr(Y \le y_1; D = 1|X = x, Z = 1) + \Pr(Y \le y_0; D = 0|X = x, Z = 1). \end{aligned} \tag{3.5}$$

In the next lemma, we show that (3.5) is indeed the first–order condition of the population objective function $Q_0(\cdot, y_1)$, which is continuously differentiable and weakly convex on $\mathbb{R}$. We also show that $Q_0(\cdot, y_1)$ is strictly convex on $\mathscr{S}^\circ_{Y|D=0,X=x}$ and minimized uniquely on $\mathbb{R}$ at $y_0 = \phi_{0x}(y_1)$ whenever $y_1 \in \mathscr{S}^\circ_{Y|D=1,X=x}$. A similar argument also holds for the population objective function $Q_1(y_0, \cdot)$.

Lemma 6. *Suppose the conditions in Theorem 6 hold. Then, for $d = 0, 1$ and $y_d \in \mathbb{R}$, the function $Q_{d'}(y_0, y_1)$ is continuously differentiable and weakly convex in $y_{d'} \in \mathbb{R}$ where $d' = 1 - d$. Moreover, if $y_d \in \mathscr{S}^\circ_{Y|D=d,X=x}$, then $Q_{d'}(y_0, y_1)$ is strictly convex in $y_{d'} \in \mathscr{S}^\circ_{Y|D=d',X=x}$, and uniquely minimized on $\mathbb{R}$ at $\phi_{d'x}(y_d)$.*

Lemma 6 provides a basis for our nonparametric estimation of the counterfactual mappings $\phi_{0x}(\cdot)$ and $\phi_{1x}(\cdot)$. It is worth pointing out that each minimization is a one–dimensional optimization problem.

We are now ready to define our estimator. For expositional simplicity, let $\mathscr{S}_{Y|D=d,X=x}$ be a compact interval $[\underline{y}_{dx}, \overline{y}_{dx}]$. For $d = 0, 1$, $(y_0, y_1) \in \mathbb{R}^2$ and $z \in \{0, 1\}$, let $d' = 1 - d$ and

$$\begin{aligned}\hat{\rho}_d(y_0, y_1; z) &= \frac{\sum_{j=1}^n |Y_j - y_d| \times \mathbb{1}(D_j = d; X_j = x; Z_j = z)}{\sum_{j=1}^n \mathbb{1}(X_j = x; Z_j = z)} \\ &- \frac{\sum_{j=1}^n \text{sign}(Y_j - y_{d'}) \times \mathbb{1}(D_j = d'; X_j = x; Z_j = z)}{\sum_{j=1}^n \mathbb{1}(X_j = x; Z_j = z)} \times y_d.\end{aligned}$$

Moreover, let

$$\hat{\phi}_{d'x}(y_d) = \underset{y_{d'} \in [\underline{y}_{d'x}, \overline{y}_{d'x}]}{\arg\min} \ \hat{Q}_{d'}(y_0, y_1), \quad \forall \ y_d \in \mathscr{S}_{Y|D=d,X=x}.$$

where $\hat{Q}_{d'}(y_0, y_1) = (-1)^{d'} \times \big[\hat{\rho}_{d'}(y_0, y_1; 1) - \hat{\rho}_{d'}(y_0, y_1; 0)\big]$. For simplicity, we assume the support $[\underline{y}_{d'x}, \overline{y}_{d'x}]$ is known. See e.g. Guerre et al. (2000) for nonparametric estimation of the support $[\underline{y}_{d'x}, \overline{y}_{d'x}]$ if it is unknown.

Given the sample $\{(Y_i, D_i, X_i', Z_i)' : i = 1, \cdots, n\}$, we can construct the counterfactual outcome for every individual in the sample from her observables (Y_i, D_i, X_i). Namely,

$$\begin{cases} \hat{h}(0, X_i, \epsilon_i) = \hat{\phi}_{0X_i}(Y_i), & \text{if } D_i = 1; \\ \hat{h}(1, X_i, \epsilon_i) = \hat{\phi}_{1X_i}(Y_i), & \text{if } D_i = 0. \end{cases}$$

Thus, we can estimate the ITE by (3.3), i.e., for $i = 1, \cdots, n$,

$$\hat{\Delta}_i = \begin{cases} Y_i - \hat{h}(0, X_i, \epsilon_i), & \text{if } D_i = 1; \\ \hat{h}(1, X_i, \epsilon_i) - Y_i, & \text{if } D_i = 0. \end{cases} \tag{3.6}$$

In particular, we can construct a pseudo sample $\{\hat{\Delta}_i : i = 1, \cdots, n\}$ from the observed sample $\{(Y_i, D_i, X_i', Z_i)' : i = 1, \cdots, n\}$.

It is worth pointing out that the first–stage estimation is computationally simple and does not suffer from an ill–posed inverse problem (see e.g. Horowitz and Lee, 2007). In particular, to solve the one–dimensional optimization problem for each individual's counterfactual outcome, the practitioner can use a grid search algorithm that is simple but highly robust. As is shown below, the first–stage estimation bias $\hat{\phi}_{dx}(\cdot) - \phi_{dx}(\cdot)$ uniformly converges to zero at the parametric rate of $\sqrt{n}$, given that all the covariates X are discrete variables.[3]

Next, we follow Guerre et al. (2000) to estimate the density function f_Δ by the kernel method. To clarify ideas, let $[\underline{\delta}, \overline{\delta}]$ be a subinterval of the ITE's support. Then, we define the density estimator:

$$\hat{f}_\Delta(\delta) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\hat{\Delta}_i - \delta}{h}\right), \quad \forall \delta \in [\underline{\delta} + h, \overline{\delta} - h],$$

where h is a bandwidth and K is a kernel with a compact support. Because the kernel estimator $\hat{f}_\Delta$ suffers from boundary issues, then we restrict the estimation of f_Δ to the inner subset $[\underline{\delta} + h, \overline{\delta} - h]$.

3.3 Monte Carlo Experiments

To illustrate the finite sample performance of the proposed estimator, we conduct a Monte Carlo study. For simplicity, we do not include other covariates X in the specification. Following the conditions in Theorem 6, the

[3] If X_i contains continuous random variables, then we need to smooth over X_i as otherwise there may not be enough observations for which $X_j = X_i$.

data generating process is given by

$$Y = h(D, \epsilon), \qquad D = \mathbb{1}(\gamma_0 + \gamma_1 \cdot Z + \nu \geq 0),$$

where $h(d, \epsilon) = (\epsilon + 1)^{2+d}$ for $d = 0, 1,$[4] and (ϵ, ν) conforms to a joint distribution with uniform marginal distributions on $[0, 1]$ and Gaussian copula with correlation coefficient 0.3.[5] Because $h(d, \cdot)$ is continuous and strictly increasing in ϵ, Condition (i) in Theorem 6 is satisfied. We set $\gamma_0 = -0.7$ and $\gamma_1 = 0.1, 0.2$ and 0.3, respectively. The value of γ_1 determines the size of the compliers group, i.e., $-\gamma_0 - \gamma_1 \leq \nu < -\gamma_0$. Hence, the larger γ_1, the more "effective" the instrumental variable Z. In our setting, $\Delta = \epsilon(\epsilon+1)^2$ is distributed on $[0, 4]$ with mean 1.417 and median 1.125 in the population. Moreover, we set $Z = \mathbb{1}\{\xi \geq 0\}$ where $\xi \sim N(0, 1)$ is independent of (ϵ, ν). Conditions (ii)–(iv) in Theorem 6 are satisfied. In particular, condition (iv) holds since $F_{\epsilon\nu|X}$ has a rectangular support as noted in Vuong and Xu (forthcoming).

Table 3.1 reports the finite sample performance of our ITE estimates in terms of the Root Mean Squared Error (RMSE). Specifically, for each size $n = 1000, 2000, 4000$ we draw $\{(\epsilon_i, \nu_i, \xi_i) : i = 1, \cdots, n\}$ to obtain a sample (Y_i, D_i, Z_i) of size n. We then compute the true ITE Δ_i by $h(1, \epsilon_i) - h(0, \epsilon_i)$ and its estimate $\hat{\Delta}_i$ by (3.6) for each individual (Y_i, D_i, Z_i). To obtain the RMSE for each such individual's ITE, we draw another 200 samples $\{(Y_i^{(r)}, D_i^{(r)}, Z_i^{(r)}) :$

[4] We also consider other functional forms for $h(d, \cdot)$, e.g., $h(0, \epsilon) = \ln(\epsilon + 1)$ and $h(1, \epsilon) = (\epsilon + 1)^2$. The results are qualitatively similar.

[5] A copula is a multivariate probability distribution of random variables, each of which is marginally uniformly distributed on $[0, 1]$. The Gaussian copula is constructed from a multivariate normal distribution. See e.g. Nelsen (2007).

$i = 1, \cdots, n\}$ from $\{(\epsilon_i^{(r)}, \nu_i^{(r)}, \xi_i^{(r)}) : i = 1, \cdots, n\}$ for $r = 1, \cdots, 2000$. These are used to repeatedly estimate the ITEs for the individuals in the original sample by $\hat{\Delta}_i^{(r)} = [Y_i - \hat{\phi}_0^{(r)}(Y_i)]D_i + [\hat{\phi}_1^{(r)}(Y_i) - Y_i](1 - D_i)$ where $\hat{\phi}_d^{(r)}$ is the estimate of ϕ_d using the r–th new drawn sample. Thus, we obtain the RMSE of $\hat{\Delta}_i$ by $\sqrt{\frac{1}{200}\sum_{r=1}^{200}\left[\hat{\Delta}_i^{(r)} - \Delta_i\right]^2}$. For comparison, we also provide the RMSE of the LATE over the 200 replications/samples within curly brackets as proposed by Imbens and Angrist (1994).[6] By comparing their RMSEs from Table 3.1, a surprising result is that estimating treatment effects at individual level (i.e. ITE) is not more difficult than to estimate treatment effects at aggregated level (e.g. LATE) for every sample size. As sample size increases, both the bias and standard error decrease at the expected $\sqrt{n}$–rate. The estimation error (i.e. its size and standard deviation) depends on the sample size n and the compliers group's proportion γ_1. Specifically in the different designs, the finite sample performance of the ITE estimator depends on the value of $n \cdot \gamma_1^2$. For example, the performance of our estimator under $(n, \gamma_1) = (1000, 0.2)$ is similar to that under $(n, \gamma_1) = (4000, 0.1)$. This observation is consistent with our asymptotic properties established in the next section.

Figures 3.1 and 3.2 illustrate the performance of the ITE estimates for the n individuals with $D = 0$ and $D = 1$, respectively. In particular, we

[6] For our Monte Carlo setting, the LATE reduces to $[\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)] \,/\, [p(1) - p(0)] = 1.5351, 1.4912, 1.4449$ for $\gamma_1 = 0.1, 0.2, 0.3$, respectively. Moreover, the LATE is estimated by $\left[\overline{Y}(1) - \overline{Y}(0)\right] / \left[\hat{p}(1) - \hat{p}(0)\right]$ for a given sample, where $\overline{Y}(z)$ and $\hat{p}(z)$ are the sample means of Y and D given $Z = z$, respectively, for $z = 0, 1$. In particular, unlike ITE and its estimate, LATE and its estimate do not vary across individuals by definition.

Table 3.1: Finite sample performance of ITE

Sample size		$\gamma_1 = 0.1$	0.2	0.3
1,000	Ave. RMSE	1.2918	0.6076	0.4071
	Std. RMSE	(0.5279)	(0.2912)	(0.2231)
	LATE RMSE	{1.0448}	{0.5159}	{0.3619}
2,000	Ave. RMSE	0.9343	0.4381	0.2670
	Std. RMSE	(0.4289)	(0.2122)	(0.1511)
	LATE RMSE	{0.6639}	{0.3759}	{0.2532}
4,000	Ave. RMSE	0.6059	0.3245	0.18313
	Std. RMSE	(0.2839)	(0.1455)	(0.0985)
	LATE RMSE	{0.5057}	{0.2220}	{0.1790}

plot the ITE estimates versus the true ITE. The green solid line is the mean and the dotted lines give the 90% confidence interval computed from the 200 repetitions. The grey solid line is the 45–degree diagonal. The ITE estimates for the group $D = 1$ behave better than the estimates for $D = 0$. This observation is also consistent with our asymptotic results in the next section: The performance of $\hat{\Delta}$ of an individual with $D = d$ depends on the density function of $h(d', x, \epsilon)$, evaluated at her quantile in the distribution, conditional on the compliers group (and $X = x$ as well). In our setting, the conditional density of $h(0, \epsilon)$ given the compliers group is larger uniformly at all quantiles than that of $h(1, \epsilon)$, which leads to a more accurate estimator $\hat{\Delta}$ for the group $D = 1$. For comparison, we also plot the true value of LATE with the 90% confidence interval of its estimate in grey color columns. Overall, estimates of ITE and LATE behave similarly. Note that for any individual in the group $D = 1$, our estimator of the ITE behaves better than LATE.

For the density estimator, we choose the bandwidth $h = (\ln n/n)^{1/7}$

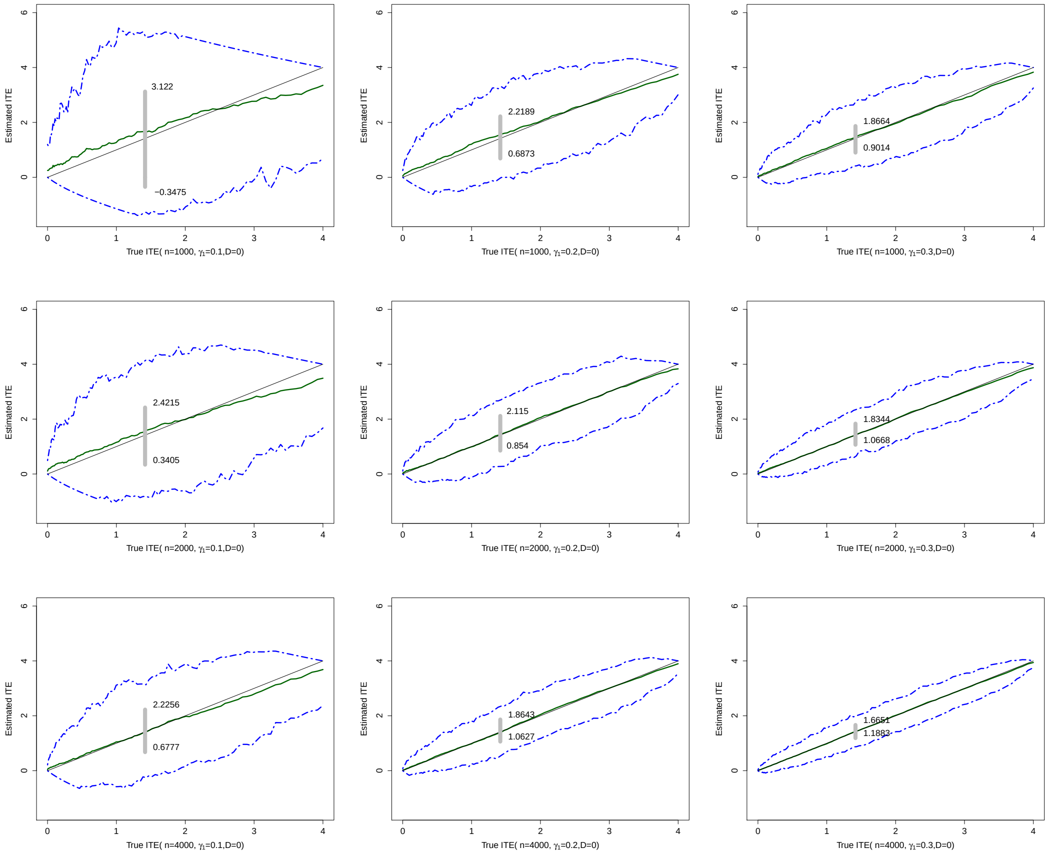


Figure 3.1: True and estimated ITE for $D = 0$

and the pdf of the standard normal as the Kernel function. Figure 3.3 shows the performance of our density estimator $\hat{f}_{\Delta}$. The black dotted line is the true density of the ITE and the green one is the average of our density estimates $\hat{f}_{\Delta}$ over the 200 repetitions. We also provide the 5% and 95% percentiles of estimated densities using blue dotted lines, which gives the (pointwise) 90% confidence band. Figure 3.3 shows again the importance of the size of the complier group through $n\gamma_1^2$.

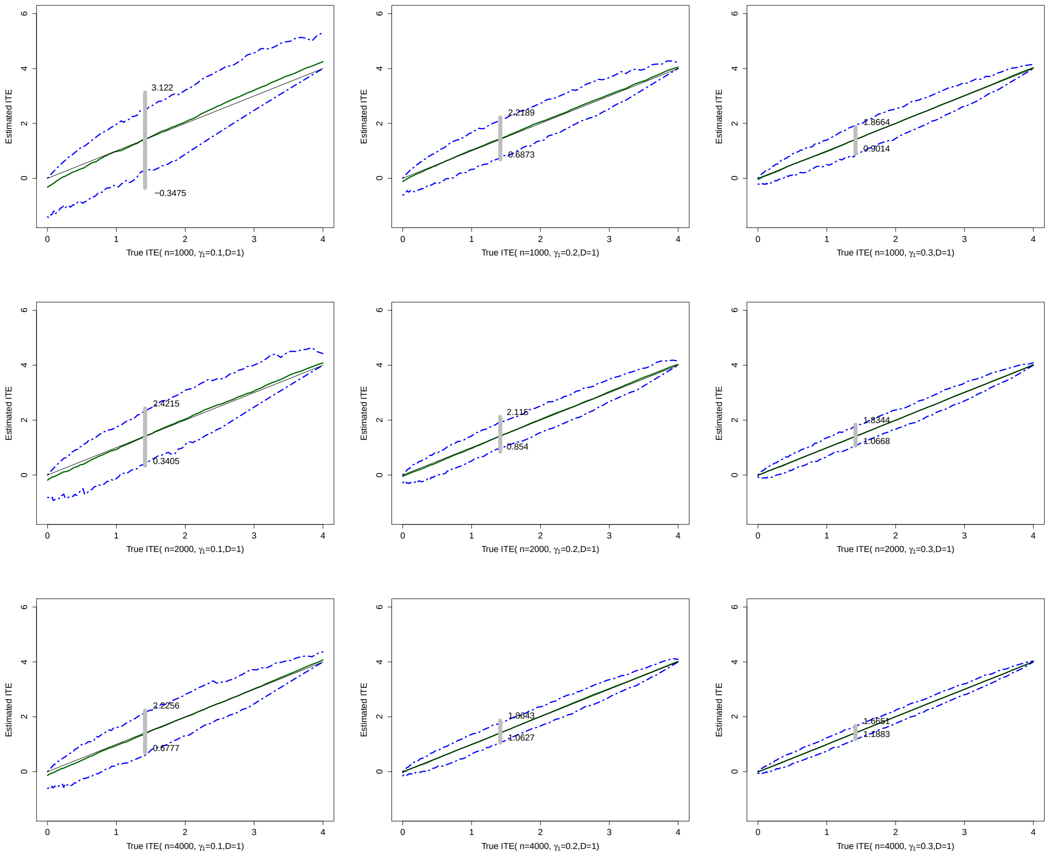


Figure 3.2: True and estimated ITE for $D = 1$

3.4 Asymptotic Properties

We now establish the asymptotic properties of our proposed nonparametric estimators. We first show the uniform $\sqrt{n}$–consistency of the counterfactual mapping estimator $\hat{\phi}_{dx}$, and we give its limiting distribution. We then establish the asymptotic properties of our density estimator $\hat{f}_\Delta$ taking into account the first-step estimation of Δ.

