ESSAYS ON DISCRETE MULTIVALUED TREATMENTS WITH ENDOGENEITY AND HETEROGENEOUS COUNTERFACTUAL ERRORS

By

Ibrahim Kekec

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2021

ABSTRACT

ESSAYS ON DISCRETE MULTIVALUED TREATMENTS WITH ENDOGENEITY AND HETEROGENEOUS COUNTERFACTUAL ERRORS

By

Ibrahim Kekec

This dissertation is composed of three chapters, and each one of them studies discrete multivalued treatments with endogeneity and heterogeneous counterfactual errors. The first chapter extends the investigations of average treatment effects (ATEs) in extensively-studied binary treatments to those in discrete multivalued treatments with both endogeneity and heterogeneous counterfactual errors and explores the behavior of control function (CF) and instrumental variables (IV) methods in this framework. Specifically, I offer identification strategies for the ATEs, suggest a consistent estimator for the ATEs, show the asymptotic properties of CF parameter estimates, and derive a score test in order to draw inferences about the ATEs and other parameters of interest. Moreover, using a Monte Carlo simulation analysis, I compare CF method with widely used IV method in terms of asymptotic efficiency, asymptotic unbiasedness, and consistency. Simulation results suggest that CF method can be asymptotically up to 12% more efficient than IV method, and asymptotic bias in parameter estimates of IV method can be as high as 43%. However, when misspecification is introduced, simulation results favor IV method. For the empirical illustration, I apply ordinary least squares (OLS), CF, IV, and nonparametric bound analysis to the estimation of how limited English proficiency (LEP) influences wages of Hispanic workers in the USA. The data come from the 1% Public Use Microdata Series of the 1990 US Census. Utilizing age at arrival as an instrumental variable, both OLS and CF methods indicate that LEP on average imposes a statistically significant wage penalty (up to 79% in some CF estimates) on Hispanic community in the USA. IV method mostly produces insignificant results, and nonparametric bound analysis provides uninformative lower bounds.

The second chapter incorporates a structure of correlated random coefficients (CRCs) into the framework introduced in the first chapter. However, in this new setting with CRCs, conventional IV method is suspected to be inconsistent for ATEs. In this chapter, I propose a consistent CF estimation procedure for the ATEs and show the asymptotic properties of CF parameter estimates. In addition, my Monte Carlo simulation analysis suggests that, in the absence of misspecification, CF method is asymptotically unbiased and consistent (but not necessarily more efficient). Whereas, IV method is generally asymptotically biased and inconsistent. In the presence of misspecification, the simulation results show that both CF and IV methods have biased estimates (more on CF estimates). With regard to efficiency, the simulation findings show that none of the methods outperforms the other one clearly.

In the third chapter, I take the treatment model from the first chapter to a specific linear high dimensional sparse setting where the high dimensional variables are irrelevant in treatment choice given the instruments and appear only in the outcome equation. Using a detailed simulation analysis, I examine the finite sample properties, model selection features, and prediction capabilities of several machine learning (ML) methods and of the CF method from the first chapter. To estimate the parameters of interest, I use four different ML methods: LASSO; post partial-out LASSO of Belloni et al. (2012); post double selection LASSO of Belloni, Chernozhukov, and Hansen (2014a); and double/debiased ML LASSO of Chernozhukov et al. (2018). The most important simulation result is that, in the presence of enough extra predictive variables that are ignorable in treatment selection and are from a set of high dimensional predictors of outcome, more complicated LASSO-based methods result in efficiency gains in ATE estimates over the simpler CF method although both LASSO-based methods and the CF method perform more or less the same as far as finite sample bias is concerned. As far as model selection goes, the simulations show that the double/debiased ML LASSO both selects the most number of potential variables and correctly selects the most number of variables with true nonzero impact on outcome in estimation. As to prediction, the simulation results suggest that LASSO has the best prediction features.

To my family, for always supporting me.

ACKNOWLEDGMENTS

I would like to thank to my advisor, Jeffrey M. Wooldridge, for his guidance during my doctoral journey and all the suggestions he offered on the previous versions of my dissertation. I also thank my committee members, Kyoo il Kim, Peter Schmidt, and Chih-Li Sung for their comments. My special thanks go to Steven Haider, Todd Elder, and Soren Anderson for their useful advices; Jay Feight, Lori Jean Nichols, and Margaret Lynch for their administrative help; and Dean Olson and the High Performance Computing Center at Michigan State University Institute for Cyber-Enabled Research for technical support.

I also would like to convey my sincere appreciation to Margie Tieslau for sowing the love of econometrics in me and my deep gratitude to Joan G. Staniswalis for encouraging me when I really needed it. My dear friends Yi Li and Fei Jia have my deep thanks for lending me their helping hands during the ups and downs of being a doctoral student.

Lastly and most importantly, I extend my most profound gratitude to my family for always supporting me unconditionally no matter what. I am indebted to my sister for urging me to start this journey and always being there for me. I am forever thankful to my parents for their never-ending love that has shaped me to be the person that I am now. To their sacrifices, I dedicate this milestone in my life. I could not have done it without your love and help.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

IDENTIFICATION, ESTIMATION, AND INFERENCE FOR MULTIVALUED ENDOGENOUS TREATMENT EFFECT MODELS: A CONTROL FUNCTION APPROACH

1.1 Introduction

In the economics literature, a great deal of interest lies in the estimation of average treatment effects (ATEs) since it gives economists a method to evaluate the effects of government programs and policies, such as school voucher programs, the effects of personal choices, such as higher education attendance, and the effects of many other institutional or personal decisions. To quantify the interest, in 2020 so far even in the midst of Covid-19 pandemic, 9 articles related to ATEs were published in the journal *Econometrica* alone.

The economics literature in ATEs is fairly large. The golden standard for estimating ATEs is through the use of randomized control trials or natural experiments for many economists; however, these estimation methods are rare mostly due to financial restraints or ethical problems (e.g., trying to assign some students into the control group against their will while investigating the impact of a tutoring program on students' academic achievements) associated with an experiment. For these (and other similar) reasons, economists often use observational data and employ estimation methods that suit observational data. Using observational data, there are three major methods in the ATE estimation arsenal of an economist: methods employing ignorability, regression discontinuity designs, and instrumental variables (IV) and control function (CF) methods. Rosenbaum and Rubin (1983) first used the ignorability assumption in the context of ATEs, and the assumption states the independence of treatment variable and counterfactual outcomes given observed control variables. With the help of this very assumption, the classical methods to estimate ATEs are regression adjustment, propensity score, and matching. For those interested in further information on these methods, both Wooldridge (2010) and Cameron and Trivedi (2005) have

their decent coverage. In certain cases where discontinuity of policy assignment is the direct result of some ad hoc institutional regulation/rule, outcome differences between those who are treated and those who are not can be indeed ascribed to treatment statuses. Regression discontinuity designs pave its way for estimating ATEs at the discontinuity point in these very cases. For more information on this method, see Hahn, Todd, and van der Klaauw (2001); van der Klaauw (2002); and Imbens and Lemieux (2008).

When the treatment status is truly endogenous in ATE models, the usual ignorability assumption fails. In such a case, the conventional methods to estimate ATEs such as regression adjustment or propensity score weighting proves ineffective. Among different estimation methods, IV method is one of the most, if not the most, commonly used estimation methods when treatment choice is endogenous. Angrist, Imbens, and Rubin (1996); Heckman (1997); and Wooldridge (1997, 2003) are all useful for reference purposes and discuss the estimation of ATEs using IV in detail. CF method where estimating equation gets augmented by the addition of functions of all relevant observed covariates (inclusive of the endogenous treatment variable) is an alternative in such endogenous setting, as well. Historically, CF method dates back as early as 1970s. Heckman (1979) and Lee (1978) can be considered two examples of the early works where the idea behind the CF method was used; nevertheless, they did not use the very words "control function." while referring to their estimation methods. Specifically, Heckman (1979) used the idea in order to take care of sample selection bias. The switching regression model of Lee (1978) where union membership changes the wage equations for workers employed the idea to examine the relationship between the labor unions and wage rates. For review purposes, Vella and Verbeek (1999) indeed give a good discussion of CF method in a setting where binary treatment variable is endogenous and analyzes the relationship between CF and IV methods. In addition, Wooldridge (2015) provides a very comprehensive view of CF methods in both linear and nonlinear models with endogenous explanatory variables and argues that CF methods are, in a complementary fashion, alternatives to traditional IV methods. Furthermore, ATE models are indeed part of program

evaluation literature, so for an overview of standard methods and advances in the theoretical aspects of program evaluation, see Imbens and Wooldridge (2009) and Abadie and Cattaneo (2018).

When the treatment status takes on more than just two values, i.e. binary treatment effect, economists step into the realm of multivalued treatment effects where the number of treatments can be more than just two but finite. Seminal works in casual inference with multivalued treatments were developed by Rubin (1978) and Robins (1989a, 1989b). Building on top of these works, Angrist and Imbens (1995) improved the casual inference with multivalued treatments by both going beyond binary treatments and utilizing the concept of counterfactuality in their IV analysis of the effect of schooling on earnings. Upon these pioneering works, academic interest in ATEs seems to go on without any reduction in intensity. Recent survey articles contain but not limited to Heckman and Vytlacil (2007a, ch. 70; 2007b, ch.71), Imbens and Wooldridge (2009), and Linden *et al.* (2016) specifically for multivalued treatments under ignorability. In the ATE literature, the latent choice model for treatment statuses in treatment effects is indeed a discrete choice model on which studies have reached maturity. Besides survey articles (see, for example, Amemiya, 1981), there are several textbooks and book chapters devoted to discrete choice analysis (a.k.a. qualitative response models). To name a few, see Daganzo (1979) specifically in multinomial probit models, Ben-Akiva and Lerman (1985), Maddala (1986), McFadden (1984, ch. 24), Amemiya (1985, ch. 9), Maddala and Flores-Lagunes (2001, ch. 17), Cameron and Trivedi (2005, chs. 14 and 15), and Wooldridge (2010, chs. 15 and 16).

One of the first studies that utilize discrete choice models with multiple choices belongs to the work of Dubin and McFadden (1984). In their paper, they studied the demand for electricity by residences. To estimate the demand for electricity model, price and income elasticities, appliance dummy variables that follow a discrete choice model were included in their model. Since the theory of economics suggests that these dummy variables are endogenous, Dubin and McFadden estimated the demand for electricity by IV and CF methods.

In their final analysis, they obtained similar elasticities in magnitude from both estimation methods.

Another paper in the subject comes from Bourguignon, Fournier, and Gurgand (2007). In their survey paper, they presented the set of available methods when the multinomial logit model (MNL) underlies the discrete choice model. The three approaches in their paper were those studied by Lee (1983), Dubin and McFadden (1984), and Dahl (2002). After having showed the advantages and disadvantages of these approaches, they then concluded that Dubin and McFadden's model and Dahl's model were more efficient compared to other specifications. In addition, Bourguignon, Fournier, and Gurgand (2007) improved the set of methods at researchers' disposal by allowing correlations between different choices.

In a binary treatment effect model with an endogenous treatment variable, if the counterfactual errors are heterogeneous in the sense that they depend on treatment status, then conventional IV estimation in general leads to inconsistent estimates including ATEs. For example, Angrist (1991, p. 15) delineates the problems of estimating ATEs by IV method while, at the same time, outlining functional form restrictions for the method to provide consistent estimates of ATEs. He shows that IV method can lead to asymptotically biased and inconsistent estimates of ATEs when the error term in the outcome equation interacts with endogenous binary treatment. Similarly, Heckman and Robb (1985b, p. 196) portray when IV method is appropriate to use for consistent estimation of ATEs in binary treatments and state that standard IV method does not estimate ATEs consistently in a framework where the error term of the estimating equation again contains components interacting with endogenous binary treatment but CF method does. Wooldridge (1997, p. 131; 2003, p.191) also mentions about this issue clearly. Since heterogeneous counterfactual errors result in composite error terms including interactions with endogenous treatment, one can deduce that consistent estimation of ATEs in multivalued treatments by IV method hinges on whether the counterfactual errors are homogeneous.

In discrete ATE literature, most of the attention has been devoted to binary treatment

models with endogeneity, which leaves behind an untapped area of research in discrete multivalued treatments with endogeneity. This chapter extends the investigations of ATEs in binary treatments to those in discrete multivalued treatments with both endogeneity and heterogeneous counterfactual errors and explores the behavior of both CF and instrumental variables (IV) methods comparatively in this framework, which has not been examined to the best of my knowledge and constitutes my main contribution to the literature. Specifically, in this chapter, I offer identification strategies for the ATEs, suggest a consistent CF estimator for the ATEs, show the asymptotic properties of CF parameter estimates, and derive a score test in order to draw inferences about the ATEs and other parameters of interest in a discrete multivalued treatment model with endogeneity and heterogeneous counterfactual errors. I follow the latent choice model setup laid out by Dubin and McFadden (1984) for the endogenous treatment variable. Under this setup, the endogenous treatment variable follows a multinomial logit reduced form. This key observation enables me to calculate expectations of heterogeneous counterfactual errors conditional on all the exogenous variables and the endogenous treatment variable, which plays a critical role in deriving the estimating equation for CF method. I argue that CF method can be more efficient than IV method when the counterfactual errors are homogeneous and that IV parameter estimates can suffer from considerable biases when the counterfactual errors are heterogeneous. However, when misspecification is introduced, my findings slightly favor IV method.

The rest of this chapter is organized as follows. In section 1.2, I introduce the model. In section 1.3, I discuss identification strategies for the model. In section 1.4, I derive the estimating equations for both CF and IV methods and propose procedures to estimate the parameters of interest and ATEs for both methods. In section 1.5, I show the asymptotic properties of CF estimates, propose a consistent estimator for the asymptotic variance matrix of CF estimates, and show how a GMM framework can be set up for the main problem. In section 1.6, I suggest a score test to draw inferences about ATEs and parameters of interest. In section 1.7, I share some simulation results. In section 1.8, I apply ordinary least squares

(OLS), CF, IV, and nonparametric bound analysis while estimating the impact of English proficiency on wages of Hispanic workers in the USA. In section 1.9, I conclude. And, in appendix A, I share the derivations, simulation tables, and empirical analysis tables that are hidden from the main body of this chapter.

1.2 The Model

Consider the following model

$$\begin{aligned} y_g &= \alpha_g + \mathbf{x}\beta_g + u_g \\ w_g^* &= \mathbf{z}\gamma_g + a_g, \end{aligned} \tag{1.1}$$

where y_g is the g^{th} counterfactual outcome variable, α_g is the scalar coefficient in the counterfactual outcome equation for y_g, $\mathbf{x} \equiv (x_1, x_2, \ldots, x_l)$ is the $1 \times l$ vector of exogenous variables in y_g, β_g is the $l \times 1$ vector of slope coefficients in y_g, u_g is the counterfactual error in y_g, w_g^* is the latent treatment variable that determines the choice of treatment status among $G+1$ alternative treatment statuses, $\mathbf{z} \equiv (z_1, z_2, \ldots, z_k)$ is the $1 \times k$ vector of instruments that includes a constant term in the choice equation for w_g^*, γ_g is the $k \times 1$ vector of parameters in w_g^*, and a_g is the scalar error term that is independently and identically Gumbel distributed (*i.i.d.*) with location parameter $\mu = 0$ and scale parameter $\beta = 1$ in w_g^* for $g = 0, 1, \ldots, G$.

Let $w \in \{0, 1, \ldots, G\}$ be the observed discrete multivalued endogenous treatment variable whose values are determined by w_g^* for $g = 0, 1, \ldots, G$. One common interpretation of w_g^* is to think of it as the utility or satisfaction obtained from treatment status g. Let the treatment statuses of w be exhaustive and mutually exclusive. Define binary treatment status indicators, $d_g = 1[w = g]$ for $g = 0, 1, \ldots, G$. So the binary treatment status indicator d_g is equal to one if the treatment status is equal to g and zero otherwise. This coupled with the mutual exclusivity of treatment statuses implies that $\sum_{g=0}^{G} d_g = 1$. Define the $1 \times (G+1)$

vector of treatment statuses $\mathbf{d} \equiv (d_0, d_1, \ldots, d_G)$. Let y be the observed outcome. Then, I can write

$$y = d_0 y_0 + d_1 y_1 + \cdots + d_G y_G, \tag{1.2}$$

where y_g is the g^{th} counterfactual outcome for $g = 0, 1, \ldots G$.

After having described the discrete multivalued endogenous treatment model above, I now will make a series of assumptions that complete the model and that are used in estimation. First, I assume that the rational economic agents choose the status of treatment from which they receive the most satisfaction out of all possible treatment statuses. That is,

- **Assumption** 1.1 **(A.**1.1**)**: One chooses treatment status g, i.e., $w = g$ if and only if $w_g^* \geq w_j^* \ \forall j \neq g$ for $g, j = 0, 1, \ldots, G$.

Second, I assume that identification of the model in (1.1) and (1.2) is contributed by exclusion of some (at least one) variables in the set of instruments $\mathbf{z}$ from the set of exogenous variables in $\mathbf{x}$. This exclusion restriction is encouraged for the estimation and identification to be more convincing and reliable even though nonlinearity in estimation suffices for identification, especially when the exogenous variables in $\mathbf{z}$ vary enough in the sample. In the literature, it is common that the set of exogenous variables in $\mathbf{x}$ is a proper subset of the set of instruments $\mathbf{z}$. That is, $\mathbf{z}$ includes all the variables in $\mathbf{x}$ and has at least one additional variable that is not in $\mathbf{x}$. The idea is that all of the characteristics influential in the outcome are also expected to be critical in determining the treatment choice. For a concrete example on this point, see Vella and Verbeek (1999, p. 473).

- **Assumption** 1.2 **(A.**1.2**)**: Identification of the model described by (1.1) and (1.2) is strengthened by exclusion of at least one variable in $\mathbf{z}$ from the set of variables in $\mathbf{x}$.

As shown by McFadden (1973), under the model in (1.1) and (1.2) the assumptions made so far allow the treatment variable w to follow a multinomial logit model with choice probabilities given as follows:

$$P(w = g|\mathbf{x}, \mathbf{z}) = P(w = g|\mathbf{z}) = exp(\mathbf{z}\gamma_g)/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_r), \tag{1.3}$$

for $g = 0, 1, \ldots, G$. The next assumption is essential to the CF estimation, which I describe in section 1.4, since this assumption coupled with the multinomial logit specification of the treatment variable w will enable me to form control function terms that account for the endogeneity in w.

- **Assumption** 1.3 **(A.**1.3**)**: $E(u_g|\mathbf{x}, \mathbf{z}, \mathbf{a}) = E(u_g|\mathbf{a}) = \sum_{j=0}^{G} \eta_{g,j} a_j + [-\sum_{j=0}^{G} \eta_{g,j} E(a_j)]$, where u_g is the counterfactual error in y_g, $\mathbf{x}$ is the $1 \times l$ vector of exogenous variables in y_g, $\mathbf{z}$ is the $1 \times k$ vector of instruments that includes a constant term in the choice equation for w_g^*, $\mathbf{a} \equiv (a_0, a_1, \ldots, a_G)$ is the $1 \times (G+1)$ vector of $i.i.d.$ Gumbel distributed errors a_j with location parameter $\mu = 0$ and scale parameter $\beta = 1$ in w_j^*, $\eta_{g,j}$ is the scalar multiple of correlation coefficient between u_g and a_j, and $E(a_j) = 0.5772$ is Euler's constant for $j, g = 0, 1, \ldots, G$.

Bourguignon, Fournier, and Gurgand (2007) refers to A.1.3 as Dubin and McFadden's linearity assumption since the conditional expectation of counterfactual error u_g given all Gumbel distributed errors $\mathbf{a}$ is linear in $\mathbf{a}$ for $g = 0, 1, \ldots, G$. A.1.3 also implies that, conditional on $\mathbf{a}$, $\mathbf{x}$ and $\mathbf{z}$ are redundant for the conditional expectation of u_g. In other words, u_g is mean independent of $\mathbf{x}$ and $\mathbf{z}$ conditional on $\mathbf{a}$.