For estimation, we strengthen Conditions (i) and (iii) in Theorem 6,

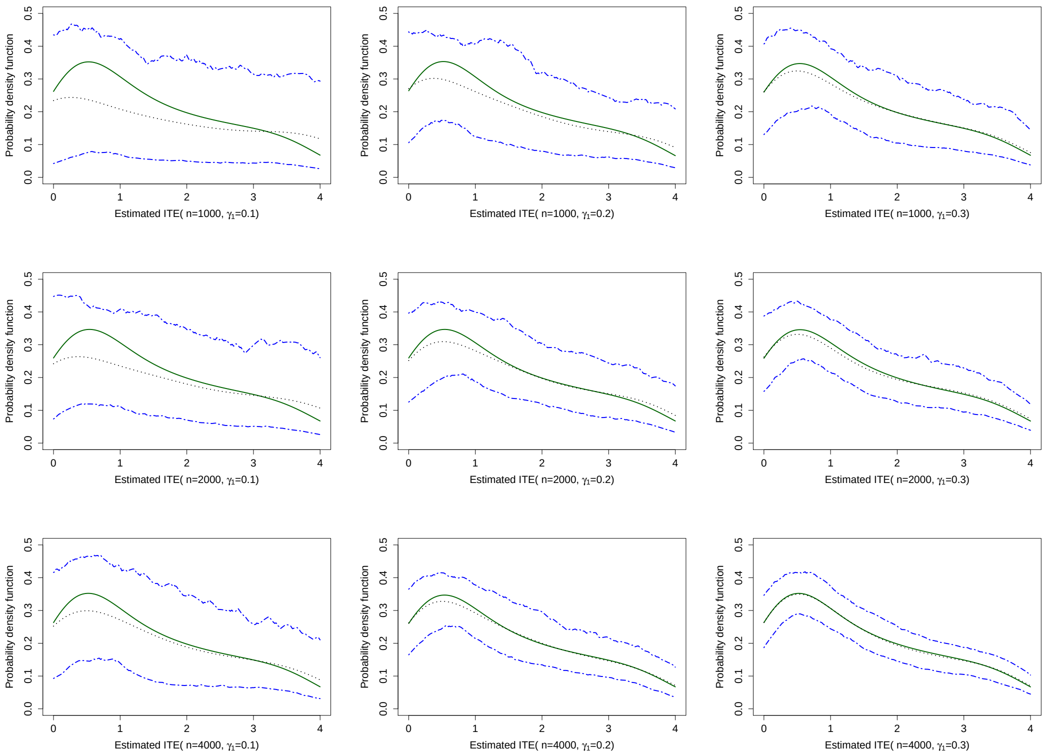


Figure 3.3: Estimated density of ITE

respectively, to

Condition (i)': h is continuously differentiable and strictly increasing in ϵ.

Condition (iii)': The conditional distribution of (ϵ, ν) given X is absolutely continuous with respect to Lebesgue measure. Moreover, the conditional density function $f_{\epsilon|X}(\cdot|x)$ is continuous for all $x \in \mathscr{S}_X$.

Under Conditions (i)', (ii) and (iii)', the conditional distribution $F_{Y|DXZ}$ is absolutely continuous with respect to Lebesgue measure and its density $f_{Y|DXZ}$

is also continuous. Therefore, the complier distribution $C_{dx}(\cdot)$ defined by (3.4) is also absolutely continuous with respect to Lebesgue measure for $d = 0, 1$. Let $c_{dx}(\cdot)$ be its density.

To simplify the exposition, we introduce the following assumption.

Assumption 1: For every $(d, x) \in \mathscr{S}_{D,X}$, (i) $\mathscr{S}_{Y|D=d,X=x} = [\underline{y}_{dx}, \overline{y}_{dx}]$, where $\underline{y}_{dx}$ and $\overline{y}_{dx}$ are finite, and (ii) $\inf_{y \in \mathscr{C}_{dx}} c_{dx}(y) > 0$. Moreover, (iii) X is a vector of discrete random variables with a finite support.

When $\mathscr{S}_{Y|D=d;X=x}$ has an unbounded support, we can always apply a known strictly increasing bounded continuous transformation to Y to satisfy Assumption 1-(i). Assumption 1-(ii) requires that the density c_{dx} be bounded away from zero on its support. It can be relaxed at the cost of technical complications due to e.g. some trimming. As indicated earlier, Assumption 1-(iii) can be relaxed to allowed for continuous variables in X by introducing some smoothing methods such as kernel ones.

The next theorem establishes the uniform consistency of the counterfactual mapping estimator $\hat{\phi}_{dx}(\cdot)$ on its full support. It also gives its $\sqrt{n}$–asymptotic distribution. For $d = 0, 1$ and $y \in \mathscr{C}_{dx}$, let $c^*_{dx}(y) = c_{dx}(y) \cdot [p(x, 1) - p(x, 0)]$ be the scale–adjusted complier density and $R_{dx}(y) = \Pr(h(d, X, \epsilon) \leq y | X = x)$ be the probability rank of y in the distribution of $h(d, x, \epsilon)$ given $X = x$. Under the monotonicity of h and the definition of $\phi_{d'x}$, we have

$$R_{dx}(y) = \Pr(Y \leq y; D = d | X = x) + \Pr(Y \leq \phi_{d'x}(y); D = d' | X = x)$$

where $d' = 1 - d$.

Theorem 7. *Suppose the conditions in Theorem 6, Conditions (i)', (iii)' and Assumption 1 hold. Then, for $d = 0, 1$ and $d' = 1 - d$, we have*

$$\sup_{y \in \mathscr{S}_{Y|D=d;X=x}} |\hat{\phi}_{d'x}(y) - \phi_{d'x}(y)| = o_p(1).$$

Moreover, the empirical process $c^*_{d'x}(\phi_{d'x}(\cdot)) \times \sqrt{n}\big(\hat{\phi}_{d'x}(\cdot) - \phi_{d'x}(\cdot)\big)$ converges in distribution to a zero–mean Gaussian process with covariance kernel

$$\Sigma_{d'x}(y, y') = \frac{R_{dx}(\min\{y, y'\}) - R_{dx}(y) \times R_{dx}(y')}{\Pr(Z = 0|X = x)\Pr(Z = 1|X = x)}.$$

The uniform convergence of $\hat{\phi}_{d'x}$ includes the boundaries, which is due to Assumption 1-(ii). Moreover, letting $y = y'$ in $\Sigma_{d'x}$ gives the asymptotic variance of $c^*_{d'x}(\phi_{d'x}(y)) \times \sqrt{n}\ \hat{\phi}_{d'x}(y)$ as follows:

$$\sigma^2_{d'x}(y) \equiv \frac{R_{dx}(y) - R^2_{dx}(y)}{\Pr(Z = 0|X = x)\Pr(Z = 1|X = x)}.$$

As y approaches its boundaries, the asymptotic variance decreases to zero. Therefore, we obtain a more accurate estimate of the counterfactual outcome when it is closer to the boundary points. We also note that the asymptotic variance of $\hat{\phi}_{d'x}(y)$ is inversely proportional to $c^{*2}_{d'x}(\phi_{d'x}(y)) = c^2_{d'x}(\phi_{d'x}(y)) \times [p(x, 1) - p(x, 0)]^2$, but is independent of the magnitude of ITE.

Theorem 7 is important for several reasons. First, given an arbitrary triplet (y, d, x), we can provide a $\sqrt{n}$–consistent estimate $\hat{\phi}_{d'x}(y)$ of the counterfactual outcome $\phi_{d'x}(y)$ whenever $y \in \mathscr{S}_{Y|D=d,X=x}$ and $x \in \mathscr{S}_X$. Its standard error is given by

$$\frac{1}{\sqrt{n} \times \hat{c}_{d'x}(\hat{\phi}_{d'x}(y))[\hat{p}(x, 1) - \hat{p}(x, 0)]} \sqrt{\frac{\hat{R}_{dx}(y) - \hat{R}^2_{dx}(y)}{\hat{\Pr}(Z = 0|X = x)\hat{\Pr}(Z = 1|X = x)}}.$$

where $\hat{R}_{dx}(y)$ and $\hat{\Pr}(Z = z|X = x)$ are sample frequencies, and

$$\hat{c}_{dx}(\cdot) \times [\hat{p}(x,1) - \hat{p}(x,0)] = (-1)^d \hat{f}_{Y|DXZ}(\cdot|d,x,z)\ \hat{\Pr}(D = d|X = x, Z = 0) \\ - (-1)^d \hat{f}_{Y|DXZ}(\cdot|d,x,z)\hat{\Pr}(D = d|X = x, Z = 1), \quad (3.7)$$

in which $\hat{f}_{Y|DXZ}(y|d,x,z)$ is a kernel density estimator and $\hat{\Pr}(D = d|X = x, Z = z)$ are sample frequencies. Equation (3.7) follows from differentiating (3.4). Second, given the uniform $\sqrt{n}$–consistency of $\hat{\phi}_{d'x}$, it follows that $\hat{\Delta}_i$ also uniformly converges to Δ_i at the $\sqrt{n}$–rate.

Next, we turn to the asymptotic properties of our density estimator $\hat{f}_\Delta$.

Assumption 2: (i) On some interval $[\underline{\delta}, \overline{\delta}]$ of $\mathscr{S}_\Delta$, the density function f_Δ admits up to P–th continuous bounded derivatives with $P \geq 1$. Moreover, $\inf_{\delta \in [\underline{\delta}, \overline{\delta}]} f_\Delta(\delta) > 0$. (ii) The kernel $K(\cdot)$ is a symmetric P-th order kernel with support $[-1, +1]$ and twice continuously bounded derivatives.[7] (iii) The bandwidth $h \propto (\ln n/n)^{1/(2P+2)}$.

The first part of Assumption 2-(i) is a high level condition requiring that the random variable $h(1, X, \epsilon) - h(0, X, \epsilon)$ has a smooth density function conditional on $X = x$. It is satisfied if $h(d, x, \cdot)$ for $d = 0, 1$ and the density of ϵ given X are$P$–th continuously differentiable. The second part of Assumption 2-(i) is standard for kernel estimation. Assumptions (ii) and (iii) relate to the choice of the kernel function K and bandwidth h, respectively. In particular,

[7] A P-th order kernel is a function integrating to one and satisfying $\int u^p K(u)du = 0$ if $1 \leq p \leq P - 1$ and $< \infty$ if $p = P$.

following Guerre et al. (2000), the bandwidth in (iii) leads to oversmoothing relative to the optimal bandwidth, i.e., $h^* \propto (\ln n/n)^{\frac{1}{2P+1}}$ (see Stone, 1982).

Given Assumption 2 and the uniform convergence of $\hat{\Delta}$ to Δ at the $\sqrt{n}$–rate, we show in the Appendix that the first–step estimation error is asymptotically negligible in $\hat{f}_{\Delta}$. Thus, we obtain the following result.

Theorem 8. *Suppose the conditions in Theorem 7 and Assumption 2 hold. Then,*

$$\sup_{\delta \in [\underline{\delta}+h, \overline{\delta}-h]} |\hat{f}_{\Delta}(\delta) - f_{\Delta}(\delta)| = O_p\big((\ln n/n)^{\frac{P}{2P+2}}\big).$$

Note that the convergence rate in Theorem 8 is uniform over the expanding interval $[\underline{\delta} + h, \overline{\delta} - h]$. It is slower than the optimal convergence rate if the ITEs were observed, which is $(\ln n/n)^{\frac{P}{2P+1}}$ (see Stone, 1982).

3.5 Individual Effects of 401(k) Programs

In this section we apply our estimation method to study the effects of 401(k) retirement programs on personal savings. The 401(k) retirement programs were introduced in the early 1980s to increase savings for retirement. Since then, they became increasingly popular in the US. It has been argued in the literature that participants might self–select into the programs non-randomly (see, e.g., Poterba et al., 1996). People with a higher preference for savings are more likely to participate and have higher savings than those with lower preferences.

Following e.g. Abadie (2003) and Chernozhukov and Hansen (2004), we use 401(k) eligibility as an instrumental variable for 401(k) participation. This is because 401 (k) plans are provided by employers. Hence, only workers in firms that offer such programs are eligible so that the monotonicity in (3.2) is satisfied.[8]

3.5.1 Data

The dataset consists of 9,275 observations from the Survey of Income and Program Participation (SIPP) of 1991 as in Abadie (2003). The observational units are household reference persons aged 25-64 and spouse if present. The included households are those with at least one member employed, with Family Income in the \$10k – \$200k interval. Eligibility for 401(k) outside the interval is rare as noted by Poterba et al. (1996).

Table 3.2 presents the summary statistics of the full sample as well as by eligibility and participation status. The dependent variable is the Family Net Financial Assets (FNFA), the treatment variable is the participation in 401(k), and the instrumental variable is the eligibility for 401(k). About 28% in the sample participate in the program and 39% are eligible for it. Other covariates include family income, age, marital status and family size. Similar to Chernozhukov and Hansen (2004), age and income are grouped into categorical

[8] Imbens and Angrist (1994) define monotonicity as: $D_i(z_1) \leq D_i(z_2)$ for all i, where $D_i(z)$ is the potential treatment status at $Z = z$. In our application, Z is 401(k) eligibility and $D_i(0) = 0$. Therefore, $D_i(0) \leq D_i(1)$ a.s., i.e., Imbens and Angrist (1994)'s monotonicity condition holds. Moreover, Vytlacil (2002) show that such a condition is observationally equivalent to the functional monotonicity in (3.2).

Table 3.2: Summary statistics

	Entire sample	By 401(k) participation		By 401(k) eligibility	
		Participants	Non-participants	Eligibles	Non-eligibles
Treatment					
401(k) Participation	0.2762			0.7044	0.0000
	(0.4472)			(0.4564)	(0.0000)
Instrument					
401(k) Eligibility	0.3921	1.0000	0.1601		
	(0.4883)	(0.0000)	(0.3668)		
Outcome variable					
FNFA	19.0717	38.4730	11.6672	30.5351	11.6768
(in thousand $)	(63.9638)	(79.2711)	(55.2892)	(75.0190)	(54.4202)
Covariates:					
Family income	39.2546	49.8151	35.2243	47.2978	34.0661
(in thousand $)	(24.0900)	(26.814.2)	(21.6492)	(25.6200)	(21.5106)
Age	41.0802	41.5133	40.9149	41.4845	40.8194
	(10.2995)	(9.6517)	(10.5323)	(9.6052)	(10.7163)
Married	0.6286	0.6956	0.6030	0.6772	0.5972
	(0.4832)	(0.4603)	(0.4893)	(0.4676)	(0.4905)
Family size	2.8851	2.9204	2.8716	2.9079	2.8703
	(1.5258)	(1.4681)	(1.5472)	(1.4770)	(1.5565)

variables 0, 1, 2 and 3 by using the 1st, 2nd and 3rd quartiles.

Table 3.3 provides the mean and standard error (in parentheses) of the outcome variable FNFA by percentiles sorted according to covariates. Clearly, FNFA is monotone increasing in family income and age. According to family size, FNFA is maximized at family size 2 and decreases with family size when it is larger than 2. Moreover, married households have higher FNFA than unmarried ones on average.

In Table 3.4, we provide OLS and 2SLS estimates as a benchmark for comparison with our ITE estimates. Our results replicate the estimates in Abadie (2003). The OLS estimates in column (1) show a significantly positive

Table 3.3: Average FNFA (in thousand $) sorted according to covariates

		Family income	Age			Married	Family size
By percentile	< 0.25	2.29	4.29	By value	0	12.83	
		(18.83)	(21.08)			(50.55)	
	0.25–0.5	7.68	14.49		1	22.76	13.59
		(29.16)	(62.78)			(70.45)	(47.59)
	0.5–0.75	16.63	21.43		2		29.11
		(53.15)	(67.33)				(82.70)
	> 0.75	49.76	36.86		3		19.17
		(104.87)	(87.31)				(66.86)
					4		17.53
							(56.83)
					> 4		12.51
							(52.46)

association between participation in 401(k) and net financial assets given covariates. Furthermore, the 2SLS estimates in column (3) confirms the positive, but attenuated treatment effects after controlling for endogeneity of participation. It turns out that FNFA increases rapidly with family income and age, and is lower for married couples and larger families.

3.5.2 ITE Estimates

To begin with, we first check the support condition for identification, i.e. Condition (iv) in Theorem 6. Because those who are not eligible for 401(k) (i.e. $Z = 0$) cannot participate in the program, then $C_{1x}(\cdot) = \Pr(Y \leq \cdot | D = 1, X = x, Z = 1)$ by (3.4) and $F_{Y|D=1,X=x} = F_{Y|D=1,X=x,Z=1}$. It follows that $\mathscr{C}_{1x} = \mathscr{S}_{Y|D=1,X=x}$ for all $x \in \mathscr{S}_X$. Hence, to check Condition (iv), it suffices to verify the support condition for $d = 0$. To do so, we estimate the density function c_{0x} by (3.7) and the density function $f_{Y|DX}(\cdot|0, x)$ directly from the data.

Table 3.4: OLS and 2SLS estimates of 401(k) participation

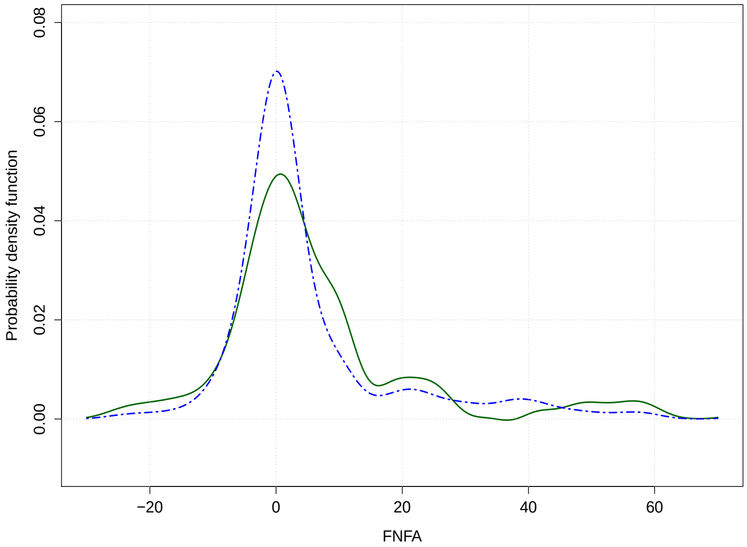
	OLS	2SLS	
		First stage	Second stage
Participation in 401(k)	13.5271		9.4188
	(1.8103)		(2.1521)
Constant	10.0421	0.0567	9.0076
	(10.9142)	(0.0464)	(10.9559)
Family income (in thousand $)	0.9769	0.0013	0.9972
	(0.0833)	(0.0001)	(0.0838)
Age	-2.3100	-0.0048	-2.2386
	(0.6177)	(0.0023)	(0.6201)
Age squared	0.0387	0.0001	0.0379
	(0.0077)	(0.0000)	(0.0077)
Married	-8.3695	-0.0005	-8.3559
	(1.8299)	(0.0079)	(1.8290)
Family size	-0.7856	0.0006	-0.8190
	(0.4108)	(0.0024)	(0.4104)
Eligibility for 401(k)		0.6883	
		(0.0080)	

Note: The dependent variable is family net financial assets (in thousand $). Family income and age enter into the regression as continuous variables. The sample includes 9,275 observations from the SIPP of 1991. The observational units are household reference persons aged 25-64, and spouse if present, with Family Income in the $10k-$200k interval. Heteroscedasticity robust standard errors are given in parentheses.

Fix the subgroup of individuals whose income is between the 25% and 50% percentile, age between 40 and 48 years old, and family size smaller than 3.[9] Figure 3.4 plots the density estimate $\hat{c}_{0x}$ using the green solid line, and the density estimate $\hat{f}_{Y|DX}(\cdot|0, x)$ using the blue dotted line. From Figure 3.4, the two distributions roughly share the same support.

[9] We repeat this for other values of covariates. The results are qualitatively similar.

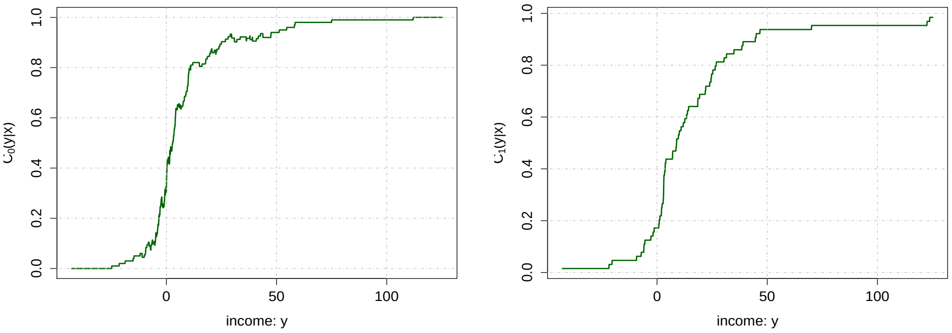
Figure 3.4: Verifying the support condition



Moreover, as shown in Vuong and Xu (forthcoming), the main restrictions imposed by our model require that $C_{dx}(\cdot)$ defined by (3.4) should be monotone increasing for $d = 0, 1$ and all $x \in \mathscr{S}_X$. We plot estimates of $C_{0x}(\cdot)$ and $C_{1x}(\cdot)$ in Figure 3.5 for the subgroup of Figure 3.4. Both of them are increasing functions globally.

Table 3.5 reports summary statistics of the ITE estimates in our sample. From Table 3.5, the ITE has a mean of \$22.45k and median Q_2 of \$8.83k, indicating a long right tail of the ITE distribution. The mean of ITE is larger than the average treatment effects (ATE) of OLS and 2SLS, which are \$13.53k and \$9.42k, respectively, while the median of ITE turns out to be smaller than these two ATEs. The differences reflect the distortion due to the linear

Figure 3.5: The model restriction



specification used in OLS and 2SLS, as well as the selection bias.

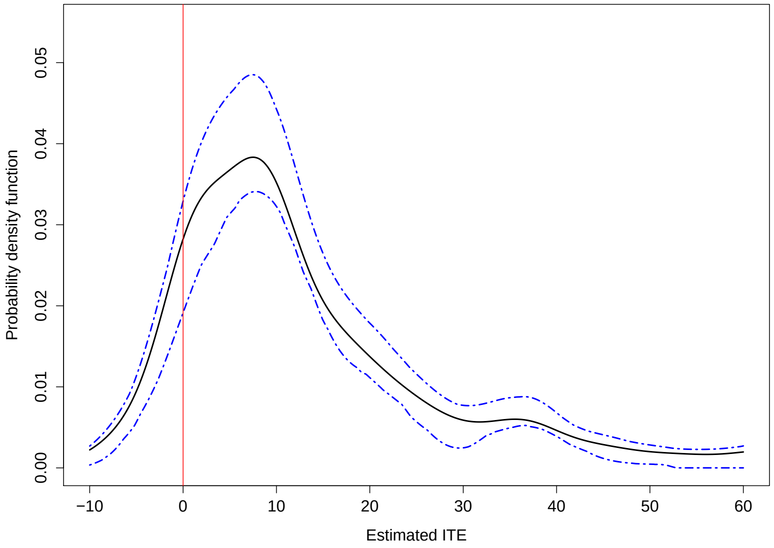
Table 3.5: Summary of ITE estimates (in thousand dollars)

Min	Max	Mean	Std.	Q_1	Q_2	Q_3.
-918	1,533	22.45	102.77	3.10	8.83	20.90

Figure 3.6 provides the ITE density estimates for the full sample along with 95% pointwise bootstrap confidence intervals. The participation effects of 401(k) on net financial assets are distributed on the interval [-\$10k, \$60k], with a mode around \$4k. As the bootstrap confidence intervals indicate, the ITE density is quite well-estimated. Figures 3.7 to 3.10 plot the ITE density estimates conditional on income, age, family size and family status, separately. In particular, the ITE density given income shifts to the right with a slight increase in variance as income increases, revealing that ITEs for individuals with high income is larger though more heterogeneous than for those whose income are low. Thus, the benefits from participating to 401(k) retirement programs on personal savings increase as Family Income increases. Though

not as pronounced, the same trend is found when conditioning on age, family size and family status.

Figure 3.6: Estimated densities of ITE for full sample



A striking feature of Figures 3.6 to 3.10 is that there exists a small but statistically significant proportion (about 8.77% in the full sample) of individuals who experience negative effects, although the majority of ITEs is positive.[10] This is especially the case for young individuals (age percentile below 0.25) where such a proportion is 15.93%. Such a finding is new. In particular, Table 3.6 provides the summary statistics of the subgroup with negative ITEs, compared with the subgroup with positive ITE and the entire

[10] For such an empirical evidence, one could investigate it alternatively by using the (conditional) quantile treatment effects for the complier group (see e.g. Abadie et al., 2002; Froelich and Melly, 2013) at low quantiles. We thank Isaiah Andrews for this point.

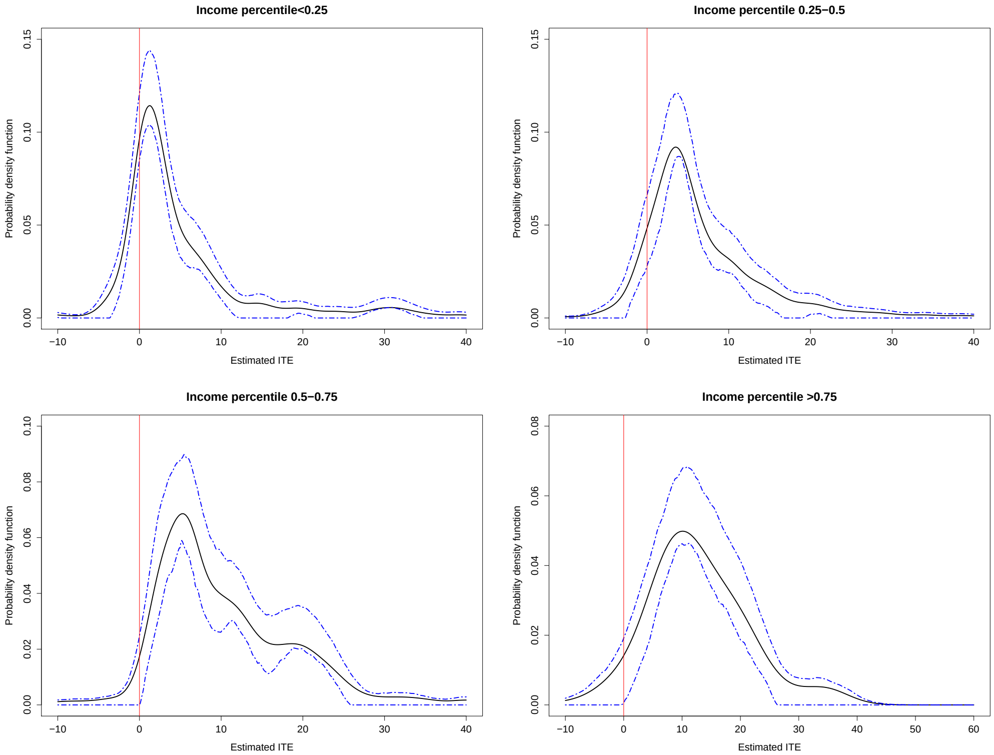


Figure 3.7: Estimated densities of ITE by income category

sample. Individuals with negative ITEs are more likely to be younger, single, and from smaller families with lower family income. A puzzling feature is that the subgroup with negative ITEs has a larger FNFA than the rest of the sample, though the large standard error (113.92) indicates a large heterogeneity among this group. Our conjecture is that the majority of this group use their savings to invest aggressively in their own businesses or in financial markets.

Figure 3.12 uses a classification tree to summarize the benefits and losses of participation decisions for all individuals in the sample: Among those who are eligible, 5.67% of them participate in 401(k) but have negative ITEs,

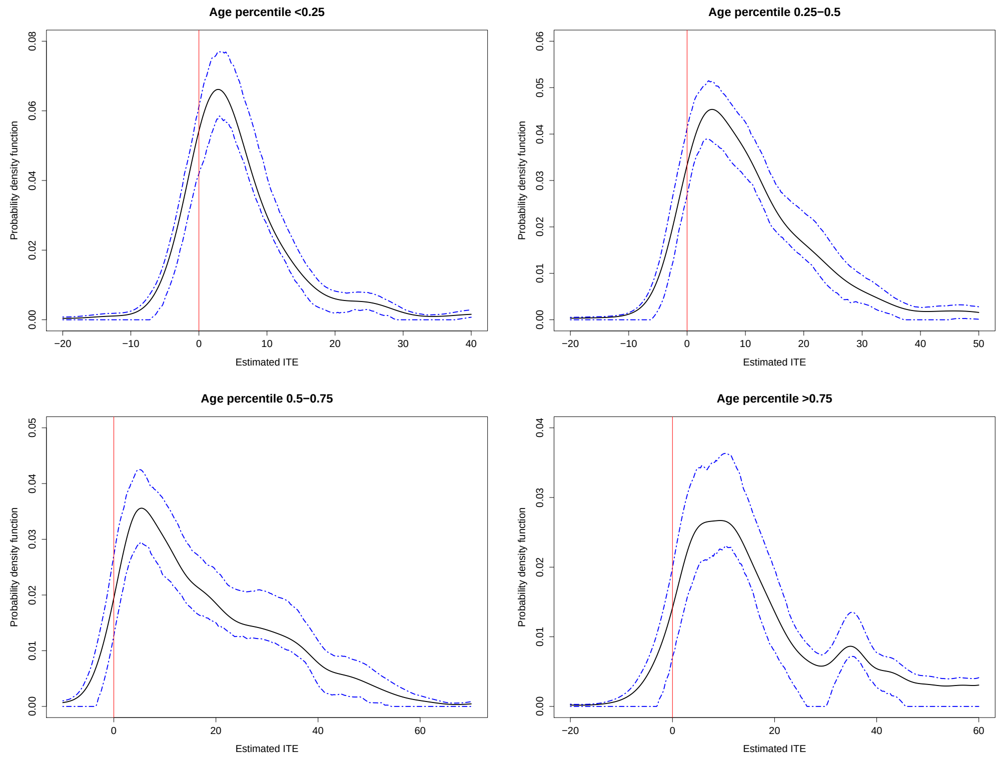


Figure 3.8: Estimated densities of ITE by age category

while 27.52% do not participate but would benefit from the 401(k) program. There are also 90.55% of non-eligible individuals who would benefit from the program if they participate. In monetary terms, the 401(k) program provides an average increase of $29.62k in FNFA to the 2,356 participants with positive ITEs and an average decrease of $19.42k in FNFA to the 206 participants with negative ITEs. That is a net increase of $65.7939 million in total in FNFA for the 401(k) program based on our sample of 9,275 households.

From Figure 3.12, about 93.12% of those who are eligible but do not participate in 401(k) programs have positive ITEs. How should one interpret

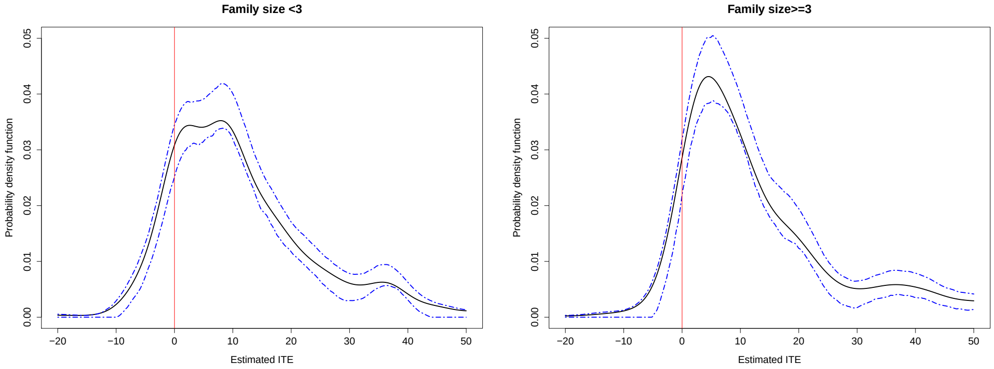


Figure 3.9: Estimated densities of ITE by family size category

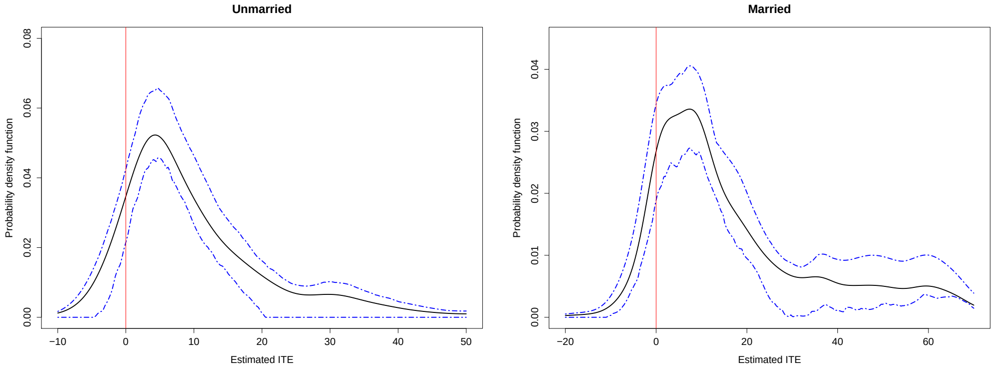


Figure 3.10: Estimated densities of ITE by marital status category

this empirical evidence? Do these eligible nonparticipants have low preference for savings, or low ability for managing their financial assets? Our ITE estimates show that the average ITE for the group of eligible nonparticipating households is $40.36k, which is significantly larger than $25.68k, the average ITE of the participating group. This evidence suggests an adverse selection issue: Households who benefit more are less likely to participate. To shed some light on this second puzzling finding, Figure 3.11 provides density estimates

Table 3.6: Summary statistics assorted according to ITE

	Negative ITE	Positive ITE	Entire sample
Participation in 401(k)	0.2534	0.2784	0.2762
	(0.4352)	(0.4482)	(0.4472)
FNFA (in thousand \$)	21.9558	18.7946	19.0717
	(113.9247)	(56.9039)	(63.9638)
Family income	30.5890	40.0872	39.2546
(in thousand \$)	(16.8846)	(24.5117)	(24.0900)
Age	34.8327	41.6805	41.0802
	(9.2949)	(10.1917)	(10.2995)
Married	0.5572	0.6354	0.6286
	(0.4970)	(0.4813)	(0.4832)
Family size	2.6421	2.9084	2.8851
	(1.4826)	(1.5280)	(1.5258)
Number	813	8,462	9,275

of the potential outcome $\hat{\phi}_{0X}(Y)$ for not participating to the 401(k) program for the participating group as well as the group of eligible nonparticipants. An interesting feature is that the distribution of participants' counterfactual FNFA (i.e., their savings without participating to 401(k) programs) are bimodal: Without participating to 401(k) programs, those participants would either do quite well or extremely poorly on their savings. In contrast, for the group of eligible but not participating households, the FNFA conforms to a unimodal distribution.

Finally, we can consider the following counterfactuals: Given that we recover the ITE for each individual, we can entertain a situation in which each eligible individual chooses his/her best option regarding participation. The 401(k) program would lead to a total increase of \$116.4681 million in FNFA coming from the 2,356 eligible households with positive ITEs and the 1,001

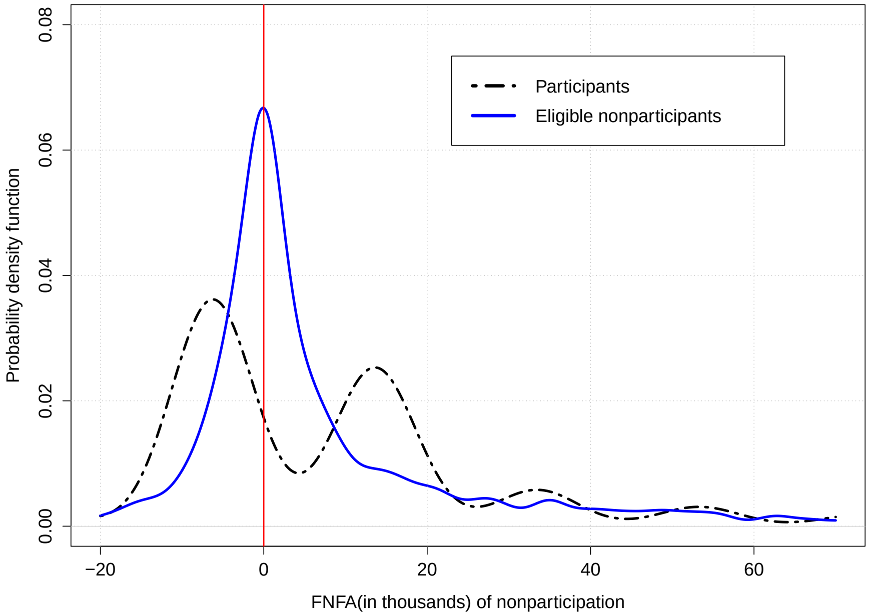
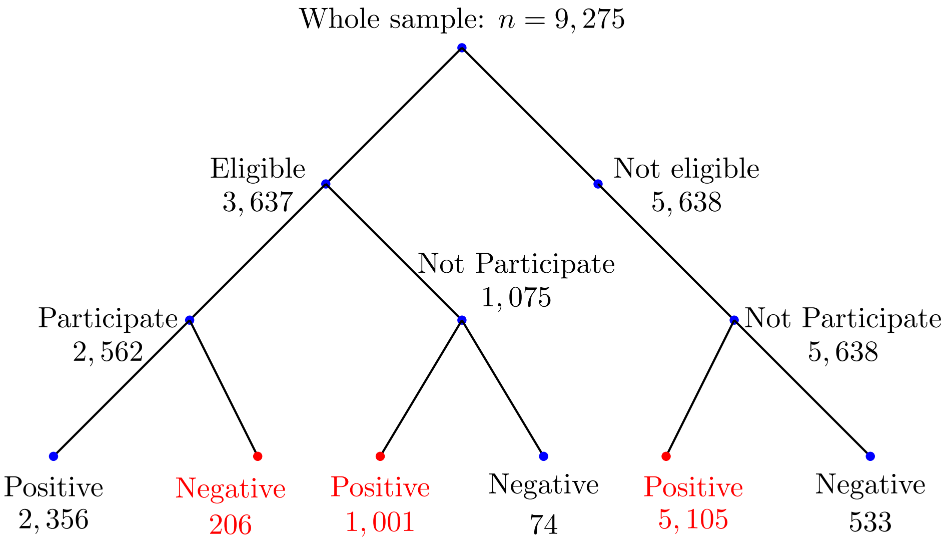


Figure 3.11: Densities of potential outcome of nonparticipation

eligible households with positive ITEs who did not participate. In addition, if the 401(k) program was available to all households, under the same scenario where each household is perfectly informed and make the correct decision, the 401(k) program will gain an additional \$120.8375 million in FNFA due to those 5,105 non-eligible households with positive ITEs. This would lead to the maximum gain of \$237.3056 million in FNFA for the 401(k) program from the 9,275 households in our sample.

Figure 3.12: Classification Tree for 401(k) Participation Decisions



3.6 Concluding Remarks

In this chapter, we invent a novel method in estimating Individual Treatment Effects (ITE) with heterogeneity and limited variations in the instruments. Our estimation approach based on "counterfactual mapping" can be extended to a lot of other scenarios. For example, one can consider continuous covariates or a mix of discrete and continuous variables. One can also study the estimation of ITE when the dependent variable is discrete. Throughout the chapter, we maintain the strict monotonicity assumption in the triangular model. Researchers can relax this assumption and conduct partial identification instead.