Under all assumptions from A.1.1 through A.1.3, the model in (1.1) and (1.2) can be consistently estimated by CF method. In section 1.4, I will propose a consistent estimator for the ATEs in this discrete multivalued endogenous treatment model.

1.2.1 The Model with $\eta_{g,j} = \eta_j$: A Special Case

This special case follows the initial model, so the basic setup and variables in (1.1) and (1.2) are the same. However, in this special case, I modify the initial model by assuming that $\eta_{g,j} = \eta_j$, $g = 0, 1, \ldots, G$, which makes the main difference in this model. The assumptions I made in this new model are pretty much the same as the ones in the initial model except A.1.3 and are as follows:

- **Assumption** 1.1′ **(A.**1.1′**)**: Same as **A.**1.1.
- **Assumption** 1.2′ **(A.**1.2′**)**: Same as **A.**1.2.
- **Assumption** 1.3′ **(A.**1.3′**)**: $E(u_g|\mathbf{x}, \mathbf{z}, \mathbf{a}) = E(u_g|\mathbf{a}) = \sum_{j=0}^{G} \eta_j a_j + [-\sum_{j=0}^{G} \eta_j E(a_j)]$, where u_g is the counterfactual error in y_g, $\mathbf{x}$ is the $1 \times l$ vector of exogenous variables in y_g, $\mathbf{z}$ is the $1 \times k$ vector of instruments that includes a constant term in the choice equation for w_g^*, $\mathbf{a}$ is the $1 \times (G+1)$ vector of *i.i.d.* Gumbel distributed errors a_j with location parameter $\mu = 0$ and scale parameter $\beta = 1$ in w_j^*, η_j is the fixed scalar multiple of correlation coefficient between u_g and a_j, and $E(a_j) = 0.5772$ is Euler's constant for $j, g = 0, 1, \ldots, G$.

The last assumption, A.1.3′, now incorporates the special condition that $\eta_{g,j} = \eta_j$which means that the correlation coefficient between u_g and a_j does not change as the treatment status g changes, but is fixed at η_j. Because of this special condition, $u_g = u$ for $g = 0, 1, \ldots, G$, which practically means that counterfactual errors u_g's are homogeneous across treatment statuses.

1.3 Identification

In plain words, identification in economics is the concept of uniquely determining some population parameters in an econometric model from what can be observed in the name of data. Unlike natural sciences, economists cannot control the conditions under which social phenomena occur or change, and the data are produced by an unknown process. Hence, an economist needs to formulate a model with its assumptions and restrictions in order to simplify the complexity of the real world and understand the population of interest. If a model is identified, then only a single set of parameter values must agree with the data under the true model, and other parameter values must point to a different data.

The origins of econometric identification go as far back in time as late 1920s. Stock and Trebbi (2003) argue that Wright (1928) was the pioneer in econometric identification because he solved an identification problem in econometrics first. Apart from these studies, Working (1925, 1927), Haavelmo (1943), Koopmans and Reiersøl (1950), Hurwicz (1950), Koopmans, Rubin, and Leipnik (1950), Wald (1950), and Fisher (1966) give the earliest examples of identification research in econometrics. For a complete review of the early research in econometric identification and its formalization, see Duo (1993, ch. 4). Since the early days in econometric identification, its literature has grown considerably and become complicated but can be roughly divided into two: point identification and set identification (a.k.a. partial identification).

Conventionally, when a parameter is identified, an economist means that it is point identified, which is mainly because the research in point identification has started earlier than in set identification. From a terminological point of view, Koopmans and Reiersøl (1950), Hurwicz (1950), Fisher (1966) and Rothenberg (1971) give the first formal definitions of point identification. Other recent definitions are available in Hsiao (1983), Bekker and Wansbeek (2001), and Matzkin (2007). In line with Lewbel (2018), let say that θ is the population parameter to be identified, ϕ is everything known about the population that data can offer, and M is a model with a set of assumptions and restrictions on the set of

possible values ϕ can take on. Then, singling out θ from ϕ given the model M establishes the point identification of θ, which is an introductory and informal way of thinking about point identification. In a similar but more informal fashion, Duo (1993, p. 95) describes the process of point identification as "... essentializing the conditions under which a certain set of values of structural parameters could be uniquely determined from the data among all the permissible sets embodied in mathematically complete theoretical model ... ".

As for set identification, in an intuitively simple way, Lewbel (2018, p. 65) describes it as "... the analysis of situations where ϕ provides some information about parameter θ, but not enough information to actually point identify θ ... ". From an informal standpoint, the true parameter θ_o is set identified if some other possible parameter values have the same chance of creating the data in hand as does the true parameter value (terminologically, these other possible parameter values are called observationally equivalent to θ_o and form the identified set together with θ_o). And, if the identified set is only composed of θ_o, then point identification of θ_o is the same as its set identification. Frisch (1934), Reiersøl (1941), and Marschak and Andrews (1944) are among the first works on set identification. Manski (1990, 1995, 2003) also analyze the subject in great detail over the years. More recent definitions are available in Matzkin (2007) and Chesher and Rosen (2017), and set identification has been reviewed in Tamer (2010). Some researchers favor set identification over point identification because, according to them, economic theory does not supply econometric models with enough information for the point identification of model parameters, leading to sophisticated tricks to obtain point identification. However, one downside of set identified models is that estimation and inference in set identified parameters get rather more complicated than in point identified parameters.

In two-stage models, the main identification strategy is generally based on either some exclusion restrictions or functional form assumptions. Applied economists also prefer using both exclusion restrictions and functional form assumptions at the very same time in order to aid and strengthen identification in their models. In two-stage models, an exclusion

restriction means the inclusion of an explanatory variable in the first stage equation that is excluded from the second stage equation. There can also be exogenous variables that do not appear in the first stage equation but are included only in the second stage equation, which is expected to produce even more variation in the model for its parameters to be identified.

The economists who are in favor of exclusion restrictions argue that identification based on functional form assumptions (a.k.a. identification by functional form) is fragile, especially in empirical settings where the data do not have enough variation in key variables. Since identification by functional form relies heavily on model assumptions which might not be true in reality, it can greatly suffer from distributional misspecifications and exclusion restrictions can be required for identification. For example, Keane (1992), Reilly (1996), Montmarquette, Viennot-Briot, and Dagenais (2007), and Shen (2013) all use some exclusion restrictions as their main identification strategy. For Olsen (1980) and Little (1985), the use of nonlinearities in identification is even unappealing because identification by functional form is not only weak but also causes high standard errors and unreliable estimates.

On the other hand, the economists who are in favor of functional form assumptions argue that identification based on exclusion restrictions is hard to achieve, especially in empirical settings where finding a true instrument that only affects the selection equation and that does not appear in the equation of interest is not very realistic. For instance, Heckman (1978), Heckman and Robb (1985a), Mendelsohn (1985), Schaffner (2002), and Lewbel (2012) all employ the nonlinearities in exogenous variables for aiding identification. In the Monte Carlo simulations of Leung and Yu (1996), they find evidence supporting the claim that two-step models are reliable in the absence of exclusion restrictions given enough variation in one of the exogenous variables in data. Wilde (2000) provide evidence in support of that parameter identification does not require exclusion restrictions in a system of equations, given each equation has at least one explanatory variable with enough variation. Given the assumption of multivariate normal distribution, his multiple equation probit model is identified without exclusion restrictions. Similarly, Escanciano *et al.* (2016) also have some

results that positively contribute to that identification by functional form is neither fragile nor unreliable in a large class of two-stage models. In certain models, some economists even show that identification of two-stage models is possible without exclusion restrictions, see, for instance, Dong (2010)'s binary choice model. It is also common that some economists use both exclusion restrictions and functional form assumptions for aiding identification in their models, see Mocan and Tekin (2003) and Appelt (2015) for further comments.

In my main model given by (1.1), (1.2), A.1.1, A.1.2, and A.1.3; the identification argument is based on both exclusion restriction(s) as in A.1.2 and nonlinearity that describes the relationship between the set of instruments $\mathbf{z}$ and the treatment variable w which follows a multinomial logit model in $\mathbf{z}$. Due to arguments stressing that it is hard to achieve identification without imposing an exclusion restriction, I rely not only on the nonlinearity but also on the exclusion restriction(s) in my model so as to come up with a stronger and improved identification argument that appeases a wide range of economists.

1.4 Estimation

In multivalued treatment cases, ATEs depend on the choice of a base treatment group out of all possible treatment groups. Upon the determination of the base treatment group, ATEs can be defined as the expectation of the gain from the treatment received with respect to the base treatment group. Note that there are G number of ATEs in my analysis since there exist $G+1$ treatments. Let $g=0$ be the base group in my analysis. Denote $ATE_{g,0}$ as the expected gain from treatment g with respect to the base treatment. In my model, under A.1.1 through A.1.3, and the law of iterated expectations, ATEs will take the following form:

$$\begin{aligned} ATE_{g,0} &= E(y_g - y_0) \\ &= E\left(\alpha_g + \mathbf{x}\beta_g + u_g - (\alpha_0 + \mathbf{x}\beta_0 + u_0)\right) \\ &= (\alpha_g - \alpha_0) + \left(E(\mathbf{x})\right)(\beta_g - \beta_0), \end{aligned} \tag{1.4}$$

where the third equality uses $E(u_g) = 0$ for $g = 1, 2, \ldots, G$.

Then, a consistent estimator of $ATE_{g,0}$ is

$$\widehat{ATE_{g,0}} = (\hat{\alpha}_g - \hat{\alpha}_0) + \bar{\mathbf{x}}(\hat{\beta}_g - \hat{\beta}_0), \tag{1.5}$$

where $\hat{\alpha}_g$, $\hat{\alpha}_0$, $\hat{\beta}_g$, $\hat{\beta}_0$, and $\bar{\mathbf{x}}= N^{-1}\sum_{i=1}^{N}\mathbf{x}_i$ are respectively consistent estimates for α_g, α_0, β_g, β_0, and $E(\mathbf{x})$. Note that when $E(\mathbf{x}) = \mathbf{0}$, $ATE_{g,0}$ simplifies to

$$ATE_{g,0} = (\alpha_g - \alpha_0). \tag{1.6}$$

Then, a consistent estimate of $ATE_{g,0}$ in (1.6) is

$$\widehat{ATE_{g,0}} = (\hat{\alpha}_g - \hat{\alpha}_0), \tag{1.7}$$

which is simply the difference between the estimates $\hat{\alpha}_g$ and $\hat{\alpha}_0$ for $g = 1, 2, \ldots, G$.

1.4.1 IV Estimation

Consider the observed outcome:

$$\begin{aligned} y &= d_0 y_0 + d_1 y_1 + \ldots + d_G y_G \\ &= \sum_{j=0}^{G} d_j \alpha_j + \sum_{j=0}^{G} d_j \mathbf{x}\beta_j + u', \end{aligned} \tag{1.8}$$

where $u' = d_0 u_0 + d_1 u_1 + \cdots + d_G u_G$. Applying IV method on (1.8) requires instruments for the binary treatment indicators d_j since $corr(d_j, u')$ is expected not to be zero for $j = 0, 1, \ldots, G$. One way to obtain instruments for d_j is to model the treatment variable w as a discrete multinomial logit model and to then use the predicted probabilities from this model as instruments for d_j for $j = 0, 1, \ldots, G$. Hence, one can prescribe the following three-stage procedure to estimate ATEs:

Procedure 1.1

1. Estimate the predicted probabilities, $\hat{\Lambda}_{j_i} = exp(\mathbf{z}_i\hat{\gamma}_j)/\sum_{r=0}^{G} exp(\mathbf{z}_i\hat{\gamma}_r)$, from a MNL of w_i on $\mathbf{z}_i$ for $j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$.

2. Estimate the parameters in (1.8) by IV method using instruments $\left(\hat{\Lambda}_{j_i}, \hat{\Lambda}_{j_i}\mathbf{x}_\mathrm{i}\right)$ for $(d_{j_i}, d_{j_i}\mathbf{x}_i)$, $j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$.

3. Plug parameter estimates from step 2 and sample average of $\mathbf{x}$ into (1.5), and estimate ATEs.

An important remark to make here is that IV estimator described in Procedure 1.1 is not a conventional IV estimator. It is in fact optimal (asymptotic variance minimizing) IV estimator with optimal instruments (predicted probabilities from the MNL of w on $\mathbf{z}$) under homoskedasticity. Whether the error term u' in (1.8) is homoskedastic is another question; however, if it is, optimal IV estimator in Procedure 1.1 is asymptotically efficient over a parametric family of conditional probabilities for w. For this reason, I prefer IV estimator described in Procedure 1.1 over other IV estimators using different variations of instruments that are functions of $\mathbf{z}$. For an example of optimal IV estimator in an endogenous dummy variable model, see Newey (1993, p. 430). For further discussion on optimal IV estimators, see Newey and McFadden (1994, Section 5.4).

Procedures similar to Procedure 1.1 in models with binary endogenous treatment are popular among the empirical economists relatively because of the straightforward implementation of IV method. For example, Robinson (1989) used a two-stage procedure where he obtained a set of predicted probabilities from a probit specification for the endogenous variable , i.e. union choice, at the first stage and used, at the second stage, that as instrument for the union choice in his model to estimate the union wage differentials. Similarly, Puhani and Weber (2007) calculated a set of predicted probabilities again from a probit specification for the endogenous variable , i.e. age of school entry, at the first stage and used, at the

second stage, that as instrument for the age of school entry in their model to explore how the age of school entry influences educational outcomes. Examples in multivalued cases are also numerous, see, for example, Ettner (1995 and 1996); Norton and Staiger (1994); Sloan, Picone, Taylor Jr., and Chou (2001).

As pointed out by Chesher and Rosen (2013) and Lewbel, Dong, and Yang (2012), although this IV approach is simple to apply and is popular in empirical work with binary endogenous treatment, it is naive and results mostly in inconsistent IV estimates (and inconsistent ATE estimates in my analysis thereof) since instruments used in Procedure 1.1 are expected to be correlated with u'. This IV method with instruments (all are nonlinear functions of $\mathbf{z}$) would most likely yield inconsistent estimates because d_j is in u' and is a function of $\mathbf{z}$ for $j = 0, 1, \ldots, G$.

Note that (1.8) must be reformulated so that one can estimate it using the canned software packages such as STATA. This reformulation is needed because I include all of the binary treatment indicators d_g in (1.2), and the canned packages also include an intercept in the first stage of the IV estimation although some allow for the exclusion of a constant term in the second stage regression, e.g., STATA. Hence, in practice, the IV estimation of (1.8) suffers from perfect multicollinearity and is not possible.

To fix this practical problem described above, lets drop one of the binary treatment indicator variables, say d_G but it could be another one, from (1.8). And then add a constant term and the variables $\mathbf{x}$ into (1.8). Then, (1.8) can be equivalently written as

$$y = (\sum_{j=0}^{G-1} d_j \tilde{\alpha}_j + \tilde{\alpha}_G) + (\sum_{j=0}^{G-1} d_j \mathbf{x} \tilde{\beta}_j + \mathbf{x} \tilde{\beta}_G) + \tilde{u}', \tag{1.9}$$

where $\alpha_g = \tilde{\alpha}_g + \tilde{\alpha}_G$, $\beta_g = \tilde{\beta}_g + \tilde{\beta}_G$, $\alpha_G = \tilde{\alpha}_G$, and $\beta_G = \tilde{\beta}_G$ for $g = 0, 1, \ldots, G-1$. Under this reformulation, $ATE_{g,0}$ for $g = 1, 2, \ldots, G-1$ is

$$ATE_{g,0} = (\tilde{\alpha}_g - \tilde{\alpha}_0) + (E(\mathbf{x}))\,(\tilde{\beta}_g - \tilde{\beta}_0) \tag{1.10}$$

and for $g = G$

$$ATE_{G,0} = (-\tilde{\alpha}_0) + (E(\mathbf{x}))\,(-\tilde{\beta}_0). \tag{1.11}$$

Therefore, consistent estimates of $ATE_{g,0}$ for $g = 1, 2, \ldots, G-1$ and $ATE_{G,0}$ under this reformulation are

$$\widehat{ATE_{g,0}} = (\hat{\tilde{\alpha}}_g - \hat{\tilde{\alpha}}_0) + \bar{\mathbf{x}}(\hat{\tilde{\beta}}_g - \hat{\tilde{\beta}}_0) \tag{1.12}$$

and

$$\widehat{ATE_{G,0}} = (-\hat{\tilde{\alpha}}_0) + \bar{\mathbf{x}}(-\hat{\tilde{\beta}}_0), \tag{1.13}$$

where $\hat{\tilde{\alpha}}_g$, $\hat{\tilde{\alpha}}_0$, $\hat{\tilde{\beta}}_g$, and $\hat{\tilde{\beta}}_0$ are respectively the consistent estimates of $\tilde{\alpha}_g$, $\tilde{\alpha}_0$, $\tilde{\beta}_g$, and $\tilde{\beta}_0$ from Procedure 1.1 applied on (1.9), and $\bar{\mathbf{x}}$ is consistent estimate of $E(\mathbf{x})$ just as in (1.5).

1.4.2 CF Estimation

CF estimation is more involved compared to the IV estimation in subsection 1.4.1. This is because I first need to derive the estimating equation of CF method. This estimating equation is based on the expectation of the observed outcome y conditional on the observed variables $(\mathbf{d}, \mathbf{x}, \mathbf{z})$, $E(y|\mathbf{d}, \mathbf{x}, \mathbf{z})$. To prevent equation clutter, I collected all the derivations in appendix A. Thus, for derivations, refer to appendix A. Having said that, (A.7) in appendix A gives me the estimating equation of CF method because I can always write

$$\begin{aligned} y \;=\; & \sum_{j=0}^{G} d_j \alpha_j + \sum_{k=0}^{G} d_k \mathbf{x} \beta_k + \\ & + \sum_{g=0}^{G} [-\eta_{g,g} d_g log(\Lambda_g)] + \sum_{g \neq 0} d_g \eta_{g,0} M_0 + \sum_{g \neq 1} d_g \eta_{g,1} M_1 + \cdots + \\ & + \sum_{g \neq G} d_g \eta_{g,G} M_G + \epsilon, \end{aligned} \tag{1.14}$$

where $E(\epsilon|\mathbf{d},\mathbf{x},\mathbf{z}) = 0$, $\Lambda_j = exp(\boldsymbol{z}\gamma_j)/\sum_{r=0}^{G} exp(\boldsymbol{z}\gamma_r)$, and $M_j = \Lambda_j log(\Lambda_j)/(1-\Lambda_j)$ for $j = 0, 1, \ldots, G$. So I can prescribe the following three-stage estimation procedure to estimate ATEs:

Procedure 1.2

1. Same as in Procedure 1.1.

2. Run the regression of y_i on d_{i0}, d_{i1},..., d_{iG}, $d_{i0}\mathbf{x}_i$, $d_{i1}\mathbf{x}_i$,..., $d_{iG}\mathbf{x}_i$, $-d_{0_i} log(\hat{\Lambda}_{0_i})$, $-d_{1_i} log(\hat{\Lambda}_{1_i})$, ..., $-d_{G_i} log(\hat{\Lambda}_{G_i})$, $d_{1_i}\hat{M}_{0_i}$, $d_{2_i}\hat{M}_{0_i}$, ..., $d_{G_i}\hat{M}_{0_i}$, $d_{0_i}\hat{M}_{1_i}$, $d_{2_i}\hat{M}_{1_i}$, $d_{3_i}\hat{M}_{1_i}$, ..., $d_{G_i}\hat{M}_{1_i}$, ..., $d_{0_i}\hat{M}_{G_i}$, $d_{1_i}\hat{M}_{G_i}$, ..., and $d_{G-1_i}\hat{M}_{G_i}$.