Appendices

Appendix A

Appendix for Chapter 1

A.1 Proofs

A.1.1 Proof of Theorem 1

Proof. Define $\mathcal{L}_2(\mathcal{W}) := \{l : \mathcal{W} \to \mathbb{R} : ||l||_2 = \sqrt{\int l(w)^2 f_W(w)dw} < \infty\}$ as the Hilbert space. Denote $M(\theta; p, h)$ and $M_n(\theta; p, h)$ to be the population moment condition and sample moment condition, respectively.

$$\begin{aligned} M(\theta, p, h) &= \mathbb{E}\left\{\left[\frac{1 - D_i}{1 - p(W_i)} A(Z_i) + \frac{D_i - p(W_i)}{1 - p(W_i)} h(W_i)\right] \epsilon_i\right\} \\ M_n(\theta, p, h) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{1 - D_i}{1 - p(W_i)} A(Z_i) + \frac{D_i - p(W_i)}{1 - p(W_i)} h(W_i)\right] \epsilon_i\right\} \end{aligned}$$

Following the sieve literature, we first give a definition of $H(\cdot, \cdot)$-smooth.

Definition 1. *A function $h(\cdot)$ is $H(\gamma, \eta)$-smooth if it belongs to a weighted Hölder ball $\Lambda_c^\gamma(\mathcal{W}, \eta)$ for some $\gamma > 0$ and $\eta \geq 0$.*

Denote such weighted Hölder ball as $\mathcal{H}$, hereafter. The following lemma states the high-level conditions required for consistency, which is an extension to Chen et al. (2003) (CLK hereafter)Theorem 1, and Pakes and Pollard (1989) Corollary 3.2.

Lemma 7. *Suppose that $\theta_0 \in \Theta$ satisfies $M(\theta_0, p_0, h_0) = 0$, and that*

(A3.1) $||M_n(\widehat{\theta}, \widehat{p}, \widehat{h})|| \leq \inf_{\theta \in \Theta} ||M_n(\theta, \widehat{p}, \widehat{h})|| + o_p(1)$*;*

(A3.2) For all $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that $\inf_{||\theta - \theta_0|| > \delta} ||M(\theta, p_0, h_0)|| \geq \epsilon(\delta) > 0$;

(A3.3) For any given $p \in \mathcal{P}$, uniformly for all $\theta \in \Theta$, $M(\theta, p, \cdot)$ is continuous w.r.t. $|| \cdot ||_{\infty\eta}$ in h at $h = h_0$; similarly, for any given $h \in \mathcal{H}$, uniformly for all $\theta \in \Theta$, $M(\theta, \cdot, h)$ is continuous w.r.t. $|| \cdot ||_{\infty\eta}$ in p at $p = p_0$;

(A3.4) $||\widehat{p} - p_0||_{\infty\eta} = o_p(1)$, $||\widehat{h} - h_0||_{\infty\eta} = o_p(1)$*;*

(A3.5) For all sequences of positive numbers $\{\delta_n\}$ with $\delta_n = o_p(1)$,

$$\sup_{\theta \in \Theta, ||p - p_0||_{\infty\eta} \leq \delta_n, ||h - h_0||_{\infty\eta} \leq \delta_n} \frac{||M_n(\theta, p, h) - M(\theta, p, h)||}{1 + ||M_n(\theta, p, h)|| + ||M(\theta, p, h)||} = o_p(1)$$

Then, $\widehat{\theta} - \theta_0 = o_p(1)$.

(A3.1) and (A3.2) are directly satisfied by identification of θ.

For $\forall p(\cdot) \in \mathcal{P}$, $\forall w \in \mathcal{W}$, $0 < \frac{1}{1-\underline{p}} \leq \frac{1}{1-p(w)} \leq \frac{1}{1-\bar{p}} < \infty$, we have

$$\begin{aligned} &|M(\theta, p, h) - M(\theta, p_0, h_0)| \\ &\qquad = |\mathbb{E}((\frac{1}{1-p(W)} - \frac{1}{1-p_0(W)})(1-D)A(Z)\epsilon \\ &\qquad + (\frac{D-p(W)}{1-p(W)}h(W)\epsilon - \frac{D-p_0(W)}{1-p_0(W)}h_0(W)\epsilon))| \end{aligned}$$

$$\begin{aligned}
\leq & \frac{1}{(1-\bar{p})^2}\mathbb{E}(||A(Z)\epsilon||(1+||W||^2)^{\frac{\eta}{2}}) \sup_{W\in\mathcal{W}} |(p(W)-p_0(W))(1+||W||^2)^{-\frac{\eta}{2}}| \\
& + |\mathbb{E}\left((\frac{D-p(W)}{1-p(W)} - \frac{D-p_0(W)}{1-p_0(W)})h(W)\epsilon \right)| \\
& + |\mathbb{E}\left(\frac{D-p_0(W)}{1-p_0(W)}(h(W)-h_0(W))\epsilon \right)| \\
\leq & \frac{1}{(1-\bar{p})^2}\mathbb{E}(||A(Z)||^2\epsilon^2)^{\frac{1}{2}}(\mathbb{E}(1+||W||^2)^{\eta})^{\frac{1}{2}}||p(\cdot)-p_0(\cdot)||_{\infty\eta} \\
& + \frac{1}{(1-\bar{p})^2}\mathbb{E}\left(||h(W)\epsilon||(1+||W||^2)^{\frac{\eta}{2}}\right) \times ||p(\cdot)-p_0(\cdot)||_{\infty\eta} \\
& + \frac{1+\bar{p}}{1-\bar{p}}\mathbb{E}|(h(W)-h_0(W)\epsilon)| \\
\leq & \frac{1}{(1-\bar{p})^2}\mathbb{E}(||A(Z)||^2\epsilon^2)^{\frac{1}{2}}(\mathbb{E}(1+||W||^2)^{\eta})^{\frac{1}{2}}||p(\cdot)-p_0(\cdot)||_{\infty\eta} \\
& + \frac{1}{(1-\bar{p})^2}(\mathbb{E}||h(W)\epsilon||^2)^{\frac{1}{2}}(\mathbb{E}(1+||W||^2)^{\eta})^{\frac{1}{2}} \times ||p(\cdot)-p_0(\cdot)||_{\infty\eta} \\
& + \frac{1+\bar{p}}{1-\bar{p}}||h(\cdot)-h_0(\cdot)||_{\infty\eta}(\sigma_{\epsilon}^2)^{\frac{1}{2}}(\mathbb{E}(1+||W||^2)^{\eta})^{\frac{1}{2}}
\end{aligned}$$

where the last inequality holds because of (A3.4) is satisfied by Chen et al. (2008) Proposition B.1(i) and Chen et al. (2005) Appendix A. In addition, for a small positive number $\delta > 0$

$$\begin{aligned}
& \mathbb{E} \sup_{||\theta-\tilde{\theta}||<\delta, ||p(\cdot)-\tilde{p}(\cdot)||_{\infty\eta}<\delta, ||h(\cdot)-\tilde{h}(\cdot)||_{\infty\eta}<\delta} |(\frac{1}{1-p(W)}\epsilon - \frac{1}{1-\tilde{p}(W)}\tilde{\epsilon})A(Z) \\
& + (\frac{D-p(W)}{1-p(W)}h(W)\epsilon - \frac{D-\tilde{p}(W)}{1-\tilde{p}(W)}\tilde{h}(W)\tilde{\epsilon})|
\end{aligned}$$

$$
\begin{aligned}
\leq & \mathbb{E}\left(\sup_{||\theta-\tilde{\theta}||<\delta, ||p(\cdot)-\tilde{p}(\cdot)||_{\infty\eta}<\delta} |(\frac{1}{1-p(W)}\epsilon - \frac{1}{1-\tilde{p}(W)}\epsilon + \frac{1}{1-\tilde{p}(W)}\epsilon - \frac{1}{1-\tilde{p}(W)}\tilde{\epsilon})A(Z)|\right) \\
& + \mathbb{E}\left(\sup_{||\theta-\tilde{\theta}||<\delta, ||p(\cdot)-\tilde{p}(\cdot)||_{\infty\eta}<\delta} |(\frac{D-p(W)}{1-p(W)} - \frac{D-\tilde{p}(W)}{1-\tilde{p}(W)})h(W)\epsilon|\right) \\
& + \frac{1+\bar{p}}{1-\bar{p}}\mathbb{E}\left(|h(W) \sup_{||\theta-\tilde{\theta}||<\delta} |g(X;\theta)-g(X;\tilde{\theta})||\right) \\
& + \frac{1+\bar{p}}{1-\bar{p}}\mathbb{E}\left(|\sup_{||\tilde{h}-h||_{\infty\eta}<\delta} \tilde{\epsilon}(\tilde{h}(W)-h(W))|\right) \\
\leq & \mathbb{E}\left(\sup_{||p(\cdot)-\tilde{p}(\cdot)||_{\infty\eta}<\delta} \frac{1}{(1-\bar{p})^2}(\sup_{W\in\mathcal{W}} |(p(\cdot)-\tilde{p}(\cdot))(1+||W||^2)^{-\frac{\eta}{2}}|||A(Z)\epsilon||(1+||W||^2)^{\frac{\eta}{2}})\right) \\
& + \mathbb{E}\left((\sup_{||\theta-\tilde{\theta}||<\delta, ||p(\cdot)-\tilde{p}(\cdot)||_{\infty\eta}<\delta} |\frac{A(Z)}{1-\tilde{p}(W)}(g(X;\tilde{\theta})-g(X;\theta))|)\right) \\
& + \frac{1+\bar{p}}{1-\bar{p}}\mathbb{E}(||h(W)||)^2\left\{\mathbb{E}\left(\sup_{||\tilde{\theta}-\theta||<\delta} |g(X^*;\theta)-g(X^*;\tilde{\theta})|^2\right)\right\}^{\frac{1}{2}} \\
& + \frac{1+\bar{p}}{1-\bar{p}}\mathbb{E}\left|\sup_{||\tilde{\theta}-\theta||<\delta} \tilde{\epsilon}(1+||W||^2)^{\frac{\eta}{2}}(1+||W||^2)^{-\frac{\eta}{2}}(\tilde{h}(W)-h(W))\right|
\end{aligned}
$$

$$
\begin{aligned}
\leq & \frac{1}{(1-\bar{p})^2}||\tilde{p}(\cdot)-p(\cdot)||_{\infty\eta}\mathbb{E}\left(||A(Z)\epsilon||(1+||W||^2)^{\frac{\eta}{2}}\right) \\
& + \frac{1}{1-\bar{p}}(\mathbb{E}||A(Z)||^2)^{\frac{1}{2}}\mathbb{E}\left(\sup_{||\theta-\tilde{\theta}||<\delta} |g(X;\tilde{\theta})-g(X;\theta)|^2\right)^{\frac{1}{2}} \\
& + \frac{1+\bar{p}}{1-\bar{p}}\mathbb{E}(||h(W)||)^2\left\{\mathbb{E}\left(\sup_{||\tilde{\theta}-\theta||<\delta} |g(X^*;\theta)-g(X^*;\tilde{\theta})|^2\right)\right\}^{\frac{1}{2}} \\
& + \frac{1+\bar{p}}{1-\bar{p}}||\tilde{h}(\cdot)-h(\cdot)||_{\infty\eta}\mathbb{E}\left(\sup_{||\tilde{\theta}-\theta||<\delta} |\tilde{\epsilon}(1+||W||^2)^{\frac{\eta}{2}}|\right) \\
\leq & const.\delta + const.a(\delta) + const.b(\delta) + const.\delta
\end{aligned}
$$

where the last inequality is due to Assumption 3. Thus we verify an alternative condition to (A3.5) (See CLK Remark 1.). Hence (A3.5) is satisfied. □

A.1.2 Proof of Theorem 2

Proof. Define $\Theta_\delta \equiv \{\theta \in \Theta : ||\theta - \theta_0|| \leq \delta\}$, $\mathcal{P}_\delta \equiv \{p \in \mathcal{P} : ||p - p_0||_{\infty\eta} \leq \delta\}$ and $\mathcal{H}_\delta \equiv \{h \in \mathcal{H} : ||h - h_0||_{\infty\eta} \leq \delta\}$ for some small δ. Let $\Gamma_1(\theta_0, p_0, h_0) = \partial M(\theta_0, p_0, h_0)/\partial\theta$ be the ordinary derivative of $M(\cdot, p_0, h_0)$ w.r.t θ evaluated at θ_0. We first introduce the following definitions of pathwise derivatives.

Definition 2. *Let $\lambda = (\lambda_p, \lambda_h)$. The pathwise derivatives w.r.t $p(\cdot)$ and $h(\cdot)$ are defined as*

$$\begin{aligned}
\Gamma_{2,p}^{(k)}(\theta_0, p_0, h_0)[\lambda_p] &= \partial_t^k M(\theta_0, p_0 + t\lambda_p, h_0)|_{t=0} \\
\Gamma_{2,h}^{(k)}(\theta_0, p_0, h_0)[\lambda_h] &= \partial_t^k M(\theta_0, p_0, h_0 + t\lambda_h)|_{t=0}, \quad k = 1, 2 \\
\Gamma_{2,ph}(\theta_0, p_0, h_0)[\lambda_p][\lambda_h] &= \partial_{t_p,t_h} M(\theta_0, p_0 + t_p\lambda_p, h_0 + t_h\lambda_h)|_{t_p=0,t_h=0}
\end{aligned}$$

The Gen-IV estimator has the semiparametric doubly robust property, in the sense that the first and second order pathwise derivatives w.r.t. $p(\cdot)$ and $h(\cdot)$ are zero. Such finding is in consistent with estimators using semiparametric doubly robust moment conditions, see e.g. Rothe and Firpo (2013).The following proposition states this property.

Proposition 1 (Double Robustness)**.** $\Gamma_{2,p}^{(k)}(\theta_0, p_0, h_0)[\lambda_p] = \Gamma_{2,h}^{(k)}(\theta_0, p_0, h_0)[\lambda_h] = 0, \quad k = 1, 2.$

Before proving the asymptotic normality of θ, we first state the high-level conditions required for asymptotic normality, summarized in the following lemma.

Lemma 8. *Suppose that $\theta_0 \in \Theta_\delta$ satisfies $M(\theta_0, p_0.h_0) = 0$ that $\widehat{\theta} - \theta_0 = o_p(1)$, and that*

(A4.1) $||M_n(\widehat{\theta}, \widehat{p}.\widehat{h})|| = \inf_{\theta\in\Theta_\delta} ||M_n(\theta, \widehat{p}.\widehat{h})|| + o_p(\frac{1}{\sqrt{n}})$

(A4.2) (i) The ordinary derivative $\Gamma_1(\theta, p_0, h_0)$ in θ of $M(\theta, p_0, h_0)$ exists for $\theta \in \Theta_\delta$, and is continuous at $\theta = \theta_0$; (ii) the matrix $\Gamma_1 \equiv \Gamma_1(\theta_0, p_0, h_0)$ is of full (column) rank.

(A4.3) For all $\theta \in \Theta_\delta$ the pathwise derivatives $\Gamma_{2,ph}(\theta, p_0, h_0)[p - p_0][h - h_0]$ of $M(\theta, p_0, h_0)$ exists in all directions $[p - p_0] \in \mathcal{P}$, $[h - h_0] \in \mathcal{H}$; and for all $(\theta, p, h) \in \Theta_{\delta_n} \times \mathcal{P}_{\delta_n} \times \mathcal{H}_{\delta_n}$ with a positive sequence $\delta_n = o(1)$: (i) $||M(\theta, p, h) - M(\theta, p_0, h_0) - \sum_{k=1,2}\sum_{j=p,h} \frac{1}{k!}\Gamma^{(k)}_{2,j}(\theta, p_0, h_0)[\cdot] - \Gamma_{2,ph}(\theta, p_0, h_0)[p-p_0][h-h_0]|| \leq c_1||p-p_0||^3_{\infty\eta} + c_2||h-h_0||_{\infty\eta}||p-p_0||^2_{\infty\eta}$ for constants c_1 and $c_2 \geq 0$; (ii) $||\sum_{k=1,2}\sum_{j=p,h} \frac{1}{k!}\Gamma^{(k)}_{2,j}(\theta, p_0, h_0)[\cdot] + \Gamma_{2,ph}(\theta, p_0, h_0)[p-p_0][h-h_0] - \Gamma_{2,ph}(\theta_0, p_0, h_0)[p-p_0][h-h_0]|| \leq o(1)\delta_n$.

(A4.4) $\widehat{p} \in \mathcal{P}$, $\widehat{h} \in \mathcal{H}$ with probability tending to one; and $||\widehat{p} - p_0||_{\infty\eta} = o_p(n^{-\frac{1}{6}})$ and $||\widehat{h} - h_0||_{\infty\eta} = o_p(n^{-\frac{1}{6}})$.

(A4.5) For all sequences of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$,

$$\sup_{\substack{||\theta-\theta_0||\leq\delta_n,\\ ||p-p_0||_{\infty\eta}\leq\delta_n,\\ ||h-h_0||_{\infty\eta}\leq\delta_n}} \frac{\sqrt{n}||M_n(\theta, p, h) - M(\theta, p, h) - M_n(\theta_0, p_0, h_0)||}{1 + \sqrt{n}(||M_n(\theta, p, h)|| + ||M(\theta, p, h)||)} = o_p(1).$$

(A4.6) For some finite matrix V_0, $\sqrt{n}\{M_n(\theta_0, p_0, h_0)+\Gamma_{2,ph}(\theta_0, p_0, h_0)[\widehat{p}-p_0][\widehat{h}-h_0]\} \Rightarrow N(0, V_0)$. Then, $\sqrt{n}(\widehat{\theta}-\theta_0) \Rightarrow N(0, \Omega)$, where $\Omega \equiv (\Gamma_1^{'} V\Gamma_1)^{-1}\Gamma_1^{'} V V_0 V\Gamma_1(\Gamma_1^{'} V\Gamma_1)^{-1}$.

Proof. The proof is very similar to that of Theorem 2 in CLK and Theorem 3.3 in Pakes and Pollard (1989) for $\sqrt{n}$-normality. The major difference is that we define the linearization $\mathcal{L}_n(\theta) = M_n(\theta_0, p_0, h_0)+\Gamma_1(\theta-\theta_0)+\Gamma_2^{ph}(\theta_0, p_0, h_0)[\widehat{p}-p_0][\widehat{h}-h_0]$.

It then can be proved that $||M_n(\bar{\theta}, \widehat{p}, \widehat{h}) - \mathcal{L}_n(\bar{\theta})|| = o_p(n^{-1/2})$, where

$$\sqrt{n}(\bar{\theta}-\theta_0) = (\Gamma_1^{'} V\Gamma_1)^{-1}\Gamma_1^{'} V\sqrt{n}\left[M_n(\theta_0, p_0, h_0) + \Gamma_2^{ph}(\theta_0, p_0, h_0)[\widehat{p}-p_0][\widehat{h}-h_0]\right]$$

is the minimizer of $\mathcal{L}_n(\theta)$, combined with $\sqrt{n}(\widehat{\theta}-\bar{\theta}) = o_p(1)$. This and condition (A4.6) imply that $\sqrt{n}(\widehat{\theta}-\theta_0) \Rightarrow N(0, \Omega)$. □

We now prove the theorem by checking conditions in Lemma 8 one by one. First, note that by Theorem 1, Assumption 4.1 and 4.2, (A4.1) and (A4.2) are satisfied.

For (A4.3)(ii),

$$\begin{aligned}&\left|\left|\sum_{k=1,2}\sum_{j=p,h}\frac{1}{k!}\Gamma_{2,j}^{(k)}(\theta, p_0, h_0)[\cdot]\right|\right|\\ &\qquad\le ||\mathbb{E}\{\frac{(1-D)(p(W)-p_0(W))(A(Z)-h_0(W))\epsilon}{(1-p_0(W))^2}\\ &\qquad + \frac{(1-D)(p(W)-p_0(W))^2(A(Z)-h_0(W))\epsilon}{(1-p_0(W))^3}\}||\end{aligned}$$

$$\begin{aligned}
&+\left|\left|\mathbb{E}\left\{\frac{D-p_0(W)}{1-p_0(W)}(h(W)-h_0(W))\epsilon\right\}\right|\right| \\
\leq&\frac{1}{(1-\bar{p})^2}\mathbb{E}\left|(A(Z)-h_0(W))(1+||W||^2)^{\frac{\eta}{2}}\epsilon\right|\sup_{W\in\mathcal{W}}|(p(W)-p_0(W))(1+||W||^2)^{-\frac{\eta}{2}}| \\
&+\frac{1}{(1-\bar{p})^3}\mathbb{E}\left|(A(Z)-h_0(W))(1+||W||^2)^{\eta}\epsilon\right|\left(\sup_{W\in\mathcal{W}}|(p(W)-p_0(W))(1+||W||^2)^{-\frac{\eta}{2}}|\right)^2 \\
&+\frac{1+\bar{p}}{1-\bar{p}}\mathbb{E}\left|(1+||W||^2)^{\frac{\eta}{2}}\epsilon\right|\sup_{W\in\mathcal{W}}|(h(W)-h_0(W))(1+||W||^2)^{-\frac{\eta}{2}}| \\
\leq&\frac{1}{(1-\bar{p})^2}\mathbb{E}\sup_{||\theta-\theta_0||\leq\delta}\left|(A(Z)-h_0(W))(1+||W||^2)^{\frac{\eta}{2}}\epsilon\right|||p-p_0||_{\infty\eta} \\
&+\frac{1}{(1-\bar{p})^3}\mathbb{E}\sup_{||\theta-\theta_0||\leq\delta}\left|(A(Z)-h_0(W))(1+||W||^2)^{\frac{\eta}{2}}\epsilon\right|||p-p_0||_{\infty\eta}^2 \\
&+\frac{1+\bar{p}}{1-\bar{p}}\mathbb{E}\sup_{||\theta-\theta_0||\leq\delta}\left|(1+||W||^2)^{\frac{\eta}{2}}\epsilon\right|||h-h_0||_{\infty\eta} \\
\leq&const.\delta+const.\delta^2+const.\delta
\end{aligned}$$

where the last inequality is due to Assumption 4.

$$\begin{aligned}
&||\Gamma_2^{ph}(\theta,p_0,h_0)[p-p_0][h-h_0]-\Gamma_2^{ph}(\theta_0,p_0,h_0)[p-p_0][h-h_0]|| \\
=&\left|\left|\mathbb{E}\frac{(1-D)(p(W)-p_0(W))(h(W)-h_0(W))(g(X^*;\theta_0)-g(X^*;\theta))}{(1-p_0(W))^2}\right|\right| \\
\leq&\frac{1+\bar{p}}{(1-\bar{p})^2}\times\sup_{W\in\mathcal{W}}|(p(W)-p_0(W))(1+||W||^2)^{-\frac{\eta}{2}}|\times... \\
&...\times\sup_{W\in\mathcal{W}}|(h(W)-h_0(W))(1+||W||^2)^{-\frac{\eta}{2}}|\times... \\
&...\times||\theta-\theta_0||\mathbb{E}\left\{||\frac{\partial g(X^*;\bar{\theta})}{\partial\theta}||(1+||W||^2)^{\eta}\right\} \\
=&||p-p_0||_{\infty\eta}||h-h_0||_{\infty\eta}||\theta-\theta_0||\mathbb{E}\left\{||\frac{\partial g(X^*;\bar{\theta})}{\partial\theta}||(1+||W||^2)^{\eta}\right\} \\
\leq&||p-p_0||_{\infty\eta}||h-h_0||_{\infty\eta}||\theta-\theta_0||\sup_{||\theta-\theta_0||<\delta}\mathbb{E}\left\{||\frac{\partial g(X;\theta)}{\partial\theta}||(1+||W||^2)^{\eta}\right\}
\end{aligned}$$

where $\bar{\theta}$ is between θ and θ_0.

Thus, under Assumption 4, Proposition B.1(i) of CLK, Proposition A.1 of CHT, (A4.3)(ii) is satisfied.

For (A4.5), we instead check a sufficient condition which is an extension of Andrews (1994)'s "type IV class" from $\theta \in \Theta$ to $(\theta, p, h) \in \Theta \times \mathcal{P} \times \mathcal{H}$. Recall $\tilde{\epsilon} = y - g(X; \tilde{\theta})$ and $\tilde{\epsilon} - \epsilon = g(X; \theta) - g(X; \tilde{\theta})$. Let δ be a small positive number.

$$
\begin{aligned}
&\mathbb{E}\{ \sup_{\substack{||\tilde{\theta}-\theta||<\delta, \\ ||\tilde{p}-p||_{\infty\eta}<\delta, \\ ||\tilde{h}-h||_{\infty\eta}<\delta}} |\frac{1-D}{1-p(W)} A(Z)\epsilon + \frac{D-p(W)}{1-p(W)} h(W)\epsilon \\
&\quad - \frac{1-D}{1-\tilde{p}(W)} A(Z)\tilde{\epsilon} + \frac{D-\tilde{p}(W)}{1-\tilde{p}(W)} \tilde{h}(W)\tilde{\epsilon}|^2\} \\
&\le \mathbb{E}\{ \sup_{||\tilde{\theta}-\theta||<\delta, ||\tilde{p}-p||_{\infty\eta}<\delta, ||\tilde{h}-h||_{\infty\eta}<\delta} 2\left| \frac{1-D}{1-p(W)} A(Z)\epsilon - \frac{1-D}{1-\tilde{p}(W)} A(Z)\tilde{\epsilon}| \right|^2 \\
&\quad + 2\left| \frac{D-p(W)}{1-p(W)} h(W)\epsilon - \frac{D-\tilde{p}(W)}{1-\tilde{p}(W)} \tilde{h}(W)\tilde{\epsilon} \right|^2 \} \\
&\le 4\mathbb{E}\left\{ \sup_{||\tilde{p}-p||_{\infty\eta}<\delta} \left| (\frac{1}{1-p(W)} - \frac{1}{1-\tilde{p}(W)}) A(Z)\epsilon \right|^2 \right\} \\
&\quad + 4\mathbb{E}\left\{ \sup_{||\tilde{\theta}-\theta||<\delta} \left| \frac{A(Z)}{1-\tilde{p}(W)} (g(X^*; \tilde{\theta}) - g(X^*; \theta)) \right|^2 \right\} \\
&\quad + 8\mathbb{E}\left\{ \sup_{||\tilde{p}-p||_{\infty\eta}<\delta} \left| (\frac{D-p(W)}{1-p(W)} - \frac{D-\tilde{p}(W)}{1-\tilde{p}(W)}) h(W)\epsilon \right|^2 \right\} \\
&\quad + 8\mathbb{E}\left\{ \sup_{||\tilde{\theta}-\theta||<\delta} \left| \frac{D-\tilde{p}(W)}{1-\tilde{p}(W)} h(W)(\epsilon - \tilde{\epsilon}) \right|^2 \right\} \\
&\quad + 4\mathbb{E}\left\{ \sup_{||\tilde{h}-h||_{\infty\eta}<\delta} \left| \frac{D-\tilde{p}(W)}{1-\tilde{p}(W)} (h(W) - \tilde{h}(W))\tilde{\epsilon} \right|^2 \right\}
\end{aligned}
$$

$$
\begin{aligned}
\leq & \frac{4}{(1-\bar{p})^2}\mathbb{E}\left\{\sup_{||\tilde{p}-p||_{\infty\eta}<\delta}|(p(W)-\tilde{p}(W))A(Z)\epsilon|^2\right\} \\
& + \frac{4}{(1-\bar{p})^2}\mathbb{E}\left\{\sup_{||\tilde{\theta}-\theta||<\delta}\left|A(Z)(g(X^*;\tilde{\theta})-g(X^*;\theta))\right|^2\right\} \\
& + \frac{8}{(1-\bar{p})^2}||\tilde{p}-p||^2_{\infty\eta}\mathbb{E}\left|(1+||W||^2)^{\frac{\eta}{2}}h(W)\epsilon\right|^2 + \frac{8(1+\bar{p})^2}{(1-\bar{p})^2}\mathbb{E}\left\{\sup_{||\tilde{\theta}-\theta||<\delta}|h(W)(\epsilon-\tilde{\epsilon})|^2\right\} \\
& + \frac{4(1+\bar{p})^2}{(1-\bar{p})^2}\mathbb{E}\left\{\sup_{||\tilde{h}-h||_{\infty\eta}<\delta}\left|(h(W)-\tilde{h}(W))(1+||W||^2)^{-\frac{\eta}{2}}(1+||W||^2)^{\frac{\eta}{2}}\tilde{\epsilon}\right|^2\right\} \\
\leq & \frac{4}{(1-\bar{p})^2}||\tilde{p}-p||^2_{\infty\eta}\left[\mathbb{E}(1+||W||^2)^{2\eta}\mathbb{E}|A(Z)\epsilon|^4\right]^{1/2} \\
& + \frac{4}{(1-\bar{p})^2}\mathbb{E}\left\{\sup_{||\tilde{\theta}-\theta||<\delta}\left|A(Z)(g(X^*;\tilde{\theta})-g(X^*;\theta))\right|^2\right\} \\
& + \frac{8}{(1-\bar{p})^2}||\tilde{p}-p||^2_{\infty\eta}(\mathbb{E}(1+||W||^2)^{2\eta}\mathbb{E}|h(W)\epsilon|^4)^{1/2} \\
& + \frac{8(1+\bar{p})^2}{(1-\bar{p})^2}(\mathbb{E}|h(W)|^4)^{1/2}\left[\mathbb{E}\left\{\sup_{||\tilde{\theta}-\theta||<\delta}|\tilde{\epsilon}-\epsilon|^4\right\}\right]^{1/2} \\
& + \frac{4(1+\bar{p})^2}{(1-\bar{p})^2}||\tilde{h}-h||^2_{\infty\eta}\mathbb{E}\left|(1+||W||^2)^{\frac{\eta}{2}}\tilde{\epsilon}\right|^2 \\
\leq & const.\delta^2 + const.\delta^{2\nu} + const.\delta^2 + const.\delta^{2\nu} + const.\delta^2
\end{aligned}
$$

where the last inequality is due to Assumption 4.3, 4.4 and CHT's Proposition B.1(i).

About the covering numbers, let $N(\delta, \Lambda_c^\gamma(\mathcal{W},\eta), ||\cdot||_{\infty\eta})$ denote the covering number of the weighted Hölder ball under the adjusted Hölder norm $||\cdot||_{\infty\eta}$ (i.e. the minimal number of N for which there exist δ-balls $\{l : ||l - u_j||_{\infty\eta} \leq \delta\}$, $j = 1, ..., N$ to cover $\Lambda_c^\gamma(\mathcal{W},\eta)$). Using Van Der Vaart and Wellner

(1996) Theorem 2.7.1, we know that

$$\log N(\delta, \Lambda_c^\gamma(\mathcal{W}, \eta), ||\cdot||_{\infty\eta}) \leq const.\delta^{-\frac{d_w}{\gamma}}$$

Then assumption... implies that

$$\int_0^\infty \sqrt{N(\delta, \Lambda_c^\gamma(\mathcal{W}, \eta), ||\cdot||_{\infty\eta})} d\delta < \infty$$

which means that the class $\{l(\cdot), l(\cdot) \in \Lambda_c^\gamma(\mathcal{W}, \eta)\}$ is a F_W-Donsker class, and which satisfies the stochastic equicontinuity condition.

Now we check (A4.3)(i) and (A4.4). It is standard in the sieve literature(see e.g. CLK, CHT, Ai and Chen (2003)) to replace (A4.3)(i) and (A4.4) by a sufficient condition

$$\begin{aligned} \left|\left| M(\theta, p, h) - M(\theta, p_0, h_0) - \sum_{k=1,2} \sum_{j=p,h} \frac{1}{k!} \Gamma_{2,j}^{(k)}(\theta, p_0, h_0)[\cdot] - \Gamma_{2,ph}(\theta, p_0, h_0)[p - p_0][h - h_0] \right|\right| \\ = o_p(n^{-\frac{1}{2}}) \end{aligned} \tag{A.1}$$

Since

$$\begin{aligned} \Gamma_{2,p}^{(1)}(\theta_0, p_0, h_0)[p - p_0] &= \mathbb{E}\left\{\left[\frac{(1-D)(p(W) - p_0(W))(A(Z) - h_0(W))}{(1 - p_0(W))^2}\right]\epsilon\right\} \\ \Gamma_{2,p}^{(2)}(\theta_0, p_0, h_0)[p - p_0] &= \mathbb{E}\left\{\left[\frac{2(1-D)(p(W) - p_0(W))^2(A(Z) - h_0(W))}{(1 - p_0(W))^3}\right]\epsilon\right\} \\ \Gamma_{2,h}^{(1)}(\theta_0, p_0, h_0)[h - h_0] &= \mathbb{E}\left[\frac{D - p_0(W)}{1 - p_0(W)}(h(W) - h_0(W))\epsilon\right] \\ \Gamma_{2,h}^{(2)}(\theta_0, p_0, h_0)[h - h_0] &= 0 \\ \Gamma_{2,ph}(\theta_0, p_0, h_0)[p - p_0][h - h_0] &= \mathbb{E}\left\{\left[\frac{(1-D)(p(W) - p_0(W))(h(W) - h_0(W))}{(1 - p_0(W))^2}\right]\epsilon\right\} \end{aligned}$$

Now

$$\left|\left|M(\theta,p,h)-M(\theta,p_0,h_0)-\sum_{k=1,2}\sum_{j=p,h}\frac{1}{k!}\Gamma_{2,j}^{(k)}(\theta,p_0,h_0)[\cdot]-\Gamma_{2,ph}(\theta,p_0,h_0)[p-p_0][h-h_0]\right|\right|$$

$$\begin{aligned}
=&||\mathbb{E}\{\frac{(1-D)(A(Z)-h_0(W))(p(W)-p_0(W))}{1-p_0(W)}\left(\frac{1}{1-p(W)}-\frac{1}{1-p_0(W)}\right)\epsilon\\
&-\left[\frac{(1-D)(p(W)-p_0(W))(A(Z)-h_0(W))}{(1-p_0(W))^2}\right]\epsilon\\
&-\left[\frac{(1-D)(p(W)-p_0(W))^2(A(Z)-h_0(W))}{(1-p_0(W))^3}\right]\epsilon\\
&+\left[\frac{(1-D)(p(W)-p_0(W))(h(W)-h_0(W))}{(1-p_0(W))^2}\right]\epsilon\}||\\
=&||\mathbb{E}\{[\frac{(1-D)(p(W)-p_0(W))^2(A(Z)-h_0(W))}{(1-p_0(W))^2(1-p(W))}\\
&-\frac{(1-D)(p(W)-p_0(W))^2(h(W)-h_0(W))}{(1-p_0(W))^2(1-p(W))}\\
&-\frac{(1-D)(p(W)-p_0(W))^2(A(Z)-h_0(W))}{(1-p_0(W))^3}]\epsilon\}||\\
=&||\mathbb{E}\{[\frac{(1-D)(A(Z)-h_0(W))(p(W)-p_0(W))^3}{(1-p_0(W))^3(1-p(W))}\\
&-\frac{(1-D)(p(W)-p_0(W))^2(h(W)-h_0(W))}{(1-p(W))(1-p_0(W))}]\epsilon\}||\\
\leq&\frac{1}{(1-\bar{p})^4}\mathbb{E}\left|\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\delta}||(1-D)(A(Z)-h_0(W))\epsilon||\times(p(W)-p_0(W))^3\right|\\
&+\frac{1}{(1-\bar{p})^3}\mathbb{E}\left|\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\delta}||(1-D)\epsilon||\times(p(W)-p_0(W))^2(h(W)-h_0(W))\right|\\
\leq&\frac{1}{(1-\bar{p})^4}\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\delta}\left\{\sup_W||(A(Z)-h_0(W))\epsilon||\times\mathbb{E}|p(W)-p_0(W)|^3\right\}\\
&+\frac{1}{(1-\bar{p})^3}\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\delta}\left\{\sup_W||\epsilon||\times\left[\mathbb{E}|p(W)-p_0(W)|^3\right]^{\frac{2}{3}}\left[\mathbb{E}|h(W)-h_0(W)|^3\right]^{\frac{1}{3}}\right\}\\
\leq&const.\,(||p(\cdot)-p_0(\cdot)||_3)^3+const.\,(||p(\cdot)-p_0(\cdot)||_3)^2\,(||h(\cdot)-h_0(\cdot)||_3)
\end{aligned}$$

By Assumption 5, $||p(\cdot)-p_0(\cdot)||_3 = o_p(n^{-\frac{1}{6}})$ and $||h(\cdot)-h_0(\cdot)||_3 = o_p(n^{-\frac{1}{6}})$, condition (A.1) holds.

Instead, we can rewrite the last two inequalities,

$$\begin{aligned}
&\left|\left|M(\theta,p,h)-M(\theta,p_0,h_0)-\sum_{k=1,2}\sum_{j=p,h}\frac{1}{k!}\Gamma^{(k)}_{2,j}(\theta,p_0,h_0)[\cdot]-\Gamma_{2,ph}(\theta,p_0,h_0)[p-p_0][h-h_0]\right|\right| \\
\leq&\frac{1}{(1-\bar{p})^4}\mathbb{E}\left|\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\delta}||(1-D)(A(Z)-h_0(W))\epsilon||\times(p(W)-p_0(W))^3\right| \\
&+\frac{1}{(1-\bar{p})^3}\mathbb{E}\left|\sup_{\theta\in\Theta:||\theta-\theta_0||\leq\delta}||(1-D)\epsilon||\times(p(W)-p_0(W))^2(h(W)-h_0(W))\right| \\
\leq& const.\left(\mathbb{E}(p(W)-p_0(W))^4\right)^{\frac{1}{2}}\left(\mathbb{E}(p(W)-p_0(W))^2\right)^{\frac{1}{2}} \\
&+const.\left(\mathbb{E}(p(W)-p_0(W))^4\right)^{\frac{1}{2}}\left(\mathbb{E}(h(W)-h_0(W))^2\right)^{\frac{1}{2}} \\
\leq& const.||p(\cdot)-p_0(\cdot)||_2^{3-\frac{d_W}{2\gamma}}+const.||p(\cdot)-p_0(\cdot)||_2^{2-\frac{d_W}{2\gamma}}||h(\cdot)-h_0(\cdot)||_2
\end{aligned}$$

where the last inequality is due to the following inequalities, for any $s\in[\frac{d_w}{4},\gamma)$,

$$\begin{aligned}
\left(\mathbb{E}(p(W)-p_0(W))^4\right)^{\frac{1}{4}}\leq& const.\left(||p(\cdot)-p_0(\cdot)||_2+||\nabla^s\{p(\cdot)-p_0(\cdot)\}||_2\right) \\
||\nabla^s\{p(\cdot)-p_0(\cdot)\}||_2\leq& const.||p(\cdot)-p_0(\cdot)||_2^{1-\frac{s}{\gamma}}
\end{aligned}$$

By optimal convergence rate property of sieve estimators (see e.g. Shen and Wong (1994), CHT), $k_{p,n}=O(n^{\frac{d_W}{(2\gamma+d_W)}})$, $k_{h,n}=O(n^{\frac{d_W}{(2\gamma+d_W)}})$, and $\gamma>\frac{1}{2}d_W$, imply that

$$\begin{aligned}
||p(\cdot)-p_0(\cdot)||_2=&O_p(n^{-\frac{\gamma}{2\gamma+d_W}}),\quad ||h(\cdot)-h_0(\cdot)||_2=O_p(n^{-\frac{\gamma}{2\gamma+d_W}}), \\
||p(\cdot)-p_0(\cdot)||_2^{3-\frac{d_W}{2\gamma}}=&o_p(n^{-1/2}),\quad ||p(\cdot)-p_0(\cdot)||_2^{2-\frac{d_W}{2\gamma}}||h(\cdot)-h_0(\cdot)||_2=o_p(n^{-1/2})
\end{aligned}$$

For (A4.6), from the above argument and $\gamma>\frac{1}{2}d_W$,

$$\Gamma_{2,ph}(\theta, p_0, h_0)[\widehat{p}-p_0][\widehat{h}-h_0] = \mathbb{E}\left\{\left[\frac{(1-D)(\widehat{p}(W)-p_0(W))(\widehat{h}(W)-h_0(W))}{(1-p_0(W))^2}\right]\epsilon\right\}$$
$$\leq const.||\widehat{p}(\cdot)-p_0(\cdot)||_2||\widehat{h}(\cdot)-h_0(\cdot)||_2$$
$$= o_p(n^{-1/2})$$

Hence, (A4.6) is satisfied since

$$\sqrt{n}\{M_n(\theta_0, p_0, h_0) + \Gamma_{2,ph}(\theta_0, p_0, h_0)[\widehat{p}-p_0][\widehat{h}-h_0]\}$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{(1-D_i)A(Z_i)}{1-p_0(W_i)} + \frac{D_i - p_0(W_i)}{1-p_0(W_i)}\mathbb{E}(A(Z_i)|W=W_i)\right\} + o_p(1)$$

A standard GMM estimator θ will satisfy

$$\sqrt{n}(\widehat{\theta}-\theta_0)$$
$$= -\left(\Gamma_1^{'}G\Gamma_1\right)^{-1}\Gamma_1^{'}G\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{(1-D_i)A(Z_i)}{1-p_0(W_i)} + \frac{D_i-p_0(W_i)}{1-p_0(W_i)}\mathbb{E}(A(Z_i)|W=W_i)\right\}+o_p(1)$$

thus Theorem 2 is established. □

A.1.3 Proof of Theorem 3

Proof. I follow the structure of the semiparametric efficiency bound derivation of Newey (1990) as well as CHT. Define a factorization of the propensity score $p(w;\psi) = \mathbb{P}(D = 1|W = w;\psi)$. The joint density function for Y, X, D, Z is given by

$$f(y, x, d, z;\psi) = f(w;\psi)p(w;\psi)^d(1-p(w;\psi))^{1-d}f(z|w;\psi)^{1-d}$$

The resulting score function is then given by

$$S_{\psi}(d,w,z) = (1-d)\frac{\partial}{\partial\psi}\log f(z|w;\psi) + \frac{d-p(w;\psi)}{p(w;\psi)(1-p(w;\psi))}\frac{\partial}{\partial\psi}p(w;\psi) + \frac{\partial}{\partial\psi}\log f(w;\psi)$$

Using results in Newey (1990), the tangent space is

$$\{(1-d)\frac{\partial}{\partial\psi}\log f(z|w;\psi) + l(w)(d-p(w;\psi)) + \frac{\partial}{\partial\psi}\log f(w;\psi)\}$$

where $l(w) \in L_2(\mathcal{W})$.