3. Same as in Procedure 1.1,

where $\hat{\Lambda}_{g_i} = exp(\mathbf{z}_i\hat{\gamma}_g)/\sum_{r=0}^{G} exp(\mathbf{z}_i\hat{\gamma}_r)$ and $\hat{M}_{g_i} = \hat{\Lambda}_{g_i} log(\hat{\Lambda}_{g_i})/(1-\hat{\Lambda}_{g_i})$ for $g = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$.

Unlike the estimates from IV method, CF method's estimates are theoretically robust to heterogeneity in u_g at least from a consistency standpoint. Under A.1.1 through A.1.3, CF method yields consistent estimates because the addition of the control function terms renders d_g exogenous in (1.14) for $g = 0, 1, \ldots, G$.

1.4.3 The Model with $\eta_{g,j} = \eta_j$: Estimation

ATEs and their estimates are calculated just the same way as in the initial model. The main estimation approaches are again IV and CF methods. Consider the observed outcome:

$$\begin{aligned} y &= d_0 y_0 + d_1 y_1 + \ldots + d_G y_G \\ &= \sum_{j=0}^{G} d_j \alpha_j + \sum_{j=0}^{G} d_j \mathbf{x}\beta_j + u', \end{aligned} \tag{1.15}$$

where $u' = d_0u_0 + d_1u_1 + \cdots + d_Gu_G$. However, by A.1.3′, $u_g = u$ for $g = 0, 1, \ldots, G$ (homogeneous counterfactual errors), and thereof $u' = u$. (1.8) and (1.15) are practically the same, so IV estimation (Procedure 1.1) can be used in this special model. However, I expect that IV approach under the condition $\eta_{g,j} = \eta_j$ would yield consistent estimates because the error term u' in (1.15) does not depend on the binary treatment status indicators d_g but only on $u_g = u$. For practical purposes, the reformulation of (1.15) is just as the reformulation of (1.8).

As before, the CF estimating equation is based on the expectation of the observed outcome y conditional on the observed variables $(\mathbf{d}, \mathbf{x}, \mathbf{z})$, $E(y|\mathbf{d}, \mathbf{x}, \mathbf{z})$. To prevent equation clutter, I again collected all the derivations in appendix A. Thus, for derivations, refer to appendix A. In this special model with $\eta_{g,j} = \eta_j$, the estimating equation of CF method gets simplified. As seen in (A.10) in appendix A, I can write the estimating equation of CF method as follows:

$$y = \sum_{g=0}^{G} d_g \alpha_j + \sum_{g=0}^{G} d_g \mathbf{x} \beta_g + \sum_{g=0}^{G} \eta_g r_g + \varepsilon, \tag{1.16}$$

where $r_g = \left[(1 - d_g)M_g - d_g log(\Lambda_g)\right]$, $M_g = \Lambda_g log(\Lambda_g)/(1-\Lambda_g)$, $\Lambda_g = exp(\boldsymbol{z}\gamma_g)/\sum_{r=0}^{G} exp(\boldsymbol{z}\gamma_r)$, and $E(\varepsilon|\mathbf{d}, \mathbf{x}, \mathbf{z}) = 0$ for $g = 0, 1, \ldots, G$. Then, I can prescribe the following three-stage estimation procedure to estimate ATEs:

Procedure 1.2′

1. Estimate the predicted probabilities $\hat{\Lambda}_{g_i}$ from a MNL of w_i on $\mathbf{z}_\mathrm{i}$ and then obtain $\hat{r}_{g_i}$.
2. Run the regression of y_i on d_{0_i}, d_{1_i},..., d_{G_i}, $d_{0_i}\mathbf{x}_i$, $d_{1_i}\mathbf{x}_i$,..., $d_{G_i}\mathbf{x}_i$, $\hat{r}_{0_i}$, $\hat{r}_{1_i}$, ..., $\hat{r}_{G_i}$.
3. Same as in Procedure 1.1,

where $\hat{r}_{g_i} = [(1 - d_{g_i})\hat{M}_{g_i} - d_{g_i} log(\hat{\Lambda}_{g_i})]$, $\hat{M}_{g_i} = \hat{\Lambda}_{g_i} log(\hat{\Lambda}_{g_i})/(1 - \hat{\Lambda}_{g_i})$, and $\hat{\Lambda}_{g_i} = exp(\mathbf{z}_i\hat{\gamma}_g)/\sum_{r=0}^{G} exp(\mathbf{z}_i\hat{\gamma}_r)$ for $g = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$. Under A.1.1′, A.1.2′, and A.1.3′, CF method yields consistent estimates because the addition of the control function terms, r_g, renders d_g exogenous in (1.16) for $g = 0, 1, \ldots, G$.

1.5 Asymptotic Normality Results

CF method is indeed a two-step M-estimator that solves the problem

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} (y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{z}_i, \hat{\gamma}), \theta))^2/2, \tag{1.17}$$

where $\hat{\gamma} = (\hat{\gamma}_0', \hat{\gamma}_1', \ldots, \hat{\gamma}_G')'$ is the $(G+1)k \times 1$ vector of $\sqrt{N} - consistent$ and asymptotically normal first stage conditional MLE (CMLE) estimates from the MNL of w_i on $\mathbf{z}_i$ for $i = 1, 2, \ldots, N$. However, the first stage estimates does not have to be consistent as long as they converge in $plim$, i.e., $\hat{\gamma} \xrightarrow{p} \gamma^*$ where $\gamma^* \in \Gamma \subset \mathbb{R}^{(\mathbf{G+1})\mathbf{k}}$. CMLE solves the problem

$$\max_{\gamma \in \Gamma} \sum_{i=1}^{N} l_i(\gamma), \tag{1.18}$$

where $\gamma = (\gamma_0', \gamma_1', \ldots, \gamma_G')'$ is the $(G+1)k \times 1$ vector of parameters, and $l_i(\gamma) \equiv log(f(w_i|\mathbf{z}_i; \gamma))$, i.e., the conditional log likelihood for observation i, is given below

$$log(f(w_i|\mathbf{z}_i, \gamma)) = \sum_{j=0}^{G} 1[w_i = j] log\left(exp(\mathbf{z}_i\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r) \right). \tag{1.19}$$

To establish that first stage MLE estimates are $\sqrt{N} - consistent$ and asymptotically normal, I will rely on Proposition 7.6 in Hayashi (2000), Theorem 13.1 in Wooldridge (2010), Proposition 7.9 in Hayashi (2000), and Theorem 13.2 in Wooldridge (2010). Theorem 1.1 (Th.1.1) below is just a combination of both the Proposition 7.6 and the Theorem 13.1 and establishes the consistency of CMLE without a compact parameter space.

- **Theorem 1.1 (Th.1.1)**: Let $\{(w_i, \mathbf{z}_i) : i = 1, 2, \ldots\}$ be a random sample with $\mathbf{z}_i \in \mathscr{Z} \subset \mathbb{R}^{\mathbf{k}}$, $w_i \in \mathscr{W} \subset \mathbb{R}$. Let $\Gamma \subset \mathbb{R}^{(\mathbf{G+1})\mathbf{k}}$ be the parameter set, and denote the parametric model for the conditional density, $p(\cdot\,|\mathbf{z})$, as $\{f(\cdot\,|\mathbf{z};\gamma) : \mathbf{z} \in \mathscr{Z}, \gamma \in \Gamma\}$. Let $l : \mathscr{W} \times \mathscr{Z} \times \Gamma \to \mathbb{R}$ be a real-valued function. Assume that (a) $f(\cdot\,|\mathbf{z};\gamma)$ is a true density function with respect to the measure $\mu(dw)$ for all $\mathbf{z}$ and γ, so that $\int_{\mathscr{W}} f(w|\mathbf{z})\mu(dw) = 1, \forall \mathbf{z} \in \mathscr{Z}$ holds; (b) for some $\gamma_o \in \Gamma$, $p_o(\cdot\,|\mathbf{z}) = f(\cdot\,|\mathbf{z};\gamma_o)$, $\forall \mathbf{z} \in \mathscr{Z}$, and the true parameter vector γ_o is the unique solution to $\underset{\gamma\in\Gamma}{max} E[l_i(\gamma)]$; (c) γ_o is an element of the interior of a convex parameter space Γ; (d) for each $\gamma \in \Gamma$, $l(\cdot\,, \gamma)$ is a Borel measurable function on $\mathscr{W} \times \mathscr{Z}$; (e) for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, $l(w, \mathbf{z}, \cdot)$ is concave in γ; and (f) $|l(w, \mathbf{z}, \gamma)| \leq b(w, \mathbf{z})$, $\forall \gamma \in \Gamma$, where $b(\cdot, \cdot)$ is a nonnegative function on $\mathscr{W} \times \mathscr{Z}$ such that $E[b(w, \mathbf{z})] < \infty$. Then there exist a solution to problem in (1.18), the CMLE $\hat{\gamma}$, and $\hat{\gamma} \xrightarrow{p} \gamma_o$.

In appendix A, I will verify the conditions stated in Th.1.1. For a generic consistency proof of extremum estimators without compactness, see Theorem 2.7 in Newey and McFadden (1994, p. 2133). Next, I will state Theorem 1.2 (Th.1.2) below, which is simply a combination of both the Proposition 7.9 and the Theorem 13.2 and establishes the asymptotic normality of CMLE.

- **Theorem 1.2 (Th.1.2)**: Let the definitions and conditions of Th.1.1 hold, and define $\mathbf{B}_o^{\mathbf{F}} \equiv Var[\nabla_\gamma' l_i(\gamma_o)]$. Furthermore, assume that (a) γ_o is an element of the interior of a parameter space Γ;—i.e., $\gamma_o \in int(\Gamma)$; (b) for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, $l(w, \mathbf{z}, \cdot)$ is twice continuously differentiable on $int(\Gamma)$; (c) $E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)] = \mathbf{0}$ and $-E[\mathbf{H}_i^{\mathbf{F}}(\gamma_o)] = Var[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)]$, where $\mathbf{s}_i^{\mathbf{F}}(\gamma) \equiv \nabla_\gamma' l_i(\gamma)$ and $\mathbf{H}_i^{\mathbf{F}}(\gamma) \equiv \nabla_\gamma[\nabla_\gamma' l_i(\gamma)]$; (d) the elements of $\nabla_\gamma[\nabla_\gamma' l(w, \mathbf{z}, \gamma)]$ are bounded in absolute value by a function $b(w, \mathbf{z})$, $\forall \gamma \in \Gamma$, where $b(\cdot, \cdot)$ is a nonnegative function on $\mathscr{W} \times \mathscr{Z}$ such that $E[b(w, \mathbf{z})] < \infty$; and (e) $\mathbf{A}_o^{\mathbf{F}} \equiv$

$-E(\nabla_\gamma[\nabla'_\gamma l_i(\gamma_o)])$ is positive definite. Then

$$\sqrt{N}(\hat{\gamma}-\gamma_o) \quad \xrightarrow{d} \quad Normal(\mathbf{0},(\mathbf{A}_o^{\mathbf{F}})^{-1}\mathbf{B}_o^{\mathbf{F}}(\mathbf{A}_o^{\mathbf{F}})^{-1}). \tag{1.20}$$

Explicitly, the score of the log likelihood for observation i is as follows:

$$\mathbf{s}_i^{\mathbf{F}}(\gamma) \equiv \nabla'_\gamma l_i(\gamma) = \left(\frac{\partial l_i}{\partial \gamma_0}(\gamma), \frac{\partial l_i}{\partial \gamma_1}(\gamma), \ldots, \frac{\partial l_i}{\partial \gamma_{\mathrm{G}}}(\gamma)\right)', \tag{1.21}$$

which is a $(G+1)k \times 1$ vector of partial derivatives of $l_i(\gamma)$ with respect to parameters in γ. The Hessian, $\mathbf{H}_i^{\mathbf{F}}(\gamma) \equiv \nabla_\gamma[\nabla'_\gamma l_i(\gamma)]$, for observation i is the $(G+1)k \times (G+1)k$ matrix of second partial derivatives of $l_i(\gamma)$ with respect to parameters in γ. Thus, using the definitions in Th. 1.2, $\mathbf{A}_o^{\mathbf{F}} \equiv -E[\mathbf{H}_i^{\mathbf{F}}(\gamma_o)]$, and $\mathbf{B}_o^{\mathbf{F}} \equiv Var[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)]$. In appendix A, I show that $E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)] = \mathbf{0}$ and $\mathbf{A}_o^{\mathbf{F}} = \mathbf{B}_o^{\mathbf{F}}$, which are used to reduce the variance expression in (1.20) to the one in (1.22) below:

$$\sqrt{N}(\hat{\gamma}-\gamma_o) \quad \xrightarrow{d} \quad Normal(\mathbf{0},(\mathbf{A}_o^{\mathbf{F}})^{-1}). \tag{1.22}$$

See appendix A for the verification of the conditions stated in Th.1.2 and see Theorem 3.1 in Newey and McFadden (1994, p. 2143) for a generic proof of asymptotic normality of extremum estimators. Now I can move to the second-stage of CF method, which is basically OLS with generated regressors. To establish that second stage estimates are $\sqrt{N}-$*consistent* and asymptotically normal, I use Theorem 1.3 (Th.1.3) and Theorem 1.4 (Th.1.4) respectively. Th.1.3 below is based off Theorem 4.3 in Wooldridge (1994, p. 2653) and establishes the consistency of CF method with a compact parameter space.

- **Theorem 1.3 (Th.1.3)**: Let $\mathtt{w} = (y, \mathbb{X}, \mathbf{v})$ be a random vector with $\mathtt{w} \in \mathtt{W} \subset \mathbb{R}^{\mathbf{M+1}}$ and $M = (l+G+2)(G+1)$. Let $\Theta \subset \mathbb{R}^{\mathbf{M}}$ and $\Gamma \subset \mathbb{R}^{\mathbf{(G+1)k}}$ be the parameter sets.

Let $q(\mathtt{w}, \theta; \gamma) : \mathtt{W} \times \Theta \times \Gamma \to \mathbb{R}$ be a real-valued function. Let $\hat{\gamma}$ be an estimator from a preliminary estimation. Assume that (a) $\hat{\gamma} \xrightarrow{p} \gamma^*$ for some $\gamma^* \in \Gamma$; (b) for a given $\gamma^* \in \Gamma$, the true parameter vector θ_o is the unique solution to $\underset{\theta \in \Theta}{min} E[q_i(\theta; \gamma^*)]$; (c) the parameter space $\Theta \times \Gamma$ is compact; (d) for each $(\theta, \gamma) \in \Theta \times \Gamma$, $q(\cdot\,, \theta; \gamma)$ is a Borel measurable function on $\mathtt{W}$; (e) for each $\mathtt{w} \in \mathtt{W}$, $q(\mathtt{w}, \cdot; \cdot)$ is continuous function on $\Theta \times \Gamma$; and (f) $E[|q(\mathtt{w}_i, \theta; \gamma)|] < \infty \ \forall (\theta, \gamma) \in \Theta \times \Gamma$. Then there exists a solution to problem in (1.17), the CF estimator $\hat{\theta}$, and $\hat{\theta} \xrightarrow{p} \theta_o$.

Compared to Th.1.1, Th.1.3 replaces the convexity assumption on parameter space with the compactness assumption and the concavity of the objective function with its continuity. In appendix A, I will verify the conditions stated in Th.1.3. In addition, see Wooldridge (1994, p. 2730) for a generic consistency proof of two-step M-estimators with compactness. Before I move into the asymptotic normality result, I will introduce some notation. From (1.17), we can see that $q(\mathtt{w}, \theta; \gamma)$ for observation i in Th.1.3 is as follows:

$$q_i(\theta; \gamma) \equiv q(\mathtt{w}_i, \theta; \gamma) \equiv (y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{z}_i, \gamma), \theta))^2/2, \tag{1.23}$$

where $m_i(\mathbf{v}_i(\gamma), \theta) \equiv m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{z}_i, \gamma), \theta) \equiv \mathbb{X}_i\delta + \mathbf{v}_i\lambda$, $\theta = (\delta', \lambda')'$ is the $M \times 1$ vector of parameters, $\mathbb{X}_i$ is the $1 \times (l+1)(G+1)$ vector of regressors in (1.17), and $\mathbf{v}_i$ is the $1 \times (G+1)(G+1)$ vector of generated regressors in (1.17). More explicitly,

$$\begin{aligned}
\mathbb{X}_i &= (\ d_{0_i},\ \cdots,\ d_{G_i},\ d_{0_i}\mathbf{x}_i,\ \cdots,\ d_{G_i}\mathbf{x}_i\ ) \\
\mathbf{v}_i &= (\ -d_{0_i} log(\Lambda_{0_i}),\ \cdots,\ -d_{G_i} log(\Lambda_{G_i}),\ d_{1_i}M_{0_i},\ d_{2_i}M_{0_i},\ \cdots,\ d_{G_i}M_{0_i}, \\
&\qquad d_{0_i}M_{1_i},\ d_{2_i}M_{1_i},\ d_{3_i}M_{1_i},\ \cdots,\ d_{G_i}M_{1_i},\ \cdots,\ d_{0_i}M_{G_i},\ d_{1_i}M_{G_i},\ \cdots, \\
&\qquad , d_{G-1_i}M_{G_i}),
\end{aligned} \tag{1.24}$$

where $\Lambda_{g_i} = exp(\mathbf{z}_i\gamma_g)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)$ and $M_{g_i} = \Lambda_{g_i} log(\Lambda_{g_i})/(1 - \Lambda_{g_i})$ for $g = 0, 1, \ldots, G$. As one can expect, expressions such as $\hat{\Lambda}_{j_i}$ and $\hat{M}_{g_i}$ are just consistent estimates of Λ_{g_i} and

M_{g_i} with $\hat{\gamma}_g$ replacing γ_g in Λ_{g_i} and M_{g_i}. Now, I will state Theorem 1.4 (Th.1.4) that is based off Theorem 4.4 in Wooldridge (1994, p. 2655) and establishes the asymptotic normality of CF method with a compact parameter space.

- **Theorem 1.4 (Th.1.4)**: Let the definitions and conditions of Th.1.3 hold. Furthermore, assume that (a) $\theta_o \in int(\Theta)$ and $\gamma^* \in int(\Gamma)$; (b) $\sqrt{N}(\hat{\gamma} - \gamma^*)$ is bounded in probability —i.e., $\sqrt{N}(\hat{\gamma} - \gamma^*) = O_p(1)$; (c) for each $(\mathtt{w}, \gamma) \in \mathtt{W} \times \Gamma$, $q(\mathtt{w}, \cdot; \gamma)$ is a twice continuously differentiable on $int(\Theta)$; (d) for each $\theta \in \Theta$, $\mathbf{s}(\cdot, \theta; \cdot) \equiv \nabla'_\theta q(\cdot, \theta; \cdot)$ is continuously differentiable on $int(\Gamma)$; (e) for each $(\theta, \gamma) \in \Theta \times \Gamma$, $\mathbf{H}(\cdot, \theta; \gamma) \equiv \nabla_\theta \mathbf{s}(\cdot, \theta; \gamma)$ is a Borel measurable function on $\mathtt{W}$; (f) for each $\mathtt{w} \in \mathtt{W}$, $\mathbf{H}(\mathtt{w}, \cdot; \cdot)$ is continuous on $\Theta \times \Gamma$; (g) $E[\| \mathbf{H}(\mathtt{w}_i, \theta; \gamma) \|] < \infty \; \forall (\theta, \gamma) \in \Theta \times \Gamma$. (h) $\mathbf{A}_o \equiv E[\mathbf{H}(\mathtt{w}_i, \theta_o; \gamma^*)]$ is positive definite; (i) for each $(\theta, \gamma) \in \Theta \times \Gamma$, $\nabla_\gamma \mathbf{s}(\cdot, \theta; \gamma)$ is a Borel measurable function on $\mathtt{W}$; (j) for each $\mathtt{w} \in \mathtt{W}$,$\nabla_\gamma \mathbf{s}(\mathtt{w}, \cdot; \cdot)$ is continuous on $\Theta \times \Gamma$; (k) $E[\| \nabla_\gamma \mathbf{s}(\mathtt{w}_i, \theta; \gamma) \|] < \infty \; \forall (\theta, \gamma) \in \Theta \times \Gamma$; (l) $E[\mathbf{s}_i(\theta_o; \gamma^*)] = \mathbf{0}$, $E[(\mathbf{A}_*^{\mathbf{F}})^{-1} \mathbf{s}_i^{\mathbf{F}}(\gamma^*)] = \mathbf{0}$, and $E[(\mathbf{A}_*^{\mathbf{F}})^{-1} \mathbf{s}_i^{\mathbf{F}}(\gamma^*) \mathbf{s}_i'(\theta_o; \gamma^*)] = \mathbf{0}$. Then,

$$\sqrt{N}(\hat{\theta} - \theta_o) \quad \xrightarrow{d} \quad Normal(\mathbf{0}, (\mathbf{A}_o)^{-1} \mathbf{D}_o (\mathbf{A}_o)^{-1}), \tag{1.25}$$

where $\mathbf{D}_o = \mathbf{B}_o + \mathbf{F}_o \mathbf{T}_o + \mathbf{T}'_o \mathbf{F}'_o + \mathbf{F}_o \mathbf{R}^* \mathbf{F}'_o$, $\mathbf{s}_i(\theta_o; \gamma^*) \equiv \nabla'_\theta q(\mathtt{w}_i, \theta_o; \gamma^*)$, $\mathbf{A}_o \equiv E[\nabla_\theta \mathbf{s}_i(\theta_o; \gamma^*)] \equiv E[\mathbf{H}_i(\theta_o; \gamma^*)]$, $\mathbf{B}_o \equiv E[\mathbf{s}_i(\theta_o; \gamma^*) \mathbf{s}_i'(\theta_o; \gamma^*)]$, $\mathbf{F}_o \equiv E[\nabla_\gamma \mathbf{s}_i(\theta_o; \gamma^*)]$, $\mathbf{T}_o \equiv E[\mathbf{r}_i(\gamma^*) \mathbf{s}_i'(\theta_o; \gamma^*)]$, $\mathbf{R}^* \equiv E[\mathbf{r}_i(\gamma^*) \mathbf{r}_i'(\gamma^*)]$, $\mathbf{r}_i(\gamma^*) = (\mathbf{A}_*^{\mathbf{F}})^{-1} \mathbf{s}_i^{\mathbf{F}}(\gamma^*)$, and $\mathbf{A}_*^{\mathbf{F}} \equiv -E(\nabla_\gamma [\nabla'_\gamma l_i(\gamma^*)])$. For the derivation of asymptotic variance of $\sqrt{N}(\hat{\theta} - \theta_o)$, refer to the subchapter 12.4 in Wooldridge (2010) or subsections 4.3 and 4.4 in Wooldridge (1994). See appendix A for the verification of the conditions stated in Th.1.4 and see Wooldridge (1994, p. 2730) for a generic asymptotic normality proof of two-step M-estimators with compactness. In appendix A, I also derive the closed forms of the population matrices $\mathbf{A}_o$, $\mathbf{B}_o$, $\mathbf{F}_o$, and $\mathbf{R}^*$ and show $E[\mathbf{r}_i(\gamma^*)] = \mathbf{0}$, $E[\mathbf{s}_i(\theta_o; \gamma^*)] = \mathbf{0}$, and $\mathbf{T}_o \equiv E[\mathbf{r}_i(\gamma^*) \mathbf{s}_i'(\theta_o; \gamma^*)] = \mathbf{0}$. Since $\mathbf{T}_o = \mathbf{0}$, $\mathbf{D}_o$ in the asymptotic variance of $\sqrt{N}(\hat{\theta} - \theta_o)$ in (1.25) simplifies to $\mathbf{B}_o + \mathbf{F}_o \mathbf{R}^* \mathbf{F}'_o$.

Let's construct the following estimators for $\mathbf{A}_o$, $\mathbf{B}_o$, $\mathbf{F}_o$, and $\mathbf{R}^*$:

$$\hat{\mathbf{A}} = N^{-1}\sum_{i=1}^{N}\mathbf{H}_i(\hat{\theta};\hat{\gamma}), \tag{1.26}$$

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\mathbf{s}_i(\hat{\theta};\hat{\gamma})\mathbf{s}_i'(\hat{\theta};\hat{\gamma}), \tag{1.27}$$

$$\hat{\mathbf{F}} = N^{-1}\sum_{i=1}^{N}\nabla_\gamma\mathbf{s}_i(\hat{\theta};\hat{\gamma}), \quad and \tag{1.28}$$

$$\hat{\mathbf{R}} = N^{-1}\sum_{i=1}^{N}\mathbf{r}_i(\hat{\gamma})\mathbf{r}_i'(\hat{\gamma}). \tag{1.29}$$

Define $\hat{\mathbf{D}} \equiv \hat{\mathbf{B}} + \hat{\mathbf{F}}\hat{\mathbf{R}}\hat{\mathbf{F}}'$. Then, using the analogy principle and Lemma 1 in appendix A, a consistent estimator for $Avar\sqrt{N}(\hat{\theta} - \theta_o)$ is $\hat{Avar}\sqrt{N}(\hat{\theta} - \theta_o) = (\hat{\mathbf{A}})^{-1}\hat{\mathbf{D}}(\hat{\mathbf{A}})^{-1}$. The asymptotic standard errors of CF estimates can be obtained from the matrix $\hat{Avar}(\hat{\theta}) = (\hat{\mathbf{A}})^{-1}\hat{\mathbf{D}}(\hat{\mathbf{A}})^{-1}/N$.

1.5.1 Method of Moments Framework

In his paper, Newey (1984) showed that two-step estimators could be interpreted as members of generalized method of moments (GMM) estimators for the purpose of obtaining asymptotic variance matrix. This perspective not only provides consistent parameter estimates but also yields consistent standard errors of parameter estimates. By jointly estimating all parameters in just one step, consistent standard errors are obtained without deriving the asymptotic variance matrix of a two-step estimator. For this reason, GMM estimation gives an alternative way to get consistent standard errors of the parameters in CF regression. GMM estimation reduces down to method of moments estimation when the number of moments is exactly equal to the number of parameters to be estimated, so I technically utilize method of moments (MoM) in my analysis.

In the first stage of CF method, $\hat{\gamma}$ is the CMLE estimator solving

$$\sum_{i=1}^{N} \mathbf{s}_i^{\mathbf{F}}(\hat{\gamma}) = \mathbf{0}, \tag{1.30}$$

where $\mathbf{s}_i^{\mathbf{F}}(\gamma) = \nabla'_\gamma \sum_{j=0}^{G} 1[w_i = j] log\left(exp(\mathbf{z}_i\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)\right)$. In the second stage, $\hat{\theta}$ is a OLS estimator solving

$$\sum_{i=1}^{N} \mathbf{s}_i(\hat{\theta}; \hat{\gamma}) = \mathbf{0}, \tag{1.31}$$

where $\mathbf{s}_i(\hat{\theta}; \hat{\gamma}) = \nabla'_\theta [y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{z}_i, \hat{\gamma}), \hat{\theta})]^2/2$. Newey (1984) proposes stacking the summands in these first order conditions into the unified function

$$g(\theta, \gamma) = \begin{pmatrix} \mathbf{s}^{\mathbf{F}}(\gamma) \\ \mathbf{s}(\theta; \gamma) \end{pmatrix} \tag{1.32}$$

and then applying MoM using the moment conditions $E[g(\theta, \gamma)] = \mathbf{0}$ to obtain consistent estimates for θ and γ, and valid asymptotic variance matrix of $\hat{\theta}$ and $\hat{\gamma}$.

1.5.2 The Model with $\eta_{g,j} = \eta_j$: Asymptotics

The asymptotic properties of CF estimates can also be obtained using theorems used for the initial model. For consistency and asymptotic normality of CMLE in the first stage, I will still use Th.1.1 and Th.1.2. Th.1.3 and Th.1.4 will be almost the same as before except now $M = (l + 2)(G + 1) + 1$. $\mathbf{v}_i$ will be the $1 \times (G + 1)$ vector of generated regressors. More explicitly, $\mathbf{v}_i = (r_{0_i}, r_{1_i}, \ldots, r_{G_i})$. Using Th.1.3 and Th.1.4 with these changes, the asymptotic standard errors of CF estimates in this special case can still be obtained from the matrix $\hat{Avar}(\hat{\theta}) = (\hat{\mathbf{A}})^{-1}\hat{\mathbf{D}}(\hat{\mathbf{A}})^{-1}/N$, where $\hat{\mathbf{A}}$ and $\hat{\mathbf{D}}$ are defined as in the initial model. A MoM estimation in this case can be applied as described in subsection 1.5.1.

1.6 Hypothesis Testing

Economists often are interested in whether a particular research question is true or not, i.e., does the limited English proficiency have an effect on the earnings of immigrant worker population in the U.S.? To answer questions of this sort, economists set up a hypothesis test with a null hypothesis H_0, a statement often against the idea one would like to accept, an alternative hypothesis H_1, a statement one would like to show evidence for, and a test statistic whose distribution can be calculated under H_0. In this section, I will devise a hypothesis testing framework with hypotheses expressed as a set of restrictions on model parameters and will construct a test statistic based on the score test (a.k.a. Lagrange multiplier test) of Rao (1948).

In my framework, I could also utilize the Wald statistic. However, I choose the score statistic because the Wald statistic generally suffer from lack of invariance to how the non-linear restrictions are constructed and, thereof, yield different hypothesis test results. In addition, note that, due to the heterogenous error terms in my estimating equations, the generalized information matrix equality (GIME) fails (i.e., the expected value of the outer product of score function is not equal to a constant multiple of the expected value of the Hessian function.) And, since GIME fails, the quasi-likelihood ratio statistic does not work in my framework. For more on the drawbacks of the Wald statistic and the quasi-likelihood ratio statistic, refer to section 7.4 of Hayashi (2000) and section 12.6 of Wooldridge (2010).

Following the notation and conditions used in subsection 4.6 of Wooldridge (1994), consider the following null hypothesis

$$H_0 : \mathbf{c}(\theta_o) = \mathbf{0}, \tag{1.33}$$

where $\mathbf{c}(\theta)$ is a $Q \times 1$ vector function of the $M \times 1$ vector θ, and some constraints may be linear while the others are nonlinear. The constrained CF estimator $\tilde{\theta}$ solves the problem

$$\underset{\theta \in \Theta}{min} \sum_{i=1}^{N} (y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{z}_i, \hat{\gamma}), \theta))^2/2 \qquad s.t. \quad \mathbf{c}(\theta) = \mathbf{0}, \tag{1.34}$$

where $\hat{\gamma}$ is as in (1.17). In the assumption below, I will state some conditions on $\mathbf{c}(\theta)$ and on a mapping defined by the restrictions in H_0. Even though it is not necessary to know the explicit form of this mapping, it will help us establish the first order representation of $\tilde{\theta}$.

- **Assumption** 1.4 **(A.**1.4**)**: Assume that (a) $Q \leq M$; (b) $\mathbf{c}(\cdot)$ is continuously differentiable on $int(\Theta)$; (c) $\theta_o \in int(\Theta)$ under H_0; (d) $\mathbf{C}(\theta) \equiv \nabla_\theta \mathbf{c}(\theta)$ is the $Q \times M$ gradient of $\mathbf{c}(\theta)$ with rank Q, and $\mathbf{C}(\theta_o)$ is bounded in probability; (e) there exists a twice continuously differentiable mapping $\mathbf{d} : \mathbb{R}^{\mathbf{M-Q}} \to \mathbb{R}^{\mathbf{M}}$ with $\theta_o = \mathbf{d}(\alpha_o)$ under H_0, where α_o is a $(M-Q) \times 1$ vector in the interior of its compact parameter space $\mathcal{A} \subset \mathbb{R}^{\mathbf{M-Q}}$ under H_0; and (f) $\mathcal{D}(\alpha) \equiv \nabla_\alpha \mathbf{d}(\alpha)$ is the $M \times (M-Q)$ gradient of $\mathbf{d}(\alpha)$ with rank $M-Q$ at $\alpha = \alpha_o$, and $\mathcal{D}(\alpha_o)$ is bounded in probability.

White (1994, p. 138) gives a famous example of where expressing restrictions in H_0 as $\theta = \mathbf{d}(\alpha)$ can be used in econometrics. In simultaneous systems of equations with overidentifying restrictions, θ represents the parameters of the reduced form, and α corresponds to the structural parameters, and $\mathbf{d}$ determines the relation between them. Hence, we can intuitively think of the mapping $\mathbf{d}$ in a similar way. Furthermore, note that the estimator of α_o, $\tilde{\alpha}$, solves the problem

$$\underset{\alpha \in \mathcal{A}}{min} \sum_{i=1}^{N} (y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{z}_i, \hat{\gamma}), \mathbf{d}(\alpha)))^2/2, \tag{1.35}$$

and the constrained CF estimator $\tilde{\theta}$ is equal to $\tilde{\theta} \equiv \mathbf{d}(\tilde{\alpha})$. Now, I will state Theorem 1.5 (Th.1.5) that offers the score test and the score statistic, the Lagrange multiplier (LM) statistic. Th.1.5 is similar to the LM statistic in subsection 4.6 of Wooldridge (1994, p. 2668); however, into this theorem below, I also incorporate the adjustment to take into consideration the estimation of the nuisance parameter γ^* in (1.18).

- **Theorem 1.5 (Th.1.5)**: Let the definitions and conditions of Th.1.2, Th.1.4 and A.1.4 hold. Then under $H_0 : \mathbf{c}(\theta_o) = 0$, (a) $\left(\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma})\right)' \mathbf{A}_o^{-1}\mathbf{C}_o'[\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o']^{-1}\mathbf{C}_o\mathbf{A}_o^{-1}$ $\left(\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma})\right)/N \xrightarrow{d} \chi^2_Q$ and (b) the LM statistic $LM_N \equiv \left(\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma})\right)' \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{C}}'[\tilde{\mathbf{C}}$ $\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{D}}\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{C}}']^{-1}\tilde{\mathbf{C}}\tilde{\mathbf{A}}^{-1}\left(\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma})\right)/N$ is asymptotically χ^2_Q,

where $\mathbf{s}_i(\tilde{\theta};\hat{\gamma})$ is as in Th.1.4 but evaluated at $(\tilde{\theta},\hat{\gamma})$, $\mathbf{A}_o$ and $\mathbf{D}_o$ are as in Th.1.4, $\mathbf{C}_o \equiv \mathbf{C}(\theta_o)$ with $\mathbf{C}(\theta)$ as in A.1.4, $\tilde{\mathbf{A}} = N^{-1}\sum_{i=1}^N \mathbf{H}_i(\tilde{\theta};\hat{\gamma})$ with $\mathbf{H}_i$ as in Th.1.4, $\tilde{\mathbf{C}} = \mathbf{C}(\tilde{\theta})$, $\tilde{\mathbf{D}} = \tilde{\mathbf{B}} + \tilde{\mathbf{F}}\hat{\mathbf{R}}\tilde{\mathbf{F}}'$, $\tilde{\mathbf{B}} = N^{-1}\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma})\mathbf{s}_i'(\tilde{\theta};\hat{\gamma})$, $\tilde{\mathbf{F}} = N^{-1}\sum_{i=1}^N \nabla_\gamma \mathbf{s}_i(\tilde{\theta};\hat{\gamma})$, and $\hat{\mathbf{R}} = N^{-1}\sum_{i=1}^N \mathbf{r}_i(\hat{\gamma})\mathbf{r}_i'(\hat{\gamma})$ as in Th.1.4. See appendix A for the proof of Th.1.5.

1.6.1 The Model with $\eta_{g,j} = \eta_j$: Hypothesis Testing

After CF estimation when $\eta_{g,j} = \eta_j$, hypothesis testing can be conducted in the same way as the initial model: Th.1.5 still holds. As noted before, nothing changes in terms of both Th.1.1 and Th.1.2. As for Th.1.3 and Th.1.4, now $M = (l+2)(G+1)+1$, and $\mathbf{v}_i$ will be the $1 \times (G+1)$ vector of generated regressors. More explicitly, $\mathbf{v}_i = (r_{0_i}, r_{1_i}, \ldots, r_{G_i})$, where r_{g_i} is as in subsection 1.4.3. Using Th.1.3, Th.1.4, and Th.1.5 with these changes, LM_N is still asymptotically χ^2_Q.

1.7 Simulations

In this section, I present some simulation results that place CF and IV methods in section 1.4 side by side and note differences and similarities in terms of their asymptotic performances, specifically asymptotic efficiency, asymptotic unbiasedness, and consistency. The setups for the main model in section 1.2 and for the special case model in subsection 1.2.1 are similar to each other; however, each setup varies a little as I change the distribution of instrument in the latent variable equation and assume $\eta_{g,j} = \eta_j$ for $g, j = 0, 1, 2$. For

the sake of computational simplicity, I adopt a scheme in which there is only one covariate in the counterfactual outcome equation, i.e., $\mathbf{x} = x$ and only one instrument in the latent variable equation, i.e., $\mathbf{z} = z$. In addition, the treatment variable w takes on only three values, and each treatment group comprises at least about 30 percent for each simulation setting. Lastly, I introduce misspecification into the main model by ignoring an instrument in the latent variable equation and examine its consequences.