The moment condition (2.1) is equivalent to the requirement that for any matrix $\mathcal{E}$ of $d_{v+1} \times d_{v+t}$ the exactly identified system of moment conditions hold $\mathcal{E}\mathbb{E}(\widetilde{Z}_i\epsilon_i) = 0$. Differentiating under the integral gives

$$\frac{\partial\theta(\psi)}{\partial\psi} = -\left(\mathcal{E}Q\right)^{-1}\mathbb{E}\left(\mathcal{E}\widetilde{Z}_i\epsilon_i\frac{\partial\log f(W_i,Z_i;\psi)}{\partial\psi'}\right)$$

where $Q = \mathbb{E}(\widetilde{Z}_i X_i^{*'})$. Therefore, any regular estimator of θ is asymptotically linear with influence function of the form

$$-\left(\mathcal{E}Q\right)^{-1}\mathcal{E}\widetilde{Z}_i\epsilon_i$$

Hence for any given matrix $\mathcal{E}$, the projection of the above influence function onto the tangent set is

$$-\left(\mathcal{E}Q\right)^{-1}\left(\frac{1-D_i}{1-p(W_i)}\widetilde{Z}_i\epsilon_i + \frac{D_i-p(W_i)}{1-p(W_i)}\mathbb{E}(\widetilde{Z}_i|W_i)\epsilon_i\right) = -\left(\mathcal{E}Q\right)^{-1}\widetilde{\mathcal{Z}}_i\epsilon_i$$

Since $\mathbb{E}(\widetilde{\mathcal{Z}}_i\epsilon_i S_{\psi}(D_i,W_i,Z_i)') = \mathbb{E}(\widetilde{Z}_i\epsilon_i\frac{\partial\log f(W_i,Z_i;\psi)}{\partial\psi'})$. Hence the asymptotic variance corresponding to the efficient influence function for given $\mathcal{E}$ is

$$\left(\mathcal{E}Q\right)^{-1}\mathcal{E}\Omega_{eff}\mathcal{E}'(Q'\mathcal{E}')^{-1} \tag{A.2}$$

where $Q = \mathbb{E}(\widetilde{Z}_i X_i^{*'})$ and

$$\Omega_{eff} = \mathbb{E}\left(\frac{1}{1-p(W_i)}\mathbb{E}(\widetilde{Z}_i\widetilde{Z}_i'|W_i)\epsilon_i^2 - \frac{p(W_i)}{1-p(W_i)}\mathbb{E}(\widetilde{Z}_i|W_i)\mathbb{E}(\widetilde{Z}_i|W_i)'\epsilon_i^2\right)$$

(A.2) is minimized at $\mathcal{E} = Q'\Omega_{eff}^{-1}$. Thus the asymptotic efficiency becomes

$$\left(Q'\Omega_{eff}^{-1}Q\right)^{-1}$$

□

A.2 Implementation Algorithms

1 Create the indicator of missingness for each observation,

$$D = \begin{cases} 1 & \text{if } \textit{instrument} \text{ is missing} \\ 0 & \text{if } \textit{instrument} \text{ is observed} \end{cases}$$

2 Do a logit or probit estimation for propensity score, or in general, let

$$\mathbb{P}(D = 1|X, Y) = p(X, Y; \psi)$$

We would maximize

$$\Pi_{i=1}^n \left\{p(w_i, \psi)\right\}^{d_i} \left\{1 - p(w_i, \psi)\right\}^{1-d_i}$$

denote by $\widehat{\psi}_n$.

3 For $D = 0$ sample, regress Z on W, get predicted value for the whole sample

$$\widehat{Z} = W'\widehat{\xi}$$

4 Form the new instrument by

$$\mathcal{Z}_i = \frac{1-d_i}{1-p(d_i, \widehat{\psi}_n)} z_i + \frac{d_i - p(d_i, \widehat{\psi}_n)}{1-p(d_i, \widehat{\psi}_n)} \widehat{z}_i$$

5 Do 2SLS, GMM, LIML, etc. based on the new IV.

Appendix B

Appendix for Chapter 2

B.0.1 Proof of Theorem 4

Proof. Let $\Upsilon = \frac{1}{n}\sum_{i=1}^{n}(\Xi_0)^{-1}\mathcal{Z}_i\nu_i$, where $\Xi_0 = diag(\zeta_1, ..., \zeta_t)$. The specified penalty loadings in (2.11) will imply that

$$\mathbb{P}(\sqrt{n}||\Upsilon||_\infty \leq \Phi^{-1}(1 - \frac{\phi_n}{2t})) \geq 1 - \phi_n$$

$$\mathbb{P}(\frac{2\lambda}{n} \geq C_0||\Upsilon||_\infty) \to 1$$

where $\Phi(\cdot)$ is the CDF of a standard normal distribution, and C_0 is some positive constant.

By Belloni et al. (2012) (BCCH, hereafter) Lemma 6 and Assumption 7.8, if $\frac{2\lambda}{n} \geq c||\Upsilon||_\infty$, where $c \equiv \frac{1}{2}C_0$, and $\widehat{\Xi}$ satisfies Assumption 9 with $u \geq 1 \geq l > \frac{1}{c}$, then

$$\sqrt{\frac{1}{n}\sum_{1}^{n}[\mathcal{Z}_i'(\widetilde{\tau} - \tau_0)]^2} \leq (u + \frac{1}{c})\frac{\lambda\sqrt{s}}{n\kappa_{c_0}} + 2c_s$$

$$||\Xi_0(\widehat{\tau} - \tau_0)||_1 \leq 3c_0\frac{\sqrt{s}}{\kappa_{2c_0}}\left((u + \frac{1}{c})\frac{\lambda\sqrt{s}}{n\kappa_{c_0}} + 2c_s\right) + \frac{3c_0 n}{\lambda}c_s^2$$

where $\kappa_{c_0} = \min_{\tau\in\mathbb{R}^t:||\Xi_0\tau_{T^c}||_1\leq c_0||\Xi_0\tau||_1,||\tau||_2\neq 0} \frac{\sqrt{s}\sqrt{\frac{1}{n}\sum_1^n(\mathcal{Z}_i'\tau)^2}}{||\Xi_0\tau_T||_1}$, and $c_0 = \frac{uc+1}{lc-1}$. Note that the quantity κ_{c_0} controls the modulus of continuity between the prediction norm $\sqrt{\frac{1}{n}\sum_1^n(\mathcal{Z}_i'\tau)^2}$ and the l_1-norm $||\tau||_1$ within a restricted region.

Since the sequence $\{\phi_n\}$ is specified as $\phi_n = o_p(1)$ and $\log(\frac{1}{\phi_n}) = O_p(\log(t \vee n))$. It turns out that $\lambda = C_0\sqrt{n}\Phi^{-1}(1 - \frac{\phi_n}{2t}) = O_p(c\sqrt{n\log(\frac{t}{\phi_n})})$. $\widehat{\Xi}$ is a asymptotically valid penalty loading. From Assumption 9 we know that $\Upsilon = O_p(\sqrt{\frac{s}{n}})$. Hence

$$\sqrt{\frac{1}{n}\sum_{1}^{n}(\mathcal{Z}_i'(\widehat{\tau} - \tau_0))^2} = O_p(\frac{1}{\kappa c_0}\sqrt{\frac{s\log(\frac{t}{\phi_n})}{n}} + \sqrt{\frac{s}{n}})$$

By the arguments in Bickel et al. (2009),

$$\kappa_{c_0} = O_p(\kappa_c)$$

where $\kappa_c \geq \frac{1}{b}\kappa_{(\frac{bc}{a})}(M)$, $a = \min_{1\leq j\leq t}|\zeta_j|$ and $b = \max_{1\leq j\leq t}|\zeta_j|$. We've just showed that

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathcal{A}}^\star - \mathcal{A}_i)^2} = O_p(\sqrt{\frac{s\log(t\vee n)}{n}}).$$

By Assumption 8.1 and Cauchy-Schwartz inequality, we know that

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathcal{A}}_i - \mathcal{A}_i)^2} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathcal{A}}_i - \widehat{\mathcal{A}}_i^\star)^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathcal{A}}^\star - \mathcal{A}_i)^2} = O_p(\sqrt{\frac{s\log(t\vee n)}{n}})$$

For the l_1-rate, by Assumption 7, Assumption 8 and Lemma 6 in Appendix C from BCCH,

$$||\widehat{\tau} - \tau||_1 \leq ||\widehat{\tau} - \widehat{\tau}^\star||_1 + ||\widehat{\tau}^\star - \tau||_1$$

Now by Theorem 4 of BCCH, Theorem 3 is established.

B.0.2 Proof of Theorem 5

Proof. Let $\theta = (\alpha, \beta)'$. I follow closely the structure of semiparametric efficiency bound derivation for general missing data problems of Tsiatis (2007).

Let $S \equiv (W, Z)$ be the full data, where $W = (X, Y)$. Introduce $G_D(S)$ as a mapping from S to a subset of its elements whenever $D = d$, i.e. $G_0(S) = S$; $G_1(S) = W$. The full-data Hilbert space $\mathcal{H}^F$ consists of all two-dimensional, mean-zero measurable functions of S with finite variance equipped with the covariance inner product. While the observed data Hilbert space $\mathcal{H}$ is the space of all all two-dimensional, mean-zero measurable functions of $\{D, G_D(S)\}$ equipped with the covariance inner product. First assume that the missing probabilities are unknown and are modeled as $\mathbb{P}(D = d|S = s) = \varpi(d, G_d(S), \psi)$, $d = 0, 1$, where ψ is an unknown parameter. The probability of missing depends on S only through the observed data, due to (1). The joint density for the full data is parametrized as $p_S(\cdot, \theta, \eta)$, where η is unknown. The observed-data likelihood for a single observation is

$$\begin{aligned}\int_{\{s:G_d(s)=g_d\}} \mathbb{P}(D = d|S = s, \psi) p_S(s, \theta, \eta) d\nu_S(s) &= \int_{\{s:G_d(s)=g_d\}} \varpi(d, G_d(S), \psi) p_S(s, \theta, \eta) d\nu_S(s) \\ &= \varpi(d, G_d(S), \psi) \int_{\{s:G_d(s)=g_d\}} p_S(s, \theta, \eta) d\nu_S(s)\end{aligned}$$

where $\nu_S(\cdot)$ is a dominating measure for S. The log-likelihood for a single observation is given by

$$\log \varpi(d, G_d(S), \psi) + \log \int_{\{s:G_d(s)=g_d\}} p_S(s, \theta, \eta) d\nu_S(s)$$

The two nuisance parameters η and ψ are separated in the log-likelihood, hence the nuisance tangent space Λ will be the direct sum of two spaces. Let Λ_ψ and Λ_η be the spaces generated by the score vector w.r.t. ψ and η, respectively. Hence

$$\Lambda = \Lambda_\psi \oplus \Lambda_\eta$$

More formally, Λ_ψ is defined as

$$\Lambda_\psi \equiv \left[B^{2\times q} S_\psi^{q\times 1}\{D, G_D(S), \psi_0\}\text{for all } B^{2\times q}\right]$$

where B is any constant $2 \times q$ matrix and

$$S_\psi^{q\times 1} = \frac{\partial \log \varpi(D, G_D(S), \psi_0)}{\partial \psi}$$

By Theorem 7.1 of Tsiatis (2007), the nuisance tangent space Λ_η can be expressed as

$$\Lambda_\eta = \mathbb{E}\{\Lambda^F | D, G_D(S)\}$$

where Λ^F denotes the full-data nuisance tangent space.

The orthogonality of the two spaces is shown in Theorem 8.2, i.e. $\Lambda_\psi \perp \Lambda_\eta$. Hence a typical element of $\Lambda^\perp$ can be found by taking an arbitrary element $h \in \Lambda_\eta^\perp$ and computing

$$h - \Pi(h|\Lambda_\psi) = \Pi(h|\Lambda_\psi^\perp) \tag{B.1}$$

Under the assumption that $\varpi(d, G_d(S), \psi) \in (0, 1)$, it is shown in Theorem 7.2 of Tsiatis (2007) that the space $\Lambda_\eta^\perp$ consists of all elements that can be written as

$$\frac{(1-D)\varphi^F(S)}{\varpi(0, G_0(S), \psi)} + \underbrace{\frac{1-D}{\varpi(0, G_0(S), \psi)}\varpi\{1, G_1(S), \psi\}L\{G_1(S)\} - DL_2\{G_1(S)\}}_{\text{augmentation term}} \tag{B.2}$$

where $L\{G_1(S)\}$ is an arbitrary two-dimensional measurable function of $G_1(S)$ and $\varphi^F(S)$ is an arbitrary element of $\Lambda^{F\perp}$. A by-product in deriving (B.2) is that $\Lambda_\eta^\perp$ can be written as the direct sum of two linear subspaces, i.e.

$$\Lambda_\eta^\perp = \frac{(1-D)\Lambda^{F\perp}}{\varpi(0, G_0(S), \psi)} \oplus \Lambda_2 \tag{B.3}$$

where Λ_2 consists of elements with expression of augmentation term in (B.2). Combine (B.1) and (B.3), the observed-data influence functions for θ can be written as

$$\varphi\{D, G_D(S)\} = \left\{ \left[\frac{(1-D)\varphi^F(S)}{\varpi(0, G_0(S), \psi)} + L_2\{1, G_1(S)\} \right] - \Pi\{[\cdot]|\Lambda_\psi\} \right\}$$

where $\Pi\{[\cdot]|\Lambda_\psi\}$ is the projection onto Λ_ψ. By Theorem 10.1, 10.2 in Tsiatis (2007), the optimal observed-data influence function is obtained by choosing $L_2\{1, G_1(S) = -\pi[\frac{(1-D)\varphi^F(S)}{\varpi(0,G_0(S),\psi)}|\Lambda_2]$. Furthermore, the projection onto the augmentation space Λ_2 is the unique element $\left\{\frac{p(W)-D}{1-p(W)}\right\} h_2^0(W) \in \Lambda_2$, where

$$h_2^0(W) = \mathbb{E}(\varphi^F(S)|W)$$

In addition, $\Lambda_\psi \in \Lambda_2$, the projection $\Pi\{[\cdot]|\Lambda_\psi\}$ is then absorbed into the augmentation term. Hence, the optimal observed-data influence function becomes

$$\frac{(1-D)\varphi^F(S)}{1-p(W,\psi)} - \frac{p(W,\psi)-D}{1-p(W,\psi)}\mathbb{E}(\varphi^F(S)|W)$$

for any fixed $\varphi^F(S) \in \Lambda^{F\perp}$, where $\varpi(0, G_0(S), \psi)$ is simplified as $1 - p(W, \psi)$, given that $\varpi(0, G_0(S), \psi) + \varpi(1, G_1(S), \psi) = 1$.

From e.g. Amemiya (1977), Newey (1990), a typical element in $\Lambda^{F\perp}$ is given as $A(Z)\epsilon$, where $A(Z)$ is a valid instrument. Define a linear operator $\mathcal{L}(\cdot) \in \Lambda^\perp$, $\mathcal{L}(\cdot) : L^2(Z) \to \mathcal{H}$,

$$\mathcal{L}(A(Z)) = \frac{(1-D)A(Z)\epsilon}{1-p(W,\psi)} - \frac{p(W,\psi)-D}{1-p(W,\psi)}\mathbb{E}(A(Z))|W)\epsilon \tag{B.4}$$

Our goal is to find the efficient influence function. It suffices to restrict our attention to the class of influence functions $\mathcal{L}(A(Z))$ with $A(Z)$ a valid instrument. Such class of linear functions will consist a linear subspace pf $\mathcal{H}$, denoted as $\mathcal{H}_{\text{opt}}$. The search for efficient influence function is within $\mathcal{H}_{\text{opt}}$. One implication of (B.4) is that the efficient score vectors have similar relationship as

$$S_{\text{eff}}\{D, G_D(S)\} = \frac{(1-D)B^F_{\text{eff}}(S)}{1-p(W,\psi)} - \frac{p(W,\psi)-D}{1-p(W,\psi)}\mathbb{E}(B^F_{\text{eff}}(S)|W)$$

In addition, by Lemma 11.1 in Tsiatis (2007), $B^F_{\text{eff}}(S)$ exists and is the unique solution to the equation

$$\Pi\left[\mathcal{M}^{-1}\{B^F_{\text{eff}}(S)\}|\Lambda^{F\perp}\right] = S^F_{\text{eff}}(S)$$

where $S^F_{\text{eff}}(S)$ is the efficient score when there is no missing. When assuming conditional homoskedasticity, i.e. $\mathbb{E}[\epsilon^2|Z] = \Omega$,

$$S^F_{\text{eff}}(S) = \mathcal{D}(Z)'\Omega^{-1}\epsilon$$

up to a constant matrix. The following lemma establishes $\mathcal{M}^{-1}$.