1.7.1 Data Generating Process

In my simulation analysis, I used five different data generating processes (DGPs); one for the main model in section 1.2 with asymmetric instrument, one for the main model with symmetric instrument, one for the special case model in subsection 1.2.1 with asymmetric instrument, one for the special case model with symmetric instrument, and one for the main model with asymmetric instrument and misspecification. The setup for the DGP of the main model with asymmetric instrument is as follows:

$$w \in \{0, 1, 2\},$$

$$d_g = 1[w = g], \quad g \in \{0, 1, 2\},$$

$$a_g \sim Gumbel(0, 1), \quad g \in \{0, 1, 2\},$$

$$\gamma_0 = 1, \quad \gamma_1 = 5, \quad and, \quad \gamma_2 = 9,$$

$$l_0 = 1, \quad l_1 = 5, \quad and, \quad l_2 = 3,$$

$$\mathbf{z} = z \sim \chi^2(2) - 2,$$

$$w_g^* = l_g + \gamma_g z + a_g, \quad g \in \{0, 1, 2\},$$

$$w = g \quad iff \quad w_g^* \geq w_j^*, \quad \forall j \neq g \quad and \quad g, j \in \{0, 1, 2\},$$

$$e_g \sim N(0, 4), \quad g \in \{0, 1, 2\},$$

$$\eta_{0,0} = 0.05, \quad \eta_{0,1} = 0.10, \quad and, \quad \eta_{0,2} = 0.15,$$

$$\eta_{1,0} = 3.05, \quad \eta_{1,1} = 3.10, \quad and, \quad \eta_{1,2} = 3.15,$$

$$\eta_{2,0} = 6.05, \quad \eta_{2,1} = 6.10, \quad and, \quad \eta_{2,2} = 6.15,$$

$$u_g = \sum_{j=0}^{2} \eta_{g,j} a_j + [-\sum_{j=0}^{2} \eta_{g,j} E(a_j)] + e_g, \quad g \in \{0, 1, 2\},$$

$$\mathbf{x} = x \sim N(0, 1),$$

$$\alpha_0 = 1, \quad \alpha_1 = 2, \quad and, \quad \alpha_2 = 3,$$

$$\beta_0 = 6, \quad \beta_1 = 7, \quad and, \quad \beta_2 = 8,$$

$$y_g = \alpha_g + x\beta_g + u_g, \quad g \in \{0, 1, 2\},$$

$$and \quad y = d_0 y_0 + d_1 y_1 + d_2 y_2.$$

For the main model with symmetric instrument, the DGP setup is very similar to the one above. However, I make the following modifications:

$$l_0 = 1, \quad l_1 = 5.2, \quad and, \quad l_2 = 2,$$

$$\mathbf{z} = z \sim N(0, 4),$$

$$\eta_{1,0} = 0.55, \quad \eta_{1,1} = 0.60, \quad and \quad \eta_{1,2} = 0.65.$$

The setup for the DGP of the special case model with asymmetric instrument is as follows:

$$w \in \{0,1,2\},$$

$$d_g = 1[w = g], \quad g \in \{0,1,2\},$$

$$a_g \sim Gumbel(0,1), \quad g \in \{0,1,2\},$$

$$\gamma_0 = 1, \quad \gamma_1 = 5, \quad and, \quad \gamma_2 = 9,$$

$$l_0 = 1, \quad l_1 = 5, \quad and, \quad l_2 = 3,$$

$$\mathbf{z} = z \sim \chi^2(2) - 2,$$

$$w_g^* = l_g + \gamma_g z + a_g, \quad g \in \{0,1,2\},$$

$$w = g \quad iff \quad w_g^* \geq w_j^*, \quad \forall j \neq g \quad and \quad g, j \in \{0,1,2\},$$

$$e \sim N(0,4),$$

$$\eta_0 = 0.05, \quad \eta_1 = 3.05, \quad and, \quad \eta_2 = 6.05,$$

$$u_g = u = \textstyle\sum_{j=0}^{2} \eta_j a_j + [-\sum_{j=0}^{2} \eta_j E(a_j)] + e, \quad g \in \{0,1,2\},$$

$$\mathbf{x} = x \sim N(0,1),$$

$$\alpha_0 = 1, \quad \alpha_1 = 2, \quad and, \quad \alpha_2 = 3,$$

$$\beta_0 = 6, \quad \beta_1 = 7, \quad and, \quad \beta_2 = 8,$$

$$y_g = \alpha_g + x\beta_g + u, \quad g \in \{0,1,2\},$$

$$and \quad y = d_0 y_0 + d_1 y_1 + d_2 y_2.$$

For the special case model with with symmetric instrument, the DGP setup is very similar to the one with asymmetric instrument above. However, I make the following modifications:

$$l_0 = 1, \quad l_1 = 5.2, \quad and, \quad l_2 = 2,$$

$$\mathbf{z} = z \sim N(0, 4),$$

$$\eta_0 = 0.05, \quad \eta_1 = 0.55, \quad and \quad \eta_2 = 6.05.$$

And lastly, for the main model with asymmetric instrument and misspecification in the latent variable equation, the DGP setup is very similar to the one without misspecification. However, I introduce an additional instrument in the latent variable equation and ignore it from the MNL regression of treatment variable on instruments at the first stage. Hence, I make the following modifications:

$$\mathbf{z} = (z_1, z_2)',$$

$$z_1 \sim \chi^2(2) - 2,$$

$$z_2 \sim \chi^2(2) - 2,$$

$$w_g^* = l_g + \gamma_g z_1 + \vartheta_g z_2 + a_g, \quad g \in \{0, 1, 2\},$$

$$\vartheta_0 = \gamma_1, \quad \vartheta_1 = \gamma_2, \quad and \quad \vartheta_2 = \gamma_0,$$

where z_1 and z_2 are scalar instruments in the choice equation for w_g^*, and ϑ_g is a scalar parameter associated with z_2 in w_g^* for $g \in \{0, 1, 2\}$. Note that missing, say, z_2 in w_g^* is practically like putting it into the $i.i.d.$ Gumbel distributed error term a_g, which creates the new error term $a_g' = \vartheta_g z_2 + a_g$ in the latent treatment equation. However, this new error term is not $i.i.d.$ and is very likely not Gumbel distributed anymore, which violates the model condition that the error term in the latent treatment equation is $i.i.d.$ Gumbel distributed. Once this model condition is infringed, w does not follow a MNL distribution and the CF terms used in (1.14) or (1.16) are not correct, which leads to that $E(\epsilon|\mathbf{d}, \mathbf{x}, \mathbf{z}) \neq 0$ in (1.14) and $E(\varepsilon|\mathbf{d}, \mathbf{x}, \mathbf{z}) \neq 0$ in (1.16). In short, it is this model condition violation as a result of

misspecification that creates huge biases on CF estimates as demonstrated in simulation results later. This is interesting in its own right because, for several two-stage estimators, similar misspecifications (i.e., excluding relevant instruments) of the first stage equation (e.g., choice/selection equation) would not cause much of a bias on the second stage (e.g., outcome stage) parameter estimates. In this regard, the most prominent estimator is two-stage least squares with excluded instruments, see McKenzie and McAleer (1994, p. 446) for example. In addition, in a unified fashion, Pagan (1986) explores when two-stage estimators are consistent and when they are not. In proposition 3.2 of this paper, he shows that two-stage estimators based on a misspecified first stage equation result in consistent second stage parameter estimates when the excluded (and included) instruments in the first stage are not correlated with the included variables in the second stage. Even though this is the case in my simulation study, my simulation results in this section below still suggest that CF estimates have big biases under misspecification, which makes my results even more interesting.

Regarding the DGP setups, note that γ_g and l_g both play a role in establishing the percentage of each treatment group in simulations for $g = 0, 1, 2$. γ_g's being distant from each other enough are also critical to obtain strong first stage estimates. With γ_g's being very close each other, one can easily run into identification problems in the first stage estimation. Having $\eta_{g,j}$'s being far away from each other, especially $\eta_{g,j}$'s from different treatment statuses, is also another critical point to create endogeneity in the main model for $g, j = 0, 1, 2$. If $\eta_{g,j}$'s from different treatment statuses get closer and closer to each other, we essentially get closer and closer to make the assumption that $\eta_{g,j} = \eta_j$, and endogeneity issue in the main model gets attenuated or maybe nearly resolved as pointed out in subsection 1.4.3. The usage of instrument z whose distribution is either asymmetric or symmetric in the latent variable equation is also important because, in endogenous binary treatment case, Wooldridge (2008, p. 106; 2010, p. 947) argues that IV method can be consistent for ATEs when the instrument is symmetrically distributed around zero. Hence, I include schemes in simulations that take into consideration this possibility.

1.7.2 Simulation Results

I present my simulation results in two parts: first, asymptotic efficiency outcomes and second, asymptotic unbiasedness and consistency outcomes. The simulation results reported in Tables A.1 through A.20 focus on comparing CF method with IV method in terms of asymptotic efficiency, asymptotic unbiasedness and consistency.

In Tables A.1 through A.20, I report the Monte Carlo (M.C.) estimates for α_g and $ATE_{h,0}$, bias in the M.C. estimate for ATEs, analytical standard errors for α_g and $ATE_{h,0}$ in CF method and uncorrected standard errors for α_g and $ATE_{h,0}$ in IV method (except in Tables A.17 through A.20), bootstrapped standard errors (BS. SEs) and Monte Carlo standard deviations (M.C. SDs) for α_g and $ATE_{h,0}$, and BS. SEs and M.C. SDs for standard errors of α_g and $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$. The analytical standard errors for α_g and $ATE_{h,0}$ in CF method are calculated as in section 1.5. The uncorrected standard errors for α_g and $ATE_{h,0}$ in IV method are directly obtained by following Procedure 1.1. And, when $\eta_{g,j} = \eta_j$ for $j, g = 0, 1, 2$, I do not even need to correct the standard errors for α_g and $ATE_{h,0}$ in IV method, see appendix A for more detailed explanation. In simulations, I use different sample sizes $n = 1000$, $n = 2000$, $n = 5000$ and $n = 10000$ for each DGP setup with the number of M.C. and BS. iterations always equal to 10000.

As for the notation, in Tables A.1-A.20, $\hat{\alpha}_g$ is the parameter estimate for α_g, $\hat{ate}_{h,0}$ is the estimate for $ATE_{h,0}$, and $bias(\hat{ate}_{h,0})$ is the bias in the estimate for $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$. Furthermore, $se(\hat{\alpha}_g)$ is the standard error of parameter estimate for α_g and $se(\hat{ate}_{h,0})$ is the standard error of the estimate for $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$. Since these tables would require a considerable amount of space in the main body of the chapter, I place all simulation tables of this chapter into appendix A.

At this point, it is also important to remember the true values for α_g and $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$ since I often refer them throughout this section. The true values are respectively as follows:

$$\alpha_0 = 1, \quad \alpha_1 = 2, \quad and, \quad \alpha_2 = 3,$$

$$ATE_{1,0} = 1, \quad and \quad ATE_{2,0} = 2.$$

1.7.2.1 Asymptotic Efficiency Outcomes

First, using simulation results from Tables A.1 through A.4, I can check how close the standard error estimator proposed in section 1.5 are to simulated results. In Table A.1, analytical standard errors of CF parameter estimates are fairly close to BS. SEs and especially to M.C. SDs of CF parameter estimates of both α_g and $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$, except the analytical standard errors of CF parameter estimates for α_1 and $ATE_{1,0}$ in Tables A.1 through A.4 and those for α_2 and $ATE_{2,0}$ in Tables A.5 through A.8. As sample size increases, the analytical standard errors get even closer to the BS. SEs and M.C. SDs, and all shrink in magnitude considerably. For example, switching from sample size of 1000 to 10000, the analytical standard error of $\hat{\alpha}_0$ decreases by 73% from .2749 to .0745, the BS. SEs by 63% from .2106 to .0790, and M.C. SDs by 70% from .2543 to .0755. A very similar pattern can be observed in Tables A.5 through A.16 as I change the distribution of instrument z and/or assume $\eta_{g,j} = \eta_j$ for $g, j = 0, 1, 2$. Hence, the analytical standard errors proposed in section 1.5 seem to be working well and to well approximate the standard deviations of the most CF parameter estimates. As a side note, the uncorrected standard errors of the IV parameter estimates in Tables A.1 through A.8 when $\eta_{g,j} \neq \eta_j$ are also not far away from BS. SEs and M.C. SDs of IV estimates of both α_g and $ATE_{h,0}$. With increased sample size, the uncorrected standard errors shrink in magnitude and get closer to the BS. SEs and M.C. SDs of the IV estimates.

Second, from an efficiency standpoint, let's first take into account the models with no misspecification. At this point, it is wiser to consider Tables A.9 through A.16 since IV method is inconsistent when $\eta_{g,j} \neq \eta_j$ as indicated in Tables A.1 through A.8 with large biases in α_g and $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$. However, in Tables A.9 through A.16,

neither CF method nor IV method has biases in parameter estimates. In Table A.9, the simulation results show that the analytical standard errors, BS. SEs, and M.C. SDs of the CF estimates are lower than the uncorrected standard errors, BS. SEs, and M.C. SDs of the IV estimates, respectively. Similarly, BS. SEs and M.C. SDs of the standard errors of CF estimates are also lower than those of the IV estimates. For instance, in Table A.9, the BS. SE of the CF parameter estimate $\hat{\alpha}_0$ is 4% lower than that of the IV estimate, and the M.C. SD of the CF parameter estimate 5% lower. Furthermore, again in Table A.9, the BS. SE of the standard error of CF parameter estimate $\hat{\alpha}_0$ is 12% lower than that of the IV estimate, and the M.C. SD of the standard error of CF parameter estimate 11% lower. This very alike pattern is persistent in Tables A.10 through A.16 as I switch the distribution of instrument z from asymmetric to symmetric and/or increase the sample size. As a result, when there is no misspecification, the simulation results demonstrate that the CF method performs better compared to the IV method from the perspective of efficiency: The results suggest that CF method estimates the parameters of interest, ATEs, and their standard errors more precisely than does IV method.

Now let's look at the models with misspecification in Tables A.17 through A.20. When sample size is 1000, the simulation results in regard to efficiency are very suggestive: IV method has sharper estimates than does the CF method for all parameters of interest, ATEs, and (almost all) their standard errors. For example, in Table A.17, the BS. SE of the IV parameter estimate $\hat{\alpha}_1$ is around 48% lower than that of the CF estimate, and the M.C. SD of the IV parameter estimate about 66% lower. Moreover, again in Table A.17, the BS. SE of the standard error of IV parameter estimate $\widehat{ate_{10}}$ is about 29% lower than that of the CF estimate, and the M.C. SD of the standard error of IV parameter estimate around 55% lower. As the sample size goes up in Tables A.18 through A.20, the same pattern is still observed in favor of IV method. And the BS. SEs and M.C. SDs of all estimates and of the standard errors of all estimates get smaller. For instance, the BS. SE (M.C. SD) of the IV parameter estimate $\hat{\alpha}_2$ goes down from 1.4534 (1.1045) in Table A.17 to .3257 (.3336) in

Table A.20. As a consequence, when there is misspecification, the simulation results show that the IV method outperforms the CF method in terms of efficiency: The results suggest that IV method estimates the parameters of interest, ATEs, and their standard errors more precisely than does CF method, which is just the opposite of the findings when there is no misspecification.

1.7.2.2 Asymptotic Unbiasedness and Consistency Outcomes

In a M.C. simulation, the average of parameter estimates over a specific number of iterations (conventionally 10000) is the M.C. simulation estimate of the expected value of those estimates. Therefore, if I repeat this M.C. simulation with a fixed number of iterations but increasing sample size and the M.C. estimates get closer and closer to true parameter values as the sample size increases, this would be suggestive of the asymptotic unbiasedness of the estimator in question. In addition, if the standard errors of the parameter estimates get smaller and smaller on top of their being closer and closer to true parameter values, this would be indicative of the consistency of the estimator in question.

First, in Tables A.1 through A.4, the simulation results show that M.C. simulation estimates from CF method of both α_g and $ATE_{h,0}$ are very close to the true values, whereas the ones from IV method are not that close at all for $g = 0, 1, 2$ and $h = 1, 2$. For example, in Table A.1, M.C. simulation estimates from IV method for α_0, α_1, and α_2 are respectively 1.3394 (about 34% higher than the true value), 1.9078 (around 5% lower than the true value), and 2.9018 (around 3% lower than the true value) and are all off the true values, causing severe biases in ATE estimates (about 43% lower in estimated $ATE_{1,0}$ and 22% lower in estimated $ATE_{2,0}$.) On the other hand, M.C. simulation estimates from CF method of both α_g and $ATE_{h,0}$ in Table A.1 are not off the true values at all with almost no biases. As the sample size increases from Table A.1 to Table A.4, M.C. simulation estimates from IV method do not improve on the biases; however, their BS. SEs and M.C. SDs get closer to zero just as

those from CF method. A very similar pattern can also be seen in Tables A.5 through A.8 as I switch the distribution of instrument z from asymmetric to symmetric and/or increase the sample size. As a result, the simulation results indicate that CF method is asymptotically unbiased and consistent while IV method is asymptotically biased and inconsistent, which is supportive of the conjecture I made in subsection 1.4.1.

Second, in Tables A.9 through A.12 when $\eta_{g,j} = \eta_j$ for $g, j = 0, 1, 2$, the simulation results demonstrate that M.C. simulation estimates of α_g and $ATE_{h,0}$ from both CF method and IV method are very close to the true values for $g = 0, 1, 2$ and $h = 1, 2$. For example, in Table A.9, M.C. simulation estimates from both CF method and IV method for α_0, α_1, and α_2 are all accurate up to two decimal places with almost no biases in ATE estimates. As the sample size increases from Table A.9 to Table A.12, M.C. simulation estimates from both CF method and IV method continue keeping their accuracy with almost no bias, and their BS. SEs and M.C. SDs shrink in size and get closer and closer to zero. This very pattern is also seen in Tables A.13 through A.16 as I switch the distribution of instrument z from asymmetric to symmetric and/or increase the sample size. As a conclusion, the simulation results demonstrate that, when $\eta_{g,j} = \eta_j$, both CF method and IV method are asymptotically unbiased and consistent, which corroborates another conjecture I made in subsection 1.4.3.

Under misspecification, the simulation results in Tables A.17 through A.20 indicate that M.C. simulation estimates of α_g and $ATE_{h,0}$ from both CF method and IV method are off the true values for $g = 0, 1, 2$ and $h = 1, 2$. For instance, in Table A.17, M.C. simulation estimates from CF method for α_0, α_1, and α_2 are respectively .9310 (about 7% lower than the true value), 1.5886 (around 21% lower than the true value), and 4.4840 (around 50% higher than the true value) and are all off the true values, causing severe biases in ATE estimates (about 36% lower in estimated $ATE_{1,0}$ and 78% higher in estimated $ATE_{2,0}$.) As indicated earlier, M.C. simulation estimates from IV method are also not close to the true values. However, biases in its estimates are lower than those in the estimates of CF method except in the estimates for α_0 and $ATE_{1,0}$. For example, in Table A.17, M.C. simulation estimates

from IV method for α_1, α_2, and $ATE_{2,0}$ are respectively 1.6833 (about 16% lower than the true value), 3.0697 (around 2% higher than the true value), and 1.7553 (around 12% lower than the true value) but these biases are all smaller than their counterparts from CF method. As the sample size increases from Table A.17 to Table A.20, M.C. simulation estimates from both IV and CF methods do not improve on the biases; however, their BS. SEs and M.C. SDs get smaller. As a result, the simulation results indicate that, under misspecification, CF method is asymptotically more biased compared to IV method, and both methods are inconsistent.

1.8 Empirical Application

In this section, I illustrate the role of limited English proficiency (LEP) in determining wages of Hispanic workers in the USA. To this end, I revisit the 1% Public Use Microdata Series (PUMS) of the 1990 U.S. Census and utilize a subsample that is constructed from data used by Gonzalez (2005). My aim is just to apply OLS, CF, IV, and nonparametric bound analysis to the estimation of how LEP influences wages of Hispanic workers in the USA and to make a comparison of their performances, not to offer a detailed or decisive evaluation of the factors that explain wages of Hispanic workers in the USA.

1.8.1 Background on the Economics of Language Skills

As a result of the growth of international immigrant flows into the host destination countries after the Second World War, researchers started investigating the immigrant behavior (e.g., how integrated the immigrants were with the members of their host society and what factors were crucial to improve the immigrants' integration period). As part of immigrant adjustment time, among other things, economists regarded language abilities as a part of human capital - for example, see Carliner (1981) and Grenier (1984) - and were interested

in questions such as: Would investing in language acquisition result in higher wages for immigrants? If it would, how proficient should immigrants be in the destination language to really benefit from higher wages? As in many other host destination countries where English is the official language of communication, Hispanic workers' deficiencies in English can lead to negative consequences, particularly lower earnings, in the U.S. labor market. Previous studies showed that Hispanic workers in the USA earn less than Non-Hispanic Whites due to low levels of schooling (see Gwartney and Long, 1978), discrimination (see Reimers, 1983), general assimilation problems for immigrants who were born outside the USA (see Borjas, 1985), and LEP (see McManus *et al.*, 1983).

In general, the human earnings analyses are rooted in the human capital earnings function, and the earnings regressions out of this function include the natural logarithm of earnings as its dependent variable and some other covariates such as education, experience (and its square), marital status, ethnicity, duration in the destination, and destination language proficiency. Early studies from 1980s used OLS to explore the impact of LEP on earnings in the USA. McManus *et al.* (1983) found that LEP curtails the common wage increases associated with schooling and experience and provided evidence for that Hispanic male workers in skill occupations where wages are generally the highest earn less than their domestic counterparts due to LEP. Reimers (1983) found that Puerto Ricans who do not speak and understand English well have wage penalties of 20%. Grenier (1984) found that the effect of LEP is a 14.6% decrease in wages, which explains that language attributes greatly explain the mean wage differential between Hispanics and non-Hispanics. Kossoudji (1988) concluded that Hispanics suffer from the economic cost to LEP more than do Asians at every skill level, with decreased earnings reaching up to 66% for sales workers and within-occupational wage gap of 18.3% even in service industry. Tainer (1988) found that each unit improvement in English speaking abilities of Hispanics leads to a 17.4% rise in their annual earnings. Chiswick and Miller (2002) indicated that the foreigners who are born in non-English speaking countries but fluent in English earn 14.4% more than those who are not

fluent in English. Lewis (2011) found that there is a wage penalty of 17% for immigrants who cannot speak English very well. Other studies around the globe also report similar findings for other countries: Immigrants benefit from the destination language proficiency in term of higher wages. Chiswick and Miller (1985) found that, in Australia, immigrants with poor English skills have earnings 4.7% lower than those with good English skills. Carliner (1981) showed that, in Montreal, Canada, immigrants who can speak neither English nor French earn 36.3% lower than those can speak English monolingually. Dustmann (1994) pointed out that, in Western Germany, male immigrants who speak German well or very well have earnings 6.9% higher than earnings of those who speak German badly or not at all, creating a considerable improvement in terms of the earnings position of male immigrants proficient in German. Chiswick and Repetto (2000) reported that, in Israel, compared to immigrant men who could not speak Hebrew at all, those who can speak Hebrew only increase their earnings by 20.8%. Lastly, Leslie and Lindley (2001) found that, in the UK, male immigrants who are fluent in English have a wage increase of 16.9% over those who have poor English skills.

One weakness of the these early studies is that they do not offer a solution to ability bias. It is probable that workers with higher innate ability are more likely to earn higher wages, to invest in language capital, and to speak English (or another foreign language) more proficiently than are workers with lower innate ability, which indicates that there may be some correlation between English skills and unobserved innate ability. Since unobserved innate ability is expected to affect both language skills and earnings, the OLS coefficient estimates of language skills might be biased. The concern over this weakness of OLS naturally made many researchers utilize alternative estimation methods such as IV while analyzing the effect of language proficiency on earnings. Chiswick and Miller (1995) found that, in the USA, the effect of language fluency on immigrant earnings goes up from 16.9% to 57.1% when IV is used in place of OLS. Chiswick (1998) showed that, in Israel, the immigrants who speak Hebrew on a daily basis as their only or primary language earn 35.08% (11%) higher than

those who cannot when IV (OLS) method is adopted. Shields and Price (2002) noted that, among male immigrants in the UK, fluency in English raises the mean occupational wage by 16.5% (8.9%) when IV (OLS) method is used for estimation. Dustmann and van Soest (2002) provided evidence for that, in Germany, the male immigrants with good or very good German speaking abilities have a wage premium of 14% (5%) over those with intermediate or poor German speaking abilities when IV (OLS) method is utilized. Budra and Swedberg (2012) pointed out that, in Spain, the influence of Spanish fluency on the earnings of male immigrants is about 27% based on a benchmark IV method, whereas OLS estimate is only 4.8%. Finally, Di Paolo and Raymond (2012) studied the immigrants in Catalonia, Spain, and their IV (OLS) estimates indicated that the monthly earnings of individuals who speak and write Catalans are about 18% (7.5%) higher than those who cannot.

Looking at the results of these studies above, one can argue that representative premiums in immigrant earnings as a result of being fluent in host destination language are mostly between 5% (20%) and 20% (50%) based off OLS (IV). Hence, there is a considerable difference between OLS and IV estimates for the earnings premium to host destination language proficiency: the OLS estimates are usually smaller than the IV estimates, and can be considered underestimates of the true values. The observation that IV estimates are usually larger than OLS estimates is actually fairly common in the literature, and one possible explanation to this phenomenon is the dominance of downward measurement error bias (e.g., misclassification errors coming from self-reported language proficiency data) over upward unobserved heterogeneity bias (e.g., innate ability common to both host destination language abilities and immigrant earnings). For more on this, see Dustmann and van Soest (2002), Dustmann and Fabbri (2003), Bleakley and Chin (2004), and Yao and van Ours (2015). Although the majority of studies probing the impact of language abilities on immigrant earnings uses OLS and IV approaches, some researchers employed different methods to shed more light on the topic and to provide a new perspective with the help of recent developments in econometrics. For instance, Dustmann and van Soest (2001) estimated the earnings equation (e.g., a

random effects panel data model) of male immigrants in Germany by maximum likelihood, together with the equation for speaking proficiency (e.g., a random effects ordered probit model) and found that speaking fluency among male immigrants in Germany can result in a wage increase of 7.3 percentage points. Berman, Lang and Siniver (2003) used the first-difference estimator in a fixed effects panel data framework while examining the effect of language acquisition on earnings and found that the earnings of male immigrants from the Soviet Union to Israel in 1994 who speak Hebrew very well are 23% higher than the earnings of those who cannot speak Hebrew at all. Dustmann and Fabbri (2003) combined a matching estimator based on the propensity score of being proficient in English with an IV estimator, and their results showed that proficiency in English leads to 35.6% higher incomes. In the spirit of Manski and Pepper (2000), Gonzalez (2005) utilized nonparametric bounds to analyze the effects of LEP on wages for Hispanic workers in the USA, and concluded that, on average, the wage premium from developing English proficiency from "not at all" to "very well" is 39% under both monotone treatment response and monotone treatment selection. Chiswick, Lee, and Miller (2005) used a first-difference inertia model to analyze the earnings growth of adult male immigrants in Australia and found that male immigrants with proficiency in English have 23.9% higher income growth compared to those who are not proficient in English. Lastly, Aldashev, Gernandt, and Thomsen (2009) employed augmented OLS to examine the results of better language skills on earnings for foreigners in West Germany by jointly modeling self-selection in participation and employment decisions (e.g., a double hurdle model) and self-selection in economic sector and occupation decisions (e.g., a bivariate probit model). They pointed out that, among the high skilled foreigners, the wage premium from speaking mainly German to their mother tongue is 26.9%. Overall, estimation methods alternative to OLS and IV also yield wage premiums for immigrants with proficiency in host destination language, going up to 39%. In conclusion, it is well accepted in economics literature that improvement in host destination language abilities is associated with increases in earnings among adult male immigrants.

1.8.2 Data

The data source in my empirical analysis is the 1% Public Use Microdata Series (PUMS) of the 1990 U.S. Census, and I utilize a subsample that is constructed from data used by Gonzalez (2005). There are 13 variables and 82250 observations in her original dataset. However, I kept only observations for individuals who reported a Hispanic first ancestry, were between 16 and 64 years old in 1989, had at least the minimum hourly wage in 1990, and worked at least 48 weeks and at least 40 hours per week in 1989. Since year of entry into the USA is in intervals (e.g., between 1950 and 1959) for Hispanic immigrants, and age of arrival into the USA is equal to age in 1989 minus the upper bound of year of entry into the USA; some observations have negative values for age of arrival into the USA. After dropping these observations from the dataset of Gonzalez (2005), the sample for my analysis has only 38779 observations.

As in many applications using Mincer earnings equation, my outcome variable is the natural log of hourly wage where hourly wage is equal to wages or salary income in 1989 divided by the product of weeks worked last year and usual hours worked per week last year. In 1990, the minimum hourly wage was \$1.335 in natural log. I drop the observations for individuals who earned less than the minimum hourly wage in 1990 and worked less than 48 weeks and less than 40 hours per week in 1989 because these individuals are not regular workers. There could be unobserved or observed factors that are specific to these irregular workers, have an impact on their earnings, and I cannot control. This would confound my estimates; however, dropping the observations for irregular workers eliminates that possibility. The treatment data in Gonzalez (2005) are based off the answers to the survey question on "ability to speak English". Hence, it is a self-reported ordinal variable, and this variable might suffer from measurement error, and thereof, misclassification of English proficiency. Taking this deficiency of the treatment variable, I collapsed the original five-category treatment variable (i.e., speaks only English at home; speaks English very well, well; does not speak English well, at all) to a new three-category treatment variable: (1-not

well) does not speak English well and at all; (2-well) speaks English very well and well; and (3-very well) speaks only English at home. The purpose behind this simple recategorization of the English proficiency is to reduce possible measurement error problem in the treatment variable, see Espenshade and Fu (1997), Dustmann and van Soest (2001), and Bleakley and Chin (2004) for more on the benefits of combining language categories. Original employment status variable in Gonzalez (2005) had six categories: civilian employed, at work; civilian employed, with a job but not at work; unemployed; armed forces, at work; armed forces, with a job but not at work; and not in labor force. In my analysis, I regrouped them into three categories: employed, unemployed and not in labor force.

As indicated in previous subsection, the endogeneity problem in wage equation (i.e., the possibility that workers with higher innate ability earn more and speak English better than do workers with lower innate ability) led several economists to use IV method for investigating the impact of English proficiency on immigrant earnings. One commonly used instrument for English proficiency in the literature is immigrants' age at arrival, see, for example, Bleakley and Chin (2004), Bleakley and Chin (2010), Miranda and Zhu (2013), and Yao and van Ours (2015). This choice of instrument is driven by scientific studies on language acquisition which provide evidence for that young people's (e.g., children's) capacity to learn and use languages is generally higher than older people's (e.g., adults'). In conformity with this idea, Akresh and Akresh (2011) showed that, for children of Hispanic immigrants, each additional year spent in the USA leads to higher scores on the passage comprehension, applied problems, and letter-word identification tests. On the other hand, age at arrival might not be a perfect instrument for English proficiency because it is argued that immigrants who come to a host destination country at a younger age might find it less costly to economically assimilate and acculturate to the host country. For example, Gonzalez (2003) noted that Mexican and Latin American immigrants who arrive at younger ages in the USA earn higher wages as a result of completing more years of schooling. Moreover, the immigrant assimilation model of Eckstein and Weiss (2004) supports the idea that age at arrival might have an effect on

earnings through channels other than language: The model claims that immigrants learn more about the host country labor market as they spend more time in the host country, become better at implementing their human capital, and earn more. See Schaafsma and Sweetman (2001) and Borjas (1985, 1995) for more on the relationship between immigrant earnings and age at immigration, and economic assimilation of immigrants.

Despite some concerns over the quality of age at arrival as an instrument for language fluency among immigrants, it is a fairly established and commonly used instrument in the literature, and I believe the wide variety of control variables in my empirical analysis would help to reduce these concerns to the minimum from a statistical point of view. Specifically, in my analysis, I create a variable based off age at arrival in the USA that takes four possible values: 0 (US born immigrants), 1 (arrived as a child-0 to 11 years old), 2 (arrived as a teenager-12 to 17 years old), and 3 (arrived as an adult-18 or older). This new variable, ordered by creation, is excluded from the second stage earnings equation and used only in the first stage language proficiency equation. By running a multinomial logistic regression in this very first stage, I obtain the predicted probabilities of being in one of the three English speaking categories (i.e., very well, well, and not well), and these predicted probabilities are used as instruments for English proficiency in the second stage earnings equation.

There are also several control variables available I use in my regressions. Previous studies show that the effect of LEP may change as individuals' characteristics (e.g., their profession, education, etc...) vary. One such control variable is education that takes on values ranging from 0 to 20 in my dataset, and Chiswick and Miller (2003) indicated that male immigrants in Canada with more years of schooling gain relatively more in earnings from being proficient in English compared to those with less years of schooling. Hence, it is possible that the effect of LEP on earnings of Hispanic immigrants in the USA might differ, depending on their education level. I also create dummy variables for region of birth based off ancestry codes from 1% PUMS, 1990 US Census. In total, there are seven such dummies for US born, Spanish, Mexican, Central American, South American, Puerto Rican, and Cuban

Hispanics. I categorize all occupations into seven groups (i.e., managerial, technical, service, repair, operators, agriculture and military) and create a dummy variable for each except agriculture. Occupation variables may yield interesting results because Berman, Lang and Siniver (2003) found that Hebrew fluency has higher effect on wages in the skilled occupations than in the unskilled occupations. Potential experience in years is another control variable, and McManus *et al.* (1983) suggested that English deficiency has wage penalties on earnings of Hispanic men in the USA, where the penalties increase with potential experience. Gender enters in my regressions as a dummy variable for female immigrants. Lastly, I divide workers into fours groups based on worker class codes from 1% PUMS, 1990 US Census (i.e., employee of a private for profit company, employee of a private for nonprofit organization, government employee and others-mostly self-employed) and create a dummy variable for each except others-mostly self-employed. For a compact version of variable descriptions used in my analysis and summary statistics, see Table A.21 in appendix A.

Before moving to regression results, I look at some of the characteristics of Hispanics in the sample and present, from a descriptive point of view, some interesting observations I gather from Table A.22 available in appendix A. The biggest treatment group is those Hispanic immigrants who speak English well, which comprises just about 61% of the whole sample. The other two treatment groups (i.e., those who do not speak English well and those who speak English very well) are about 20% of the sample each. The average of log hourly wages is 2.18 and is more or less the same across all English proficiency levels with a standard deviation of 0.5. In a descriptive sense, wages are increasing in English proficiency: Hispanics who have a better command of English in speaking on average earn higher wages. Hispanics who speak English better have more schooling: Average years of education increase from 7.8 among Hispanics with deficiency in speaking English to 12.79 among Hispanics with high proficiency in speaking English. Whereas, experience goes down in English proficiency: Hispanics who mastered in speaking English tend to have less potential experience, with 23.35 years for those who do not speak English well to 15.92 years for those who speak

English very well. In line with my expectations, the percentage of female Hispanics goes up from 27.73% among those who do not speak English well to 37.79% among those who speak English very well. Hispanics who speak English better are more prone to work in high skilled occupations: the percentage of Hispanics working in managerial positions (as operators) increases (decreases) from 3.61% (41.34%) for treatment 1 to 23.69% (16.86%) for treatment 3. The percentage of Mexicans goes down from 62.23% among those who do not speak English well to 55.57% among those who speak English very well. Hispanics who have a better mastery in speaking English tend to suffer from unemployment less: the unemployment rate of Hispanics among those who do not speak English well is 1.2 percentage points higher than that among those who speak English very well. Finally, age is slightly decreasing in English proficiency: Younger Hispanics are slightly more inclined to speak English better.

1.8.3 Regression Results

First, I present the first stage regression findings for English speaking proficiency among Hispanics. Second, I move to the estimates from the second stage regression for earnings among Hispanics. Lastly, I share the results that come from nonparametric bound analysis. However, before discussing these results, I have to admit that there are some limitations in my analysis. One of them is that there may be a selection bias on my estimates because of non-randomly selected sample, only considering those who are earning some income. Another limitation is that, just as I expect language proficiency causes to produce increases in earnings, earning more (i.e., being able to save some money for language courses) could reversely cause improvement in language skills, as well, which opens the door to the simultaneity problem. Yet another limitation is that self-reported language measures might suffer from measurement errors that come from either individual respondents (i.e., exaggerating their language skills for personal reasons) or interviewers (i.e., misclassifying respondents'

language ability), which can create a degree of subjectivity in my English proficiency variable. And lastly as pointed out previously, the validity of immigrants' age at arrival as an instrument for English fluency is argued, especially in immigrant assimilation models. Overall, the readers need to keep in mind these limitations of my analysis as I present my results.

Tables A.23 and A.24 available in appendix A report the estimated parameters on Hispanic workers' arrival age in the USA from a multinomial logit regression of English proficiency for several different models/specifications and their likelihood ratio chi square test statistics of goodness of fit. In all the regressions, the outcome is a discrete variable of English proficiency that has three categories: Not well, well, and very well with not well being the base outcome. The arrival age, critical predictor and instrumental variable included in these first stage regressions but excluded from the second stage regressions, takes four values: 0 for US born, 1 for arrived as a child (0 to 11 years old), 2 for arrived as a teenager (12-17 years old) , and 3 for arrived as an adult (18 or older). As to the models/specifications, they all use the same multinomial logit model but with different sets of predictors. Models 1 and 2 have only arrival age included. Model 3 controls for both arrival age and education. Model 4 contains arrival age, education, and gender. Models 3a (4a) has exactly the same variables as Model 3 (4) except that arrival age is not included in Model 3a (4a). Model 5 controls for arrival age, education, gender, and occupation. Model 6 includes arrival age, education, gender, occupation, and ancestry. Model 7 contains arrival age, education, gender, occupation, ancestry, and employment status. Model 8 controls for arrival age, education, gender, occupation, ancestry, employment status, and worker class, which is the full specification regression. Lastly, Models 5a, 6a, 7a, and 8a have exactly the same variables as Models 5, 6, 7, and 8 respectively except that arrival age is not included in Models 5a, 6a, 7a, and 8a. We can think of Model 8 as a language proficiency equation and of its predictors as the determinants of language proficiency. In all models, I use the whole sample with 38779 observations in it.

Even though I have shared only the estimated parameters on arrival age, I can make them available upon request. However, to mention a few of the other parameter estimates (almost all of them statistically significant) from the full specification regression (Model 8), one more year of schooling causes the odds ratio for speaking English very well to not well to increase by about 34%, holding all other variables constant. In the USA, those Hispanic immigrants working in high skill occupations such as managerial positions and in government sector have a significantly higher probability to speak English better. The gender impact is negative against my expectations but statistically insignificant for females: being a female is associated with a little over 5% decrease in the odds ratio for speaking English very well to not well to, ceteris paribus. However, this might be simply because of that female Hispanic immigrants have lower incentives to learn English. In the USA, those Hispanic immigrants from Mexico and Puerto Rico are more likely to speak English more fluently. This supports the idea that immigrants can acquire speaking fluency by exposure to the host destination language through their relatives and friends already working and living in the USA.

In IV method, it is necessary that instruments be significantly correlated with the endogenous variable, English proficiency in my models. To check the quality of instruments, there are a few empirical ways: significance of instruments in the model for the endogenous variable, instruments' contribution to the explanatory power of endogenous variable model, and the Sargan overidentification test, see, for instance, Bound, Jaeger, and Baker (1995) for more on instrument checks. I specifically use the first two ways to assess the quality of the instrument, age at arrival, for the fluency in English. In all the models in Tables A.23 and A.24, the estimates on age at arrival are statistically significant at the 1% level and fairly stable for all English speaking categories, which is a positive indicator for the instrument, age at arrival. Furthermore, when I look at the χ^2 statistics, it is obvious that, in all models, excluding age at arrival from the English proficiency equation results in a significant reduction in the explanatory power of multinomial logistic regressions. Comparing Model 8 to 8a in Table A.24, I see that adding age at arrival into the English proficiency

equation increases the explanatory power of the English proficiency model by about 9.4%. The improvements resulting from the inclusion of age at arrival in the explanatory power of the English proficiency model are much higher in other regressions in Tables A.23 and A.24. Hence, the instrument age at arrival passes the quality check in terms of its correlation with English proficiency.

Now let's consider the ATE results in Tables A.25 and A.26 available in appendix A that report the estimated ATEs of English proficiency on log hourly wages among Hispanic immigrants in the USA and their standard errors in parentheses, so the dependent variable in all the second stage regressions is log hourly wages. I use three different estimation methods (i.e., CF, IV, and OLS) and several different specifications in order to compare the performance of these estimation methods to each other. CF is the control function estimation with control function terms in Procedure 1.2. IV is the instrumental variables estimation in Procedure 1.1. Since English proficiency has three levels (i.e., not well, well, and very well), I create binary English proficiency indicators for each level and label them ep_1, ep_2, and ep_3 respectively. The $ep_2 - ep_1$ and $ep_3 - ep_1$ in Tables A.25 and A.26 denote the differences between the corresponding estimates on binary English proficiency indicators and are simply the estimated ATEs due to the usage of demeaned control variables in the models. In IV regressions, I employ the predicted probabilities from the first stage regressions (e.g., those probabilities from Model 1 in Table A.23) as instruments for the binary endogenous English proficiency indicators (e.g., instruments in the IV regression under Model 1 in Table A.25). As for the control variables in the models, Model 1 has no exogenous variables controlled for. Model 2 introduces only potential experience that is totally excluded from the first stage regressions. Model 3 controls for both potential experience and education. Model 4 includes potential experience, education, and gender. Model 5 controls for potential experience, education, gender, and occupation. Model 6 contains potential experience, education, gender, occupation, and ancestry. Model 7 includes potential experience, education, gender, occupation, ancestry, and employment status. Lastly, Model 8 controls for potential experience,

education, gender, occupation, ancestry, employment status, and worker class. The standard errors in CF regressions come from the analytical formula in Th.1.4, and the standard errors in IV regressions are bootstrapped. Model 8 in Table A.26 is the full specification model and, thereof, can be thought as a earnings equation with its predictors as the determinants of earnings. As in the first stage, in all models, I use the whole sample with 38779 observations in it.