Lemma 9. *The inverse operator* $\mathcal{M}^{-1}$ *is given by*

$$a^F(S) = \mathcal{M}^{-1}\{B^F(S)\} = \frac{B^F(S)}{1-p(W,\psi)} - \frac{p(W,\psi)}{1-p(W,\psi)}\mathbb{E}(B^F(S)|W)$$

Proof. The existence and uniqueness of $\mathcal{M}^{-1}$ has been shown in Theorem 10.6, where

$$\mathcal{M}(a^F) \equiv (1-p(W,\psi))a^F + p(W,\psi)\mathbb{E}(a^F|W)$$

We only need to show that $\mathcal{M}(a^F) = B^F$. Plug in a^F, we get

$$\begin{aligned}
\mathcal{M}(a^F) &= (1-p(W,\psi))\left[\frac{B^F(S)}{1-p(W,\psi)} - \frac{p(W,\psi)}{1-p(W,\psi)}\mathbb{E}(B^F(S)|W)\right] + p(W,\psi)\mathbb{E}([\cdot]|W) \\
&= B^F - p(W,\psi)\mathbb{E}(B^F|W) + p(W,\psi)\frac{1-p(W,\psi)}{1-p(W,\psi)}\mathbb{E}(B^F|W) \\
&= B^F
\end{aligned}$$

□

And from previous literature on optimal instruments, we know that

$$\Pi[a^F(S)|\Lambda^{F\perp}] = \mathbb{E}[a^F(S)\epsilon|Z]\Omega^{-1}\epsilon$$

for any $a^F(S) \in \mathcal{H}^F$. Hence, $B^F_{\text{eff}}(S)$ is the solution to

$$\mathbb{E}\left[\left(\frac{B^F(S)}{1-p(W,\psi)} - \frac{p(W,\psi)}{1-p(W,\psi)}\mathbb{E}(B^F(S)|W)\right)\epsilon|Z\right]\Omega^{-1}\epsilon = \mathcal{D}(Z)'\Omega^{-1}\epsilon$$

$$\Rightarrow \mathbb{E}\left[\left(\frac{B^F(S)}{1-p(W,\psi)} - \frac{p(W,\psi)}{1-p(W,\psi)}\mathbb{E}(B^F(S)|W)\right)\epsilon|Z\right] = \mathcal{D}(Z)' \qquad \text{(B.5)}$$

Since $B^F_{\text{eff}}(S)$ has the form $A(Z)\epsilon$, (B.5) becomes

$$\mathbb{E}\left[\left(\frac{A(Z)}{1-p(W,\psi)} - \frac{p(W,\psi)}{1-p(W,\psi)}\mathbb{E}(A(Z)|W)\right)\epsilon^2|Z\right] = \mathcal{D}(Z)'$$

We denote the solution to this equation as $A^*(Z)$. Hence the observed-data efficient score vector is

$$\begin{aligned}
S_{\text{eff}}\{D, G_D(S)\} &= \frac{(1-D)A^*(Z)\epsilon}{1-p(W,\psi)} - \frac{p(W,\psi)-D}{1-p(W,\psi)}\mathbb{E}(A^*(Z)|W)\epsilon \\
&= \left[\frac{(1-D)A^*(Z)}{1-p(W,\psi)} - \frac{p(W,\psi)-D}{1-p(W,\psi)}\mathbb{E}(A^*(Z)|W)\right]\epsilon
\end{aligned}$$

The efficient influence functions is then

$$\varphi_{\text{eff}}(S) = -\left(\mathbb{E}\left[A^*(Z)\mathcal{D}(Z)\right]\right)^{-1} S_{\text{eff}}\{D, G_D(S)\}$$

The asymptotic covariance matrix for θ is $\mathbb{E}\left[\varphi_{\text{eff}}(S)\varphi_{\text{eff}}(S)'\right]$, i.e.

$$(\mathbb{E}\left[A^*(Z)\mathcal{D}(Z)\right])^{-1}\,\mathbb{E}\left[S_{\text{eff}}\{D,G_D(S)S_{\text{eff}}\{D,G_D(S)'\right]\left(\mathbb{E}\left[A^*(Z)'\mathcal{D}(Z)'\right]\right)^{-1}$$

□

B.0.3 Proof of Lemma 4

Given the existence and uniqueness of the optimal instruments, by Theorem 2.3.1(Successive Approximation) from Zemyan (2012), the resolvent kernel

$$R(z,t;\widetilde{\Omega}^{-1})=\sum_{m=1}^{\infty}\widetilde{\Omega}^{1-m}K_m(z,t)$$

where $K_1(z,t)=K(z,t)$, and for $m\geq 2$, $K_m(z,t)=\int K_{m-1}(z,s)K(s,t)ds$. We proceed the proof by induction. For $m=2$,

$$\begin{aligned}
K_2(z,t) &= \int K_1(z,s)K(s,t)ds \\
&= \int\left[\int Q(w)f_{W|Z}(w|z)f_{Z|W}(s|w)dw\int Q(v)f_{W|Z}(v|s)f_{Z|W}(t|v)dv\right]ds \\
&= \int\int Q(w)Q(v)f_{W|Z}(w|z)\left(\int f_{Z|W}(s|w)f_{W|Z}(v|s)ds\right)f_{Z|W}(t|v)dwdv \quad \text{(Fubini's)} \\
&= \int\int Q(s_1)Q(s_2)f_{W|Z}(s_1|z)\left(\mathbb{E}[f_{W|Z}(s_2|Z)|W=s_1]\right)f_{Z|W}(t|s_2)ds_1ds_2
\end{aligned}$$

Suppose for $m=n$,

$$\begin{aligned}
&K_n(z,t) \\
&\quad=\int Q(s_n)f_{W|Z}(s_1|z)f_{Z|W}(t|s_n)\prod_{i=1}^{n-1}Q(s_i)\mathbb{E}\left[f_{W|Z}(s_{i+1}|Z)|W=s_i\right]ds^n
\end{aligned}$$

Then for $m = n + 1$,

$$
\begin{aligned}
K_{n+1}(z,t) &= \int K_n(z,s)K(s,t)ds \\
&= \int \left[\int Q(s_n) f_{W|Z}(s_1|z) f_{Z|W}(s|s_n) \prod_{i=1}^{n-1} Q(s_i)\mathbb{E}\left[f_{W|Z}(s_{i+1}|Z)|W = s_i \right] ds^n \right] \times ... \\
&\quad ... \times \left[\int Q(s_{n+1}) f_{W|Z}(s_{n+1}|s) f_{Z|W}(t|s_{n+1}) ds_{n+1} \right] ds \\
&= \int Q(s_{n+1})Q(s_n) f_{W|Z}(s_1|z) \left[\int f_{Z|W}(s|s_n) f_{W|Z}(s_{n+1}|s) ds \right] \times ... \\
&\quad ... \times \prod_{i=1}^{n-1} Q(s_i)\mathbb{E}\left[f_{W|Z}(s_{i+1}|Z)|W = s_i \right] f_{Z|W}(t|s_{n+1}) ds^{n+1} \\
&= \int Q(s_{n+1}) f_{W|Z}(s_1|z) f_{Z|W}(t|s_{n+1}) \times ... \\
&\quad ... \times \prod_{i=1}^{n} Q(s_i)\mathbb{E}\left[f_{W|Z}(s_{i+1}|Z)|W = s_i \right] ds^{n+1}
\end{aligned}
$$

with $Q(s_i) = \frac{p(s_i)}{1-p(s_i)}\epsilon^2$. □

Appendix C

Appendix for Chapter 3

C.0.1 Proof of Lemma 6

Proof. First, we differentiate $Q_0(y_0, y_1)$ with respect to y_0. Noting that $\partial\mathbb{E}|W - w|/\partial w = 2F_W(w) - 1$ for a continuous distribution $F_W(\cdot)$, we obtain

$$\begin{aligned} & \frac{\partial}{\partial y_0}\mathbb{E}\big[|Y - y_0|(1 - D)|X = x, Z = z\big] \\ = & \frac{\partial}{\partial y_0}\mathbb{E}\big(|Y - y_0|\big|D = 0, X = x, Z = z\big) \times \Pr(D = 0|X = x, Z = z) \\ = & 2\Pr(Y \leq y_0; D = 0|X = x, Z = z) - \Pr(D = 0|X = x, Z = z). \end{aligned}$$

Moreover, we have

$$\begin{aligned} & \mathbb{E}[\text{sign}(Y - y_1) \cdot D|X = x, Z = z] \\ & \qquad = -2\Pr(Y \leq y_1; D_1 = 1|X = x, Z = z) + \Pr(D = 1|X = x, Z = z). \end{aligned}$$

It follows that

$$\begin{aligned} \frac{\partial}{\partial y_0}Q_0(y_0, y_1) &= 2[\Pr(Y \leq y_0; D = 0|X = x, Z = 0) - \Pr(Y \leq y_0; D = 0|X = x, Z = 1)] \\ &+ 2[\Pr(Y \leq y_1; D = 1|X = x, Z = 0) - \Pr(Y \leq y_1; D = 1|X = x, Z = 1)] \\ &= 2[p(x, 1) - p(x, 0)] \times [C_{0x}(y_0) - C_{1x}(y_1)], \end{aligned}$$

where the last step comes from the definition of C_{dx} in (3.4). Fix $y_1 \in \mathbb{R}$. Note that $C_{0x}(\cdot)$ is weakly increasing on $\mathbb{R}$ and strictly increasing on $\mathscr{C}^o_{dx} = \mathscr{S}^\circ_{Y|D=0,X=x}$

by Theorem 1. Moreover, because $p(x,0) < p(x,1)$,[1] then $Q_0(\cdot, y_1)$ has a weakly and strictly increasing derivative on $\mathbb{R}$ and $\mathscr{C}^o_{dx}$, respectively. Therefore, $Q_0(\cdot, y_1)$ is weakly and strictly convex on $\mathbb{R}$ and $\mathscr{C}^o_{dx}$, respectively, for arbitrary $y_1 \in \mathbb{R}$. Furthermore, if $y_1 \in \mathscr{S}^\circ_{Y|D=1,X=x}$, we have $C_{0x}(y_0) = C_{1x}(y_1)$ if and only if $y_0 = \phi_{0x}(y_1)$ by Theorem 6. Thus, $y_0 = \phi_{0x}(y_1)$ uniquely solves the first–order condition $\frac{\partial}{\partial y_0} Q_0(y_0, y_1) = 0$ whenever $y_1 \in \mathscr{S}^\circ_{Y|D=1,X=x}$. A similar argument also applies to the population objective function $Q_1(y_0, \cdot)$. □

C.0.2 Proof of Theorem 7

Proof. Fix $X = x$. All the following argument is conditional on $X = x$. For simplicity, we suppress the dependence on x, e.g., we use ϕ_d for ϕ_{dx}, omit the term $\mathbb{1}(X_i = x)$ in the estimation, and $X = x$ in the conditional probability $\Pr(Y \leq y; D = d|X = x; Z = z)$. Moreover, we only show the results for $d = 0$. The proof for the case $d = 1$ can be derived similarly.

First, we show uniform consistency. By Angrist et al. (2006), it suffices to show that $\sup_{(y_0,y_1)\in\mathcal{B}} \|\hat{Q}_0(y_0, y_1) - Q_0(y_0, y_1)\| = o_p(1)$ for any compact set $\mathcal{B} \subset \mathbb{R}^2$. By the law of large number, we have pointwise convergence, i.e., $\|\hat{Q}_0(y_0, y_1) - Q_0(y_0, y_1)\| = o_p(1)$. Then, it suffices to show the stochastic equicontinuity of the empirical process $\hat{\rho}_0(\cdot, \cdot; z) - \rho_0(\cdot, \cdot; z)$, which directly follows the general argument in Koenker and Xiao (2002). Next, we establish the limiting distribution of the

[1] When such a rank of $p(x, z)$ is unknown, we can modify the objective function by $\tilde{Q}_0(y_0, y_1) = [p(x, 1) - p(x, 0)] \times Q_0(y_0, y_1)$. The additional term $p(x, 1) - p(x, 0)$ changes the sign of $\tilde{Q}_0(\cdot, y_1)$ based on the relative rank of $p(x, z)$ while its scale does not matter for the optimization of $\tilde{Q}_0(\cdot, y_1)$.

process.

Taking the directional derivative, we have

$$\frac{d}{dt}\hat{Q}_0(y_0+t,y_1)\Big|_{t\downarrow 0} = \frac{2\sum_{i=1}^n \mathbb{1}(Y_i \leq y_0; D_i = 0; Z_i = 0)}{\sum_{i=1}^n \mathbb{1}(Z_i = 0)} - \frac{2\sum_{i=1}^n \mathbb{1}(Y_i \leq y_0; D_i = 0; Z_i = 1)}{\sum_{i=1}^n \mathbb{1}(Z_i = 1)}$$
$$+ \frac{2\sum_{i=1}^n \mathbb{1}(Y_i \leq y_1; D_i = 1; Z_i = 0)}{\sum_{i=1}^n \mathbb{1}(Z_i = 0)} - \frac{2\sum_{i=1}^n \mathbb{1}(Y_i \leq y_1; D_i = 1; Z_i = 1)}{\sum_{i=1}^n \mathbb{1}(Z_i = 1)} + \xi_n(y_0).$$

where the remainder term $\xi_n(y_0)$ is bounded by

$$\frac{n \cdot \sum_{i=1}^n \mathbb{1}(Y_i = y_0)}{\sum_{i=1}^n \mathbb{1}(Z_i = 0) \times \sum_{i=1}^n \mathbb{1}(Z_i = 1)}.$$

By the computational properties of linear programming in Koenker and Bassett (1978, Theorem 3.3), we have $\xi_n(y_0) = O_p(n^{-1})$ uniformly in $y_0 \in \mathbb{R}$. We can derive a similar expression for $\frac{d}{dt}\hat{Q}_0(y_0 - t, y_1)\Big|_{t\downarrow 0}$. Note that $\frac{d}{dt}\hat{Q}_0(\hat{\phi}_0(y_1) + t, y_1)\Big|_{t\downarrow 0} \geq 0$ and $\frac{d}{dt}\hat{Q}_0(\hat{\phi}_0(y_1) - t, y_1)\Big|_{t\downarrow 0} \geq 0$ as $\hat{\phi}_0(y_1)$ minimizes $\hat{Q}_0(\cdot, y_1)$. Hence, we have

$$\frac{\sum_{i=1}^n \mathbb{1}(Y_i \leq \hat{\phi}_0(y_1); D_i = 0; Z_i = 0)}{\sum_{i=1}^n \mathbb{1}(Z_i = 0)} - \frac{\sum_{i=1}^n \mathbb{1}(Y_i \leq \hat{\phi}_0(y_1); D_i = 0; Z_i = 1)}{\sum_{i=1}^n \mathbb{1}(Z_i = 1)}$$
$$+ \frac{\sum_{j=1}^n \mathbb{1}(Y_i \leq y_1; D_i = 1; Z_i = 0)}{\sum_{i=1}^n \mathbb{1}(Z_i = 0)} - \frac{\sum_{j=1}^n \mathbb{1}(Y_i \leq y_1; D_i = 1; Z_i = 1)}{\sum_{i=1}^n \mathbb{1}(Z_i = 1)} = O_p(n^{-1})$$

uniformly in y_1.

Following the convention, we introduce some notation from the empirical process literature: For $W = (Y, D, Z)'$ and a generic function g, let $\mathbb{E}_n[g(W)] = n^{-1}\sum_{i=1}^n g(W_i)$ and $\mathbb{G}_n[g(W)] = n^{-1/2}\sum_{i=1}^n \left\{g(W_i) - \mathbb{E}[g(W_i)]\right\}$. Hence, the above condition can be rewritten as

$$\sqrt{n}\left\{\frac{\mathbb{E}_n \mathbb{1}(Y \leq \hat{\phi}_0(y_1); D = 0; Z = 0)}{\mathbb{E}_n \mathbb{1}(Z = 0)} + \frac{\mathbb{E}_n \mathbb{1}(Y \leq y_1; D = 1; Z = 0)}{\mathbb{E}_n \mathbb{1}(Z = 0)}\right\}$$
$$- \sqrt{n}\left\{\frac{\mathbb{E}_n \mathbb{1}(Y \leq \hat{\phi}_0(y_1); D = 0; Z = 1)}{\mathbb{E}_n \mathbb{1}(Z = 1)} + \frac{\mathbb{E}_n \mathbb{1}(Y \leq y_1; D = 1; Z = 1)}{\mathbb{E}_n \mathbb{1}(Z = 1)}\right\} = o_p(1)$$

uniformly in $y_1 \in \mathbb{R}$. It follows that

$$\begin{aligned}
&\frac{\sqrt{n}\ \mathbb{E}\Big\{\mathbb{1}(Y \le \hat{\phi}_0(y_1); D = 0; Z = 0) + \mathbb{1}(Y \le y_1; D = 1; Z = 0)\Big\}}{\mathbb{E}_n \mathbb{1}(Z = 0)} \\
&\qquad - \frac{\sqrt{n}\ \mathbb{E}\Big\{\mathbb{1}(Y \le \hat{\phi}_0(y_1); D = 0; Z = 1) + \mathbb{1}(Y \le y_1; D = 1; Z = 1)\Big\}}{\mathbb{E}_n \mathbb{1}(Z = 1)} \\
&\qquad + \frac{\mathbb{G}_n\Big[\mathbb{1}(Y \le \hat{\phi}_0(y_1); D = 0; Z = 0) + \mathbb{1}(Y \le y_1; D = 1; Z = 0)\Big]}{\mathbb{E}_n \mathbb{1}(Z = 0)} \\
&\quad - \frac{\mathbb{G}_n\Big[\mathbb{1}(Y \le \hat{\phi}_0(y_1); D = 0; Z = 1) + \mathbb{1}(Y \le y_1; D = 1; Z = 1)\Big]}{\mathbb{E}_n \mathbb{1}(Z = 1)} = o_p(1). \quad \text{(C.1)}
\end{aligned}$$

Because $\mathbb{E}_n \mathbb{1}(Z = z) = \Pr(Z = z) + O_p(n^{-1/2})$, then by Taylor expansion,

$$\frac{1}{\mathbb{E}_n \mathbb{1}(Z = z)} = \frac{1}{\Pr(Z = z)} - \frac{1}{\Pr^2(Z = z)} \times [\mathbb{E}_n \mathbb{1}(Z = z) - \Pr(Z = z)] + O_p(n^{-1}).$$

Thus,

$$\begin{aligned}
&\frac{\sqrt{n}\ \mathbb{E}\Big\{\mathbb{1}(Y \le \hat{\phi}_0(y_1); D = 0; Z = z) + \mathbb{1}(Y \le y_1; D = 1; Z = z)\Big\}}{\mathbb{E}_n \mathbb{1}(Z = z)} \\
= &\ \sqrt{n}\ \mathbb{E}\big\{\mathbb{1}(Y \le \hat{\phi}_0(y_1); D = 0) + \mathbb{1}(Y \le y_1; D = 1)\big| Z = z\big\} \\
- &\ \mathbb{E}\Big\{\mathbb{1}(Y \le \hat{\phi}_0(y_1); D = 0) + \mathbb{1}(Y \le y_1; D = 1)\Big| Z = z\Big\} \times \frac{\mathbb{G}_n \mathbb{1}(Z = z)}{\Pr(Z = z)} + o_p(1) \\
= &\ \sqrt{n}\ \mathbb{E}\big\{\mathbb{1}(Y \le \hat{\phi}_0(y_1); D = 0) + \mathbb{1}(Y \le y_1; D = 1)\big| Z = z\big\} \\
- &\ \mathbb{E}\Big\{\mathbb{1}(Y \le \phi_0(y_1); D = 0) + \mathbb{1}(Y \le y_1; D = 1)\Big| Z = z\Big\} \times \frac{\mathbb{G}_n \mathbb{1}(Z = z)}{\Pr(Z = z)} + o_p(1)
\end{aligned}$$

where the last $o_p(1)$ term is uniform in y_1 due to the uniform convergence of $\hat{\phi}_0$ to ϕ_0.