To avoid clutter in Tables A.25 and A.26 due to a large number of control variables used in models, I have shared only the estimated ATEs in these tables. However, I can make the full results available upon request. By discussing the results of regressions in Tables A.25 and A.26 that relate English proficiency to earnings of immigrant Hispanics in the USA, my goal is to be able to say something about whether deficiency in speaking English negatively influences earnings for immigrant Hispanic workers in the USA, that is, whether the average Hispanic immigrant who improves in its English speaking skills ends up with a higher wage than it would have earned had he not improved its English speaking skills. After controlling for experience and education variables in Model 3, CF estimates for ATEs start getting smaller in size and significant. Especially after adding occupation, ancestry, and employment status variables into the models one by one, CF estimates for ATEs shrink in magnitude greatly and CF estimates for $ep_3 - ep_1$ (the leap from speaking English not well to very well) become all statistically significant, which indicates the significant contribution of these control variables into the earnings models. OLS estimates are also very stable and statistically significant in all models, and shrink in magnitude as I control for more and more determinants of earnings. For example, CF estimates from Model 8 in Table A.26 reveal that the wage increase from speaking English not well to well is around 30% and 79% from speaking English not well to very well. The same wage premiums from OLS are about 12% and 22% respectively. On the other hand, IV estimates do not perform well in these earnings models of Tables A.25 and A.26. IV captures a positive effect of one treatment (the leap from speaking English not well to well) on earnings but a negative effect of the

other treatment (the leap from speaking English not well to very well) on earnings, which is a conflicting result. The estimated ATEs from IV are also all insignificant in Models 6 through 8 which I am in favor of for the sake of a more complete earnings model because Models 6 through 8 include extra control variables frequently used in the literature.

In Tables A.27 through A.32 available in appendix A, I also add the results from nonparametric bound analysis in the sense of Manski and Pepper (2000). Assuming both monotone treatment response (MTR) and monotone treatment selection (MTS), this nonparametric bound analysis provides identification regions for the ATEs. When drawing conclusions about the returns to English proficiency on log hourly wages among Hispanic immigrants in the USA, the MTR assumption means that wages among Hispanic immigrants in the USA increase as a function of English proficiency levels, ceteris paribus. On the other hand, the MTS assumption states that these Hispanic immigrants with higher levels of English proficiency have weakly higher mean wage functions than do those with lower levels of English proficiency. As shown in Manski and Pepper (2000) and Gonzalez (2005), combining these MTR and MTS assumptions produce tighter identification regions with smaller upper bounds on the returns to schooling and English proficiency. Hence, I follow their approach and construct the upper and lower bounds of the estimated ATEs of English proficiency on log hourly wages among Hispanic immigrants in the USA by using the combined MTR and MTS assumptions. The MTR+MTS bounds in Tables A.27 through A.32 are calculated based off the inequalities (21) in Manski and Pepper (2000). In these tables, I essentially report the estimated ATEs of English proficiency on log hourly wages among Hispanic immigrants in the USA, their standard errors in parentheses (for CF, IV, and OLS estimates only), the estimated nonparametric bounds for the ATEs, and their 95% confidence intervals in brackets. As in Tables A.25 and A.26, the dependent variable is log hourly wages. I use four different estimation methods (i.e., CF, IV, OLS, and nonparametric bounds) and the full specification model (i.e., Model 8 in Table A.26) in order to compare the performance of these estimation methods to each other. CF and IV estimations are as described in Pro-

cedure 1.2 and Procedure 1.1, respectively. As in Table A.26, the $ep_2 - ep_1$ and $ep_3 - ep_1$ in Tables A.27 and A.32 are simply the estimated ATEs due to the usage of demeaned control variables in the models. Again, in IV regressions, I employ the predicted probabilities from the first stage regressions (e.g., those probabilities from Model 8 in Table A.24) as instruments for the binary endogenous English proficiency indicators (e.g., instruments in the IV regression under Model 8 in Table A.26). As for the control variables in CF, IV, and OLS, I control for potential experience, education, gender, occupation, ancestry, employment status, and worker class. The standard errors in CF regressions come from the analytical formula in Th.1.4, and the standard errors in IV regressions are bootstrapped. The 95% confidence intervals are also bootstrapped. In Table A.27, I use the whole sample with 38779 observations in it. In Tables A.28, A.29, A.30, A.31, and A.32, I pay attention to the subsamples of males, females, operators, repair workers, and service employees with 25568, 13211, 9622, 6209, and 5417 observations in them, respectively.

For the sake of tidiness in Tables A.27 through A.32, I have chosen to share only the estimated ATEs in these tables. However, I can make the full results available upon request. In Table A.27, CF estimate for $ep_2 - ep_1$ (.30 but statistically insignificant) is within the estimated nonparametric bounds but CF estimate for $ep_3 - ep_1$ (.79 and statistically significant) is well over the bounds. OLS estimates are all statistically significant and well within the estimated nonparametric bounds. However, the lower bound estimates are all zero and, thereof, not informative in the sense that they do not narrow the expected lower bounds, which are positive with respect to human capital theory. Specifically, the estimated bounds for $ep_2 - ep_1$ are 0 and .32, and those for $ep_3 - ep_1$ 0 and .43. These bounds in Table A.27, nevertheless, suggest that the wage premiums from speaking English better is positive. IV estimates do not perform well: They are all statistically insignificant in Table A.27.

In Tables A.28 through A.32, I investigate how the penalties imposed by LEP (or the gains resulting from improvements in English proficiency) may vary across gender and occupation. In Tables A.28 and A.29, the bounds on ATEs are similar for both men and women are

relatively similar with a slightly higher upper bound for $ep_3 - ep_1$ among men. However, this is not the case for CF and OLS estimates: They reveal that the wage premiums due to improvements in English proficiency are considerably lower among women. For example, CF estimates in Tables A.28 and A.29 show that the wage increase from speaking English not well to well (not well to very well) among men is over 50% (84%) more than that among women. The same applies to OLS findings, as well. Even though the upper bounds on $ep_3 - ep_1$ are smaller than the point estimates of CF, the empirical results suggest that CF estimates (and OLS estimates) detect the wage inequality between men and women.

In Tables A.30 through A.32, I looked at different occupations (i.e., operators, repair workers, and service employees) and explored if heterogeneity in their use of language may lead to differences in wage premiums due to speaking English better among Hispanic immigrants in the USA. All the results from CF, OLS, and the nonparametric bound analysis imply that the wage gains due to improvements in English proficiency varies greatly across occupations. For instance, in managerial and repair occupations, the highest wage increases from speaking English not well to very well (137% and 114% by CF estimation, respectively) are observed. Whereas, the lowest wage premium values from speaking English not well to very well are attained in service occupation (34% by nonparametric upper bounds). The IV estimates are all statistically insignificant.

Overall, only CF, OLS, and nonparametric bound analysis produce statistically significant results in line with the literature. CF estimates for ATEs are well above the OLS estimates and generally outside the nonparametric bounds, which indicates that OLS estimates might be biased downwards and that CF method overestimates, especially when the size of treatment groups are imbalanced. However, as noted before, earlier studies provide substantial evidence for this overestimation: Estimation methods that explore the effect of English proficiency on earnings and control for endogeneity and measurement error produce estimates greater in magnitude than does OLS. In addition, nonparametric bound analysis has its own disadvantage: Its lower bound estimates for ATEs are always zero. Therefore,

CF method I propose offers an attractive alternative estimation method to ATE literature and outperforms IV in this empirical application.

1.9 Conclusion

In this chapter, I introduce an econometric model with a discrete multivalued endogenous treatment variable and show how to consistently estimate ATEs by a three step estimation procedure of CF method in such a model. In addition, I show the asymptotic distribution of the CF estimates follows a normal distribution, and the CF estimates are $\sqrt{N} - consistent$. I propose a consistent estimator for the asymptotic variance matrix of CF estimates, which takes into consideration the nonlinear first stage estimation. Using GMM, I also indicate how one can consistently estimate ATEs and obtain valid standard errors for the parameters of interest. I offer a hypothesis testing framework with hypotheses expressed as a set of restrictions on model parameters and construct a Lagrange multiplier statistic which is asymptotically χ^2.

I also demonstrate how CF method can be applied to one special case: the model with fixed correlation between counterfactual error terms and latent model errors. As expected, the asymptotic distribution of the CF estimates still follows a normal distribution, and the CF estimates are still $\sqrt{N} - consistent$ in this special case. A consistent estimator for the asymptotic variance matrix of CF estimates in this case still follows the conventional sandwich form, and the GMM solution follows the same structure proposed for the more general case.

In my simulation analysis, I compare CF method with IV method. The simulation results suggest that, under no misspecification, CF method is asymptotically unbiased and consistent and can be more efficient than IV method. Whereas, IV method is generally asymptotically biased and inconsistent. Therefore, the simulation results also indicate that, without misspecification, CF method consistently estimates ATEs while IV method often

cannot. On the other hand, when the correlation between counterfactual error terms and latent model errors is constant, i.e. when counterfactual errors are homogeneous, the simulation results indicate that, under no misspecification, both CF method and IV method are asymptotically unbiased and consistent with CF method being slightly more efficient. After introducing some misspecification, the simulation results show that IV method outperforms CF method in terms of efficiency and that both CF and IV methods have biased estimates. However, biases in IV estimates are generally lower than those in the estimates of CF method.

In my empirical application, I illustrate the role of limited English proficiency (LEP) in determining wages of Hispanic workers in the USA. Utilizing age at arrival as an instrumental variable, both OLS, CF, and nonparametric bound analysis indicate that LEP on average imposes a statistically significant wage penalty on immigrant Hispanic workers in the USA. In line with the existing literature, CF estimates are greater in magnitude than the OLS estimates, and nonparametric bound analysis provides uninformative lower bounds. IV estimates mostly produce insignificant results or results that are against expectations.

In future, it can be worth further researching CF method and its large sample properties in a discrete multivalued endogenous treatment model with a nonlinear second stage (outcome) equation. We can theoretically examine how the nonlinearity embedded in the first stage (choice) equation can improve identification in a discrete multivalued endogenous treatment model with weak instruments. Furthermore, under the current model, we can systematically compare CF method to IV method in terms of their asymptotic efficiency when counterfactual errors are homogeneous and can explore how to estimate the local average treatment effects and quantile effects. Lastly, we can theoretically study the large sample properties of CF and IV method in a high dimensional (i.e., settings where sample size is less than the number of parameters to be estimated) discrete multivalued endogenous treatment model.

CHAPTER 2

ESTIMATION AND INFERENCE FOR MULTIVALUED ENDOGENOUS TREATMENT EFFECT MODELS WITH CORRELATED RANDOM COEFFICIENTS

2.1 Introduction

When the parameter of interest in an economic model changes in a population, economists turn to random coefficient (RC) models. For example, the effect of tutoring on exam performance may be quite different across students: some students with decent exam preparation under their belt can greatly benefit from tutoring, and for some students who are totally lost in class tutoring can really be a waste of time and energy. The regression models capturing this idea of RCs date back as early as 1950s. In the econometrics textbook of Klein (1953, p. 216), he points out the lack of complexity in linear regression equations (especially those with a limited number of covariates) using cross sectional data to decipher the differences among people in their responses to outcomes. And then he suggests the usage of RC models to take these differences truly into account. However, RC models came to the mainstream economics with the work of Zellner (1969) on aggregation problem in models with random coefficients, see Swamy and Tavlas (2001) for a broad summary of random coefficient models. Swamy (1970) offers a consistent and an asymptotically efficient estimator for the mean of RCs in a panel data setting and applies its theoretical findings to the analysis of annual gross investment of firms. Using again panel data, Swamy and Mehta (1977) estimate the demand model for liquid asset in which the effect of time deposits, demand deposits and savings, and loan association share varies by both time and state in the USA (in addition to showing the asymptotic properties of estimators for the mean of RCs and their variance-covariance matrix.)

In nonlinear settings, Bjorklund and Moffitt (1987) generate a RC self-selection model for the effect of some activities such as education, training, and unions on wages and apply their

model to the government manpower training program in Sweden to show the heterogeneity in wage gains to the program. Akin, Guilkey, and Sickles (1979) extend the ordered response probit model to the RC probit model where the coefficients are allowed to be random in the latent variable equation and examine family moving decisions using the Panel Study of Income Dynamics survey data. As explained by Heckman and Robb (1985b, p. 173), RC models also have links to switching regression models with which union/nonunion wage gaps are estimated as surveyed in Lewis (1983) or college/high school wage differentials are explored (among other things) as in Willis and Rosen (1979). RC models have been intensely used in the human capital researches, giving informative insight into the understanding of, for example, the relationship between economic earnings and schooling and how this relationship varies across individuals. Becker and Chiswick (1966), Chiswick and Mincer (1972), and Chiswick (1974, ch. 3), for instance, are influential researches on earnings function relating personal earnings to schooling and other employment variables. Traditionally, it is generally assumed in these researches that economic return changes across people but is independent (or uncorrelated) of the level of schooling. Further examples over this come from Becker (1967) and Mincer (1974, chs. 2 and 3), and they use models that are in line with this assumption.

Correlated random coefficient (CRC) models come into play at occasions where researchers would like to allow for at least some correlation between a RC of interest and the variable of interest through some unobservables. As Wooldridge (2015, p. 430) mentions, CRC models can also allow for both heterogeneous treatment effects and self-selection into treatment. Heckman and Vytlacil (1998) bring in the usage of CRC among economists and specifically mention that CRC models compared to RC models can be more plausible with empirical observations and economic theory by relaxing the assumption of no correlation between the variable of interest and its rate of return, see Rosen (1977, p. 14 and 17) for a summary of the relationship among schooling, ability, and earnings and the problems associated with extracting the marginal effect of schooling on earnings. They develop the

classic RC model as in Becker and Chiswick (1966) and turn that into a CRC model in which some unmeasured ability/motivation factors (and observed characteristics) can influence the return to schooling and can also be correlated with the level of schooling, creating the correlation between a RC of interest and the variable of interest. Heckman and Vytlacil (1998) use a two-step estimator of the average return to schooling in their wage equation whereas Card (2001) (he summarizes different methods, inclusive of RC models, of causal modeling of the return to education) and Meghir and Palme (2001) prefer using instrumental variables (IV) method.

Contrary to the popularity of estimating ATE by IV method in continuous treatment effects (see, for example, Moffitt (1999) and Wooldridge (2003) for arguments on the grounds of robustness), Wooldridge (2008) explains a case where IV estimation can produce inconsistent estimates in the framework of CRC model with an endogenous binary treatment. This is due to that, in a CRC model, the binary treatment variable and unobserved heterogeneity factors in random coefficients are allowed to be correlated, and the binary treatment variable interacts with the unobserved heterogeneity factors. In a CRC model with multiple endogenous treatments, Wooldridge (2003, p. 191) points out that consistency condition of conventional IV does not hold when treatment variables are discrete. In the presence of heterogeneity and endogenous treatment, Heckman and Li (2004) also mention that traditional IV method fails to provide true average treatment effect (ATE) of schooling on earnings and employ a semi-parametric method for estimating treatment effects. Hence, in a CRC model with discrete endogenous multivalued treatments, conventional IV method is generally expected to be inconsistent for ATEs because of the existence of CRCs (if there are also heterogeneous counterfactual errors in the structural equation as in Chapter 1, then inconsistency of IV can get even worse). Control function (CF) method naturally comes as an alternative estimation method to IV. For instance, after pinpointing the drawbacks of ordinary least squares (OLS) and IV methods, Gebel and Pfeiffer (2007) estimate average returns to education in the West German labor market using CF method in a CRC setting.

In continuous treatment case, Amann and Klein (2012) use CF method to estimate the ATE of tenure on hourly wages in Germany within a CRC framework, allowing heterogeneous returns to tenure across individuals and feedback between these returns and tenure decision. Unfortunately, in discrete treatment cases, CF method has received little to no attention under the framework of CRC models.

In this chapter, I extend my work from Chapter 1 to CRC framework. I focus on estimating ATEs in a discrete multivalued endogenous treatment model with CRCs and heterogeneous counterfactual errors and investigate the behavior of both CF and IV methods comparatively in this setting. This has not been studied to the best of my knowledge and is my main contribution to the literature. Specifically, in this chapter, I suggest a consistent CF estimator for the ATEs and show the asymptotic properties of CF parameter estimates in a discrete multivalued endogenous treatment model with CRCs and heterogeneous counterfactual errors. Using a simulation analysis, I also claim that, without misspecification, IV method is generally asymptotically biased and inconsistent to a great degree whereas CF method is not. However, when misspecification is introduced, my simulation findings suggest that IV method perform better than CF method when it comes to unbiasedness. As for efficiency (with or without misspecification), the findings from simulations show that neither IV method nor CF method is necessarily more efficient than the other.

The rest of this chapter is organized as follows. In section 2.2, I introduce the model. In section 2.3, I derive the estimating equations for both CF and IV methods and propose procedures to estimate the parameters of interest and ATEs for both methods. In section 2.4, I show the asymptotic properties of CF estimates, propose a consistent estimator for the asymptotic variance matrix of CF estimates, and show how a GMM framework can be set up for the main problem. In section 2.5, I share some simulation results. In section 2.6, I conclude. And, in appendix B, I share the derivations and simulation tables that are hidden from the main body of this chapter.

2.2 The Model

Consider the following model with CRCs

$$\begin{aligned} y_g &= m_g + \mathbf{x}\mathbf{b}_g + u_g \\ w_g^* &= \mathbf{z}\gamma_{\mathrm{g}} + a_g, \end{aligned} \tag{2.1}$$

where y_g is the g^{th} counterfactual outcome variable, m_g is the scalar random coefficient in the counterfactual outcome equation for y_g, $\mathbf{x} \equiv (x_1, x_2, \ldots, x_l)$ is the $1 \times l$ vector of exogenous variables in y_g, $\mathbf{b}_g$ is the $l \times 1$ vector of slope random coefficients in y_g, u_g is the counterfactual error in y_g, w_g^* is the latent treatment variable that determines the choice of treatment status among $G+1$ alternative treatment statuses, $\mathbf{z} \equiv (z_1, z_2, \ldots, z_k)$ is the $1 \times k$ vector of instruments that includes a constant term in the choice equation for w_g^*, γ_g is the $k \times 1$ vector of parameters in w_g^*, and a_g is the scalar error term that is independently and identically Gumbel distributed (*i.i.d.*) with location parameter $\mu = 0$ and scale parameter $\beta = 1$ in w_g^* for $g = 0, 1, \ldots, G$. Note that (2.1) is almost the same as the counterfactual outcome equation together with the choice equation from Chapter 1; however, here I incorporate CRCs into the model.

As in Chapter 1, let $w \in \{0, 1, \ldots, G\}$ be the observed discrete multivalued endogenous treatment variable whose values are determined by w_g^* for $g = 0, 1, \ldots, G$. One common interpretation of w_g^* is to think of it as the utility or satisfaction obtained from treatment status g. Let the treatment statuses of w be exhaustive and mutually exclusive. Define binary treatment status indicators, $d_g = 1[w = g]$ for $g = 0, 1, \ldots, G$. So the binary treatment status indicator d_g is equal to one if the treatment status is equal to g and zero otherwise. This coupled with the mutual exclusivity of treatment statuses implies that $\sum_{g=0}^{G} d_g = 1$. Define the $1 \times (G+1)$ vector of treatment statuses $\mathbf{d} \equiv (d_0, d_1, \ldots, d_G)$.