Let $\varphi(\cdot, y_1) = \mathbb{1}(Y \leq \cdot; D = 0) + \mathbb{1}(Y \leq y_1; D = 1)$. Therefore, (C.1) implies

$$\begin{aligned}
&\sqrt{n}\, \mathbb{E}\big[\varphi(\hat{\phi}_0(y_1), y_1)|Z = 0\big] - \sqrt{n}\, \mathbb{E}\big[\varphi(\hat{\phi}_0(y_1), y_1)|Z = 1\big] \\
= \quad & -\frac{\mathbb{G}_n \left[\varphi(\hat{\phi}_0(y_1), y_1) \times \mathbb{1}(Z = 0)\right]}{\mathbb{E}_n \mathbb{1}(Z = 0)} + \frac{\mathbb{G}_n \left[\varphi(\hat{\phi}_0(y_1), y_1) \times \mathbb{1}(Z = 1)\right]}{\mathbb{E}_n \mathbb{1}(Z = 1)} \\
+ \quad & \frac{\mathbb{E}\left[\varphi(\phi_0(y_1), y_1)|Z = 0\right]}{\Pr(Z = 0)} \times \mathbb{G}_n \mathbb{1}(Z = 0) - \frac{\mathbb{E}\left[\varphi(\phi_0(y_1), y_1)|Z = 1\right]}{\Pr(Z = 1)} \times \mathbb{G}_n \mathbb{1}(Z = 1) + o_p(1).
\end{aligned}$$

Note that $\mathbb{E}\left[\varphi(\phi_0(y_1), y_1)|Z = z\right] = R_1(y_1)$ which does not depend on z. Hence,

$$\begin{aligned}
&\sqrt{n}\, \mathbb{E}\big[\varphi(\hat{\phi}_0(y_1), y_1)|Z = 0\big] - \sqrt{n}\, \mathbb{E}\big[\varphi(\hat{\phi}_0(y_1), y_1)|Z = 1\big] \\
&\quad = -\frac{\mathbb{G}_n \left[\varphi(\hat{\phi}_0(y_1), y_1) \times \mathbb{1}(Z = 0)\right]}{\mathbb{E}_n \mathbb{1}(Z = 0)} + \frac{\mathbb{G}_n \left[\varphi(\hat{\phi}_0(y_1), y_1) \times \mathbb{1}(Z = 1)\right]}{\mathbb{E}_n \mathbb{1}(Z = 1)} \\
&\qquad + \frac{R_1(y_1)}{\Pr(Z = 0)} \times \mathbb{G}_n \mathbb{1}(Z = 0) - \frac{R_1(y_1)}{\Pr(Z = 1)} \times \mathbb{G}_n \mathbb{1}(Z = 1) + o_p(1).
\end{aligned}$$

Moreover, the derivative of $\mathbb{E}\left[\varphi(\cdot, y_1)|Z = z\right]$ is the derivative of $\Pr(Y \leq \cdot; D = 0|Z = z)$. Thus, using (3.4) and the definition of $c^*_{dx}(\cdot)$, a Taylor expansion gives

$$\sqrt{n}\, \mathbb{E}\big[\varphi(\hat{\phi}_0(y_1), y_1)|Z = 0\big] - \sqrt{n}\, \mathbb{E}\big[\varphi(\hat{\phi}_0(y_1), y_1)|Z = 1\big] = c^*_0(\tilde{\phi}_0(y_1)) \times \sqrt{n}\, [\hat{\phi}_0(y_1) - \phi_0(y_1)]$$

where $\tilde{\phi}_0(y_1)$ is between $\phi_0(y_1)$ and $\hat{\phi}_0(y_1)$. Note that $c^*_0(\tilde{\phi}_0(y_1)) = c^*_0(\phi_0(y_1)) + o_p(1)$ uniformly in y_1. It follows that

$$\begin{aligned}
&[c^*_0(\phi_0(y_1)) + o_p(1)] \times \sqrt{n}\, [\hat{\phi}_0(y_1) - \phi_0(y_1)] \\
&\quad = -\frac{\mathbb{G}_n \left[\varphi(\hat{\phi}_0(y_1), y_1) \times \mathbb{1}(Z = 0)\right]}{\mathbb{E}_n \mathbb{1}(Z = 0)} + \frac{\mathbb{G}_n \left[\varphi(\hat{\phi}_0(y_1), y_1) \times \mathbb{1}(Z = 1)\right]}{\mathbb{E}_n \mathbb{1}(Z = 1)} \\
&\qquad + \frac{R_1(y_1)}{\Pr(Z = 0)} \times \mathbb{G}_n \left[\mathbb{1}(Z = 0)\right] - \frac{R_1(y_1)}{\Pr(Z = 1)} \times \mathbb{G}_n \left[\mathbb{1}(Z = 1)\right] + o_p(1).
\end{aligned}$$

Because φ is Donsker, by the empirical process theorem (see e.g. Van Der Vaart and Wellner, 1996), we have the equicontinuity of the function class $\varphi(\cdot, \cdot)$. Hence, uniformly in y_1,

$$\mathbb{G}_n\left[\varphi(\hat{\phi}_0(y_1), y_1) \times \mathbb{1}(Z = z)\right] = \mathbb{G}_n\left[\varphi(\phi_0(y_1), y_1) \times \mathbb{1}(Z = z)\right] + o_p(1),$$

which converges to a zero-mean Gaussian process. Thus, we obtain

$$\begin{aligned}&[c_0^*(\phi_0(y_1)) + o_p(1)] \times \sqrt{n}\ [\hat{\phi}_0(y_1) - \phi_0(y_1)]\\ &\quad = -\mathbb{G}_n\left\{[\varphi(\phi_0(y_1), y_1) - R_1(y_1)] \times \left[\frac{\mathbb{1}(Z = 0)}{\Pr(Z = 0)} - \frac{\mathbb{1}(Z = 1)}{\Pr(Z = 1)}\right]\right\} + o_p(1) \qquad \text{(C.2)}\end{aligned}$$

where the right–hand side converges to a zero-mean Gaussian process. Therefore, $c_0^*(\phi_0(\cdot)) \times \sqrt{n}[\hat{\phi}_0(\cdot) - \phi_0(\cdot)]$ converges in distribution to a zero-mean Gaussian process.

Its covariance kernel $\Sigma_0(y, y')$ for $y \leq y'$ is obtained as

$$\begin{aligned}\Sigma_0(y, y') &= \mathbb{E}\left\{[\varphi(\phi_0(y), y) - R_1(y)] \times [\varphi(\phi_0(y'), y') - R_1(y')] \times \left[\frac{\mathbb{1}(Z = 0)}{\Pr(Z = 0)} - \frac{\mathbb{1}(Z = 1)}{\Pr(Z = 1)}\right]^2\right\}\\ &= \mathbb{E}\left\{[\varphi(\phi_0(y), y) - R_1(y)] \times \varphi(\phi_0(y'), y') \times \left[\frac{\mathbb{1}(Z = 0)}{\Pr(Z = 0)} - \frac{\mathbb{1}(Z = 1)}{\Pr(Z = 1)}\right]^2\right\}\\ &= \mathbb{E}\left\{\left[\varphi(\phi_0(y), y) - R_1(y)\varphi(\phi_0(y'), y')\right] \times \left[\frac{\mathbb{1}(Z = 0)}{\Pr(Z = 0)} - \frac{\mathbb{1}(Z = 1)}{\Pr(Z = 1)}\right]^2\right\}\\ &= \left[R_1(y) - R(y)R_1(y')\right] \times \mathbb{E}\left[\frac{\mathbb{1}(Z = 0)}{\Pr(Z = 0)} - \frac{\mathbb{1}(Z = 1)}{\Pr(Z = 1)}\right]^2\end{aligned}$$

where the second and third equalities use the definition of $\varphi(\phi_0(y_1), y_1)$, and the fourth equality uses $\mathbb{E}\left[\varphi(\phi_0(y_1), y_1)|Z = z\right] = R_1(y_1)$. The expression for $\Sigma_0(y, y')$ given in the theorem follows upon noting that

$$\mathbb{E}\left[\left(\frac{\mathbb{1}(Z = 0)}{\Pr(Z = 0)} - \frac{\mathbb{1}(Z = 1)}{\Pr(Z = 1)}\right)^2\right] = \frac{1}{\Pr(Z = 0)\Pr(Z = 1)}. \quad \square$$

C.0.3 Proof of Theorem 8

Proof. We have $\hat{f}_\Delta(\delta) - f_\Delta(\delta) = [\hat{f}_\Delta(\delta) - \tilde{f}_\Delta(\delta)] + \tilde{f}_\Delta(\delta) - f_\Delta(\delta)$, where

$$\tilde{f}_\Delta(\delta) = \frac{1}{nh}\sum_{i=1}^{n} K\big(\frac{\Delta_i - \delta}{h}\big), \quad \forall \delta \in [\underline{\delta} + h, \overline{\delta} - h],$$

is the infeasible kernel estimator of $f_\Delta(\delta)$. From standard kernel estimation, we have

$$\sup_{\delta \in [\underline{\delta}+h, \overline{\delta}-h]} |\tilde{f}_\Delta(\delta) - f_\Delta(\delta)| = O_p\big(h^P\big)$$

since $h = (\ln n/n)^{\frac{1}{2P+2}}$ leads to oversmoothing. Thus, it suffices to show that the same uniform convergence rate holds for $|\hat{f}_\Delta(\delta) - \tilde{f}_\Delta(\delta)|$. We actually show that

$$\sup_{\delta \in [\underline{\delta}+h, \overline{\delta}-h]} |\hat{f}_\Delta(\delta) - \tilde{f}_\Delta(\delta)| = o_p\big(h^P\big)$$

so that the first step estimation error is negligible given our choice of bandwidth.

From a second-order Taylor expansion we have

$$\hat{f}_\Delta(\delta) - \tilde{f}_\Delta(\delta) = \frac{1}{nh^2}\sum_{i=1}^{n} K'\big(\frac{\Delta_i - \delta}{h}\big)(\hat{\Delta}_i - \Delta_i) + \frac{1}{2nh^3}\sum_{i=1}^{n} K''\big(\frac{\Delta_i^\dagger - \delta}{h}\big)(\hat{\Delta}_i - \Delta_i)^2$$

where $\Delta_i^\dagger$ is between $\hat{\Delta}_i$ and Δ_i. Since $\sup_i |\hat{\Delta}_i - \Delta_i| = O_p(n^{-1/2})$ from Theorem 8, we have

$$\Big|\frac{1}{nh^2}\sum_{i=1}^{n} K'\big(\frac{\Delta_i - \delta}{h}\big)(\hat{\Delta}_i - \Delta_i)\Big| \leq O_p(n^{-\frac{1}{2}}h^{-1}) \times \frac{1}{nh}\sum_{i=1}^{n}\Big|K'\big(\frac{\Delta_i - \delta}{h}\big)\Big|$$

where the summation is a nonparametric estimator of $f_\Delta(\delta) \times \int |K'(u)|du$. Therefore,

$$\frac{1}{nh^2}\sum_{i=1}^{n} K'\big(\frac{\Delta_i - \delta}{h}\big)(\hat{\Delta}_i - \Delta_i) = O_p(n^{-\frac{1}{2}}h^{-1}) = O_p(h^P/(\ln n)^{1/2})$$

which is an $o_p\big(h^P\big)$. Furthermore, because K'' is bounded, we have

$$\Big|\frac{1}{nh^3}\sum_{i=1}^{n} K''\big(\frac{\Delta_i^{\dagger}-\delta}{h}\big)(\hat{\Delta}_i-\Delta_i)^2\Big| = O_p(n^{-1}h^{-3})$$

which is also an $o_p\big(h^P\big)$ provided $P \geq 1$. Therefore, the first-step estimation error is negligible. $\square$

Bibliography

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *The Journal of Econometrics*, 113(2):231–263, 2003.

Alberto Abadie, Joshua Angrist, and Guido Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117, 2002.

Jason Abrevaya and Stephen Donald. A gmm approach for dealing with missing data on regressors. *Review of Economics and Statistics*, forthcoming.

Daron Acemoglu and A Robinson. The colonial origins of comparative development: An empirical investigation. *The American Economic Review*, 91(5):1369–1401, 2001.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Takeshi Amemiya. Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. *Econometrica: Journal of the Econometric Society*, pages 999–1012, 1974.

Takeshi Amemiya. The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica: Journal of the Econometric Society*, pages 955–968, 1977.

Donald WK Andrews. Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica: Journal of the Econometric Society*, pages 307–345, 1991.

Donald WK Andrews. Empirical process methods in econometrics. *Handbook of econometrics*, 4:2247–2294, 1994.

Joshua Angrist, Victor Chernozhukov, and Iván Fernández-Val. Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563, 2006.

Joshua Angrist, Victor Lavy, and Analia Schlosser. Multiple experiments for the causal link between the quantity and quality of children. *Journal of Labor Economics*, 28(4):773–824, 2010.

Joshua D Angrist. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pages 313–336, 1990.

Joshua D Angrist and William N Evans. Children and their parents' labor supply: evidence from exogenous variation in family size. *The American Economic Review*, 88(3):450–477, 1998.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Paul A Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681, 1994.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernandez-Val, and Christian B Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, forthcoming.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

Herman J Bierens. Consistent model specification tests. *Journal of Econometrics*, 20(1):105–134, 1982.

Richard Blundell and James L Powell. Endogeneity in nonparametric and semiparametric regression models. *Econometric Society Monographs*, 36:312–357, 2003.

Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, page 0962280215597579, 2015.

David Card. Using geographic variation in college proximity to estimate the return to schooling. *Aspects of labour market behavior: Essays in honour of John Vanderkamp*, pages 201–222, 1995.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

John C Chao, Norman R Swanson, Jerry A Hausman, Whitney K Newey, and Tiemen Woutersen. Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory*, 28(01):42–86, 2012.

Saraswata Chaudhuri and David K Guilkey. Gmm with multiple missing variables. Technical report, Technical report, University of North Carolina, Chapel Hill, 2013.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Xiaohong Chen and Demian Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015.

Xiaohong Chen and Xiaotong Shen. Sieve extremum estimates for weakly dependent data. *Econometrica*, pages 289–314, 1998.

Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5): 1591–1608, 2003.

Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366, 2005.

Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in gmm models of nonclassical measurement errors, missing data and treatment effects. 2008.

Victor Chernozhukov and Christian Hansen. The effects of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis. *The Review of Economics and Statistics*, 86(3):735–751, 2004.

Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132: 491–525, 2006.

Victor Chernozhukov and Christian Hansen. Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1):379–398, 2008.

Victor Chernozhukov, Whitney K Newey, and Andres Santos. Constrained conditional moment restriction models. *arXiv preprint arXiv:1509.06311*, 2015.

Andrew Chesher. Identification in nonseparable models. *Econometrica*, 71(5):1405–1441, 2003.

Andrew Chesher. Nonparametric identification under discrete variation. *Econometrica*, 73(5):1525–1550, 2005.

Pierre-Andre Chiappori, Ivana Komunjer, and Dennis Kristensen. Nonparametric identification and estimation of transformation models. *Journal of Econometrics*, 188(1):22–39, 2015.

Gordon Dahl and Stefano DellaVigna. Does movie violence increase violent crime? *The Quarterly Journal of Economics*, pages 677–734, 2009.

Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

Stephen G Donald, Guido W Imbens, and Whitney K Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117(1):55–93, 2003.

Juan Carlos Escanciano, D Jacho-Chavez, and A Lewbel. Uniform convergence for semiparametric two step estimators and tests. *Unpublished manuscript*, 2011.

Luca Flabbi, Stefano Paternostro, and Erwin R Tiongson. Returns to education in the economic transition: A systematic assessment using comparable data. *Economics of Education Review*, 27(6):724–740, 2008.

Markus Froelich and Blaise Melly. Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics*, 31(3):346–357, 2013.

Patrick Gagliardini and Olivier Scaillet. Nonparametric instrumental variable estimation of structural quantile effects. *Econometrica*, 80(4):1533–1562, 2012.

Bryan S Graham. Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 79(2):437–452, 2011.

Bryan S Graham, Cristine Campos De Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies*, 79(3):1053–1079, 2012.

Zvi Griliches. Wages of very young men. *Journal of Political Economy*, 84(4): S69–S85, 1976.

Zvi Griliches. Estimating the returns to schooling: Some econometric problems. *Econometrica: Journal of the Econometric Society*, pages 1–22, 1977.

Emmanuel Guerre, Isabelle Perrigne, and Quang Vuong. Optimal nonparametric estimation of first-price auctions. *Econometrica*, 68(3):525–574, 2000.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Jinyong Hahn. Optimal inference with many instruments. *Econometric Theory*, 18 (01):140–168, 2002.

Jinyong Hahn and Geert Ridder. Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica*, 81(1):315–340, 2013.

Jaroslav Hájek. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(4):323–330, 1970.

Christian Hansen, Jerry Hausman, and Whitney Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 2012.

James J Heckman and Xuesong Li. Selection bias, comparative advantage and heterogeneous returns to education: evidence from china in 2000. *Pacific Economic Review*, 9(3):155–171, 2004.

James J Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation1. *Econometrica*, 73(3):669–738, 2005.

James J Heckman and Edward J Vytlacil. Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of econometrics*, 6:4875–5143, 2007.

James J Heckman, Jeffrey Smith, and Nancy Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Joel L Horowitz. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica: Journal of the Econometric Society*, pages 103–137, 1996.

Joel L Horowitz and Sokbae Lee. Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75(4):1191–1208, 2007.

Marian Hristache and Valentin Patilea. Semiparametric efficiency bounds for conditional moment restriction models with different conditioning variables. *Econometric Theory*, pages 1–30, 2014.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

Guido W Imbens and Donald B Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4): 555–574, 1997.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Zhijie Xiao. Inference on the quantile regression process. *Econometrica*, 70(4):1583–1612, 2002.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data.* John Wiley & Sons, 2014.

Enno Mammen, Christoph Rothe, Melanie Schienle, et al. Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics*, 40(2): 1132–1170, 2012.

Vadim Marmer and Artyom Shneyerov. Quantile-based nonparametric inference for first-price auctions. *Journal of Econometrics*, 167(2):345–357, 2012.

Magne Mogstad and Matthew Wiswall. Instrumental variables estimation with partially missing instruments. *Economics Letters*, 114(2):186–189, 2012.

Chris Muris. Efficient gmm estimation with a general missing data pattern. Technical report, mimeo, Simon Fraser University, 2011.

Roger B Nelsen. *An introduction to copulas.* Springer Science & Business Media, 2007.

Whitney K Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica: Journal of the Econometric Society*, pages 809–837, 1990.

Whitney K Newey. Two-step series estimation of sample selection models. *The Econometrics Journal*, 12(s1):S217–S229, 2009.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1778, 2003.

Ariel Pakes and David Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the Econometric Society*, pages 1027–1057, 1989.

Thierry Post, Martijn J Van den Assem, Guido Baltussen, and Richard H Thaler. Deal or no deal? decision making under risk in a large-payoff game show. *The American economic review*, pages 38–71, 2008.

James M. Poterba, Steven F. Venti, and David A. Wise. How retirement saving programs increase saving. *The Journal of Economic Perspectives*, 10(4):91–112, 1996.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Christoph Rothe and Sergio Firpo. Semiparametric estimation and inference using doubly robust moment conditions. 2013.

Andrea Rotnitzky and James M Robins. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4):805–820, 1995.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

RM Sakia. The box-cox transformation technique: a review. *The statistician*, pages 169–178, 1992.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615, 1994.

George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics*, 23(R1):R89–R98, 2014.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

Mark J Van der Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.

Stephanie von Hinke Kessler Scholder, George Davey Smith, Debbie A Lawlor, Carol Propper, and Frank Windmeijer. Mendelian randomization: the use of genes in instrumental variable analyses. *Health economics*, 20(8):893–896, 2011.

Quang Vuong and Haiqing Xu. Counterfactual mapping and individual treatment effects in nonseparable models with discrete endogeneity. *Quantitative Economics*, forthcoming.

Le Wang. Estimating returns to education when the iv sample is selective. *Labour Economics*, 21:74–85, 2013.

Jeffrey M Wooldridge. Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1(2):117–139, 2002.

Jeffrey M Wooldridge. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2):1281–1301, 2007.

S.M. Zemyan. *The Classical Theory of Integral Equations: A Concise Treatment*. Birkhäuser Boston, 2012. ISBN 9780817683481. URL https://books.google.com/books?id=_u86LgEACAAJ.

Vita

Qian Feng was born in Nanyang, Henan. After graduating from No.2 High School in Nanyang, she went to Shanghai where she received the Bachelor of Arts degree in Economics in 2009. In the Fall of 2011, she entered the graduate program in Economics at the University of Texas at Austin.

Permanent address: 10904 Rock Island Drive
Austin, Texas 78717

This dissertation was typeset with LaTeX[†] by the author.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.