Let y be the observed outcome. Then, I can write

$$y = d_0 y_0 + d_1 y_1 + \cdots + d_G y_G, \tag{2.2}$$

where y_g is the g^{th} counterfactual outcome for $g = 0, 1, \ldots G$.

After having described the discrete multivalued endogenous treatment model with CRCs above, I now will make a series of assumptions that complete the model, and are used in estimation. Notice that some of these assumptions will be the same as the ones used in Chapter 1. First, I assume that the rational economic agents choose the status of treatment from which they receive the most satisfaction out of all possible treatment statuses. That is,

- **Assumption 2.1 (A.2.1)**: One chooses treatment status g, i.e., $w = g$ if and only if $w_g^* \geq w_j^* \ \forall j \neq g$ for $g, j = 0, 1, \ldots, G$.

Second, I assume that identification of the model in (2.1) and (2.2) is contributed by exclusion of some (at least one) variables in the set of instruments $\mathbf{z}$ from the set of exogenous variables in $\mathbf{x}$. This exclusion restriction is encouraged for the estimation and identification to be more convincing and reliable even though nonlinearity in estimation suffices for identification, especially when the exogenous variables in $\mathbf{z}$ vary enough in the sample. The set of exogenous variables in $\mathbf{x}$ can all be included in the set of instruments $\mathbf{z}$.

- **Assumption 2.2 (A.2.2)**: Identification of the model described by (2.1) and (2.2) is strengthened by exclusion of at least one variable in $\mathbf{z}$ from the set of variables in $\mathbf{x}$.

As in Chapter 1, the identification argument is based on both exclusion restriction(s) and the above nonlinearity that describes the conditional probability of treatment status g

as a function of the set of instruments $\mathbf{z}$. In addition, as shown by McFadden (1973), under the model in (2.1) and (2.2) the assumptions made so far allow the treatment variable w to follow a multinomial logit model with choice probabilities given as follows:

$$P(w = g|\mathbf{x}, \mathbf{z}) = P(w = g|\mathbf{z}) = exp(\mathbf{z}\gamma_{\mathrm{g}})/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_{\mathrm{r}}), \tag{2.3}$$

for $g = 0, 1, \ldots, G$. The next assumption is essential to the CF estimation, which I describe in section 2.3, since this assumption coupled with the multinomial logit specification of the treatment variable w will play a role in creating CF terms that account for the endogeneity in w.

- **Assumption 2.3 (A.2.3)**: $E(u_g|\mathbf{x}, \mathbf{z}, \mathbf{a}) = E(u_g|\mathbf{a}) = \sum_{j=0}^{G} \eta_{g,j} a_j + [-\sum_{j=0}^{G} \eta_{g,j} E(a_j)]$, where u_g is the counterfactual error in y_g, $\mathbf{x}$ is the $1 \times l$ vector of exogenous variables in y_g, $\mathbf{z}$ is the $1 \times k$ vector of instruments that includes a constant term in the choice equation for w_g^*, $\mathbf{a} \equiv (a_0, a_1, \ldots, a_G)$ is the $1 \times (G+1)$ vector of $i.i.d.$ Gumbel distributed errors a_j with location parameter $\mu = 0$ and scale parameter $\beta = 1$ in w_j^*, $\eta_{g,j}$ is the scalar multiple of correlation coefficient between u_g and a_j, and $E(a_j) = 0.5772$ is Euler's constant for $j, g = 0, 1, \ldots, G$.

Bourguignon, Fournier, and Gurgand (2007) refers to A.2.3 as Dubin and McFadden's linearity assumption since the conditional expectation of counterfactual error u_g given all Gumbel distributed errors $\mathbf{a}$ is linear in $\mathbf{a}$ for $g = 0, 1, \ldots, G$. A.2.3 also implies that, conditional on $\mathbf{a}$, $\mathbf{x}$ and $\mathbf{z}$ are redundant for the conditional expectation of u_g. In other words, u_g is mean independent of $\mathbf{x}$ and $\mathbf{z}$ conditional on $\mathbf{a}$. The next assumption puts exogeneity and linearity restrictions on the structure of RCs in (2.1).

- **Assumption 2.4 (A.2.4)**: $m_g = \psi_{og} + \mathbf{x}\psi_g$ and $\mathbf{b}_g = \kappa_{og} + \mathbf{\Gamma}_g\mathbf{x}' + \mathbf{v}_g$ are the random coefficients in y_g, where ψ_{og} is the scalar parameter in the random scalar coefficient m_g, $\mathbf{x}$ is the $1 \times l$ vector of exogenous variables in y_g, ψ_g is the $l \times 1$ vector of slope parameters in m_g, κ_{og} is the $l \times 1$ vector of parameters in the random vector of slope coefficients $\mathbf{b}_g$, $\mathbf{\Gamma}_g$ is the $l \times l$ matrix of slope parameters in $\mathbf{b}_g$, and $\mathbf{v}_g$ is the $l \times 1$ vector of error terms with $E(\mathbf{v}_g|\mathbf{x}, \mathbf{z}) = \mathbf{0}$ in $\mathbf{b}_g$ for $g = 0, 1, \ldots, G$.

A.2.4 implies that $E(m_g|\mathbf{x}, \mathbf{z}) = E(m_g|\mathbf{x})$ and $E(\mathbf{b}_g|\mathbf{x}, \mathbf{z}) = E(\mathbf{b}_g|\mathbf{x})$ for $g = 0, 1, \ldots, G$. That is, both the scalar random coefficient m_g in y_g and the random vector of slope coefficients $\mathbf{b}_g$ in y_g are mean independent of $\mathbf{z}$ conditional on $\mathbf{x}$. In addition, for convenience in estimation, the RCs m_g and $\mathbf{b}_g$ are assumed to be linear in $\mathbf{x}$. The conditional expectation $E(\mathbf{v}_g|\mathbf{x}, \mathbf{z}) = \mathbf{0}$ might look a little restrictive; however, it still allows for possible correlation between the treatment variable w and the random vector of slope coefficients $\mathbf{b}_g$, especially in higher moments. The following assumption indeed establishes the existence of that arbitrary correlation through error terms in w_g^* but restricts its form, enabling the model to incorporate CRCs.

- **Assumption 2.5 (A.2.5)**: $E(\mathbf{v}_g|\mathbf{x}, \mathbf{z}, \mathbf{a}) = E(\mathbf{v}_g|\mathbf{a}) = \mathbf{P}\mathbf{a}'$, where $\mathbf{v}_g$ is the $l \times 1$ vector of error terms with $E(\mathbf{v}_g|\mathbf{x}, \mathbf{z}) = 0$ in $\mathbf{b}_g$, $\mathbf{a}$ is the $1 \times (G + 1)$ vector of $i.i.d.$ Gumbel distributed errors a_g with location parameter $\mu = 0$ and scale parameter $\beta = 1$ in w_g^*, $\mathbf{x}$ is the $1 \times l$ vector of exogenous variables in y_g, $\mathbf{z}$ is the $1 \times k$ vector of instruments that includes a constant term in the choice equation for w_g^*, and $\mathbf{P}$ is the $l \times (G + 1)$ matrix of constant parameters for $g = 0, 1, \ldots, G$.

By A.2.5 I impose that, conditional on $\mathbf{a}$, $\mathbf{x}$ and $\mathbf{z}$ are redundant for the conditional expectation of $\mathbf{v}_g$. In other words, $\mathbf{v}_g$ is mean independent of $\mathbf{x}$ and $\mathbf{z}$ conditional on $\mathbf{a}$.

Again, for convenience in estimation, the conditional expectation of $\mathbf{v}_g$ is assumed to be linear in $\mathbf{a}$. In essence, A.2.5 is similar to A.2.3: just as A.2.3 formulates the endogeneity in w, A.2.5 expresses the correlation between w and $\mathbf{b}_g$. In this regard, the constant parameters in $\mathbf{P}$ can be interpreted as scalar multiple of correlation coefficients between $\mathbf{v}_g$ and a_j for $j, g = 0, 1, \ldots, G$. Just as A.2.3, A.2.5 is also central to the CF method in deriving the CF estimating equation in section 2.3.

Under all assumptions from A.2.1 through A.2.5, the model in (2.1) and (2.2) can be consistently estimated by CF method. In section 2.3, I will propose a consistent estimator for the ATEs in this discrete multivalued endogenous treatment model with CRCs.

2.3 Estimation

The main theme of interest is again the estimation of ATEs in the discrete multivalued endogenous treatment model with CRCs and heterogeneous counterfactual errors that is described by (2.1) and (2.2) under the assumptions from A.2.1 through A.2.5. Following the definitions from Chapter 1, denote $ATE_{g,0}$ as the expected gain from treatment g with respect to the base treatment $g = 0$ for $g = 1, \ldots, G$. In my model, under A.2.1 through A.2.4, and the law of iterated expectations, I can write

$$\begin{aligned} ATE_{g,0} &= E(y_g - y_0) \\ &= E\left(m_g + \mathbf{x}\mathbf{b}_g + u_g - (m_0 + \mathbf{x}\mathbf{b}_0 + u_0)\right) \\ &= (\psi_{og} - \psi_{o0}) + E(\mathbf{x})\left[(\psi_g + \kappa_{og}) - (\psi_0 + \kappa_{o0})\right] + E(\mathbf{x} \otimes \mathbf{x}) vec(\mathbf{\Gamma}_g - \mathbf{\Gamma}_0), \end{aligned} \tag{2.4}$$

where $vec(\cdot)$ is the column vectorization operator and $vec(ABC) = (C' \otimes A)vec(B)$ for conformable matrices A, B, and C. Note that the last equality above uses $E(\mathbf{v}_g|\mathbf{x}, \mathbf{z}) = 0$ for $g = 1, \ldots, G$.

Then, using the analogy principle of Manski (1988, ch. 1), a consistent estimator of $ATE_{g,0}$ is

$$\widehat{ATE_{g,0}} = (\hat{\psi}_{og} - \hat{\psi}_{o0}) + \overline{\mathbf{x}}[(\widehat{\psi_g + \kappa_{og}}) - (\widehat{\psi_0 + \kappa_{o0}})] + (\overline{\mathbf{x} \otimes \mathbf{x}})vec(\hat{\mathbf{\Gamma}}_g - \hat{\mathbf{\Gamma}}_0), \tag{2.5}$$

where $\hat{\psi}_{og}$, $\hat{\psi}_{o0}$, $(\widehat{\psi_g + \kappa_{og}})$, $(\widehat{\psi_0 + \kappa_{o0}})$, $\hat{\mathbf{\Gamma}}_g$, $\hat{\mathbf{\Gamma}}_0$, $\bar{\mathbf{x}}= N^{-1}\sum_{n=1}^{N}\mathbf{x}_i$, and $\overline{\mathbf{x} \otimes \mathbf{x}}= N^{-1}\sum_{n=1}^{N}\mathbf{x}_i \otimes \mathbf{x}_i$ are respectively consistent estimates for ψ_{og}, ψ_{o0}, $(\psi_g + \kappa_{og})$, $(\psi_0 + \kappa_{o0})$, $\mathbf{\Gamma}_g$, $\mathbf{\Gamma}_0$, $E(\mathbf{x})$, and $E(\mathbf{x} \otimes \mathbf{x})$ for $g = 1, \ldots, G$. Here, it is important to reemphasize that by $(\widehat{\psi_g + \kappa_{og}})$, I mean a consistent estimate for the sum $(\psi_g + \kappa_{og})$ not the sum of consistent estimates for ψ_g and κ_{og}. Note that when there is only one exogenous variable (i.e., $\mathbf{x} = x$) in y_g with $E(x) = 0$ and $E(x^2) = 1$, $ATE_{g,0}$ simplifies to

$$ATE_{g,0} = (\psi_{og} - \psi_{o0}) + (\Gamma_g - \Gamma_0). \tag{2.6}$$

Then, a consistent estimate of $ATE_{g,0}$ in (2.6) is

$$\widehat{ATE_{g,0}} = (\hat{\psi}_{og} - \hat{\psi}_{o0}) + (\hat{\Gamma}_g - \hat{\Gamma}_0), \tag{2.7}$$

where $\hat{\psi}_{og}$, $\hat{\psi}_{o0}$, $\hat{\Gamma}_g$, and $\hat{\Gamma}_0$ are defined as in (2.5) for $g = 1, 2, \ldots, G$. It is rather important to state this simplification here because I use this version of $ATE_{g,0}$ in (2.6), instead of the one in (2.5), in my simulation analysis later in this chapter.

2.3.1 IV Estimation

Consider the observed outcome

$$\begin{aligned} y &= d_0 y_0 + d_1 y_1 + \ldots + d_G y_G \\ &= \sum_{j=0}^{G} \psi_{oj} d_j + \sum_{j=0}^{G} d_j \mathbf{x}(\psi_j + \kappa_{oj}) + \sum_{j=0}^{G} d_j (\mathbf{x} \otimes \mathbf{x}) vec\mathbf{\Gamma}_g + \varepsilon, \end{aligned} \tag{2.8}$$

where $\varepsilon = \sum_{j=0}^{G} d_j \mathbf{x}\mathbf{v}_j + \sum_{j=0}^{G} d_j u_j$. Applying IV method on (2.8) requires instruments for the binary treatment indicators d_j and $\mathbf{x}$ since both $corr(d_j, \varepsilon)$ and $corr(x_k, \varepsilon)$ are expected not to be zero, where $x_k \in \mathbf{x}$ and $\mathbf{x} \equiv (x_1, x_2, \ldots, x_l)$ for $k = 1, 2, \ldots, l$ and $j = 0, 1, \ldots, G$. One might think of $\mathbf{z}$ as instruments for $\mathbf{x}$. And as to d_j, one can model the treatment variable w as a discrete multinomial logit model and then use the predicted probabilities from this model as instruments for d_j, $j = 0, 1, \ldots, G$. Hence, one can prescribe the following three-stage procedure to estimate ATEs:

Procedure 2.1

1. Estimate the predicted probabilities, $\hat{\Lambda}_{j_i} = exp(\mathbf{z}_i \hat{\gamma}_{\mathrm{j}}) / \sum_{r=0}^{G} exp(\mathbf{z}_i \hat{\gamma}_{\mathrm{r}})$, from a MNL of w_i on $\mathbf{z}_i$ for $j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$.

2. Estimate the parameters in (2.8) by IV method using instruments $(\hat{\Lambda}_{j_i}, \hat{\Lambda}_{j_i} \mathbf{z}_i, \hat{\Lambda}_{j_i} (\mathbf{z}_i \otimes \mathbf{z}_i))$ for $(d_{j_i}, d_{j_i} \mathbf{x}_i,\ d_{j_i} (\mathbf{x}_i \otimes \mathbf{x}_i))$, $j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$.

3. Plug parameter estimates from step 2 and sample averages of $\mathbf{x}$ and $\mathbf{x} \otimes \mathbf{x}$ into (2.5), and estimate ATEs.

Procedures similar to Procedure 2.1 are not uncommon in economical applications, see, for example, Puhani and Weber (2007); and Sloan, Picone, Taylor Jr., and Chou (2001). Therefore, some economists can use conventional IV method in a discrete multivalued endogenous treatment model with CRCs and heterogeneous counterfactual errors. However,

conventional IV estimation of (2.8) is likely not to produce consistent parameter and ATE estimates as pointed out by Wooldridge (2003, p. 191 and 2008, p. 98) and Card (2001, p. 1819). Conventional IV method generally fails because it is contaminated by the two different sets of interaction terms in the error term ε of (2.8): $d_j\mathbf{x}\mathbf{v}_j$ and $d_j u_j$ for $j = 0, 1, \ldots, G$. The instruments used by conventional IV method (all are nonlinear functions of $\mathbf{z}$) as in Procedure 2.1 are expected to be correlated with ε through these interaction terms that are correlated with $\mathbf{z}$ through d_j for $j = 0, 1, \ldots, G$. Also note that IV estimator described in Procedure 2.1 would be optimal IV estimator if ε were homoskedastic, following the discussion in Chapter 1.

Just as in Chapter 1, to estimate (2.8) by IV method in canned software packages, one need to reformulate it. To this end, lets drop one of the binary treatment indicator variables, say d_G, from (2.8). And then add a constant term, $\mathbf{x}$, and $\mathbf{x} \otimes \mathbf{x}$ into (2.8). Then, (2.8) can be equivalently written as

$$\begin{aligned} y &= \left(\sum_{j=0}^{G-1} \tilde{\psi}_{oj} d_j + \tilde{\psi}_{oG}\right) + \left(\sum_{j=0}^{G-1} d_j \mathbf{x}(\widetilde{\psi_j + \kappa_{oj}}) + \mathbf{x}(\widetilde{\psi_G + \kappa_{oG}})\right) + \\ &\quad + \left(\sum_{j=0}^{G-1} d_j (\mathbf{x} \otimes \mathbf{x}) vec\widetilde{\mathbf{\Gamma}}_j + (\mathbf{x} \otimes \mathbf{x}) vec\widetilde{\mathbf{\Gamma}}_G\right) + \widetilde{\varepsilon}, \end{aligned} \tag{2.9}$$

where $\psi_{oj} = \tilde{\psi}_{oj} + \tilde{\psi}_{oG}$, $\psi_j + \kappa_{oj} = (\widetilde{\psi_j + \kappa_{oj}}) + (\widetilde{\psi_G + \kappa_{oG}})$, $vec\mathbf{\Gamma}_j = vec\widetilde{\mathbf{\Gamma}}_j + vec\widetilde{\mathbf{\Gamma}}_G$, $\psi_{oG} = \tilde{\psi}_{oG}$, $\psi_G + \kappa_{oG} = (\widetilde{\psi_G + \kappa_{oG}})$, and $vec\mathbf{\Gamma}_G = vec\widetilde{\mathbf{\Gamma}}_G$ for $j = 0, 1, \ldots, G-1$. Under this reformulation, the ATEs are as follows:

$$ATE_{g,0} = (\tilde{\psi}_{og} - \tilde{\psi}_{o0}) + E(\mathbf{x})\left[(\widetilde{\psi_g + \kappa_{og}}) - (\widetilde{\psi_0 + \kappa_{o0}})\right] + E(\mathbf{x} \otimes \mathbf{x}) vec(\widetilde{\mathbf{\Gamma}}_g - \widetilde{\mathbf{\Gamma}}_0), \tag{2.10}$$

for $g = 1, 2, \ldots, G-1$ and

$$ATE_{G,0} = (-\tilde{\psi}_{o0}) + E(\mathbf{x})(-\widetilde{\psi_0 - \kappa_{o0}}) + E(\mathbf{x} \otimes \mathbf{x}) vec(-\widetilde{\mathbf{\Gamma}}_0) \tag{2.11}$$

for $g = G$.

Therefore, consistent estimates of $ATE_{g,0}$ for $g = 1, 2, \ldots, G-1$ and $ATE_{G,0}$ under this reformulation are as follows:

$$\widehat{ATE_{g,0}} = (\widehat{\tilde{\psi}}_{og} - \widehat{\tilde{\psi}}_{o0}) + \overline{\mathbf{x}} \left[\widehat{(\psi_g + \kappa_{og})} - \widehat{(\psi_0 + \kappa_{o0})} \right] + (\overline{\mathbf{x} \otimes \mathbf{x}}) vec(\widehat{\widetilde{\mathbf{\Gamma}}}_g - \widehat{\widetilde{\mathbf{\Gamma}}}_0), \tag{2.12}$$

and

$$\widehat{ATE_{G,0}} = (-\widehat{\tilde{\psi}}_{o0}) + \overline{\mathbf{x}} \widehat{(-\psi_0 - \kappa_{o0})} + (\overline{\mathbf{x} \otimes \mathbf{x}}) vec(-\widehat{\widetilde{\mathbf{\Gamma}}}_0), \tag{2.13}$$

where $\widehat{\tilde{\psi}}_{og}$, $\widehat{\tilde{\psi}}_{o0}$, $\widehat{(\psi_g + \kappa_{og})}$, $\widehat{(\psi_0 + \kappa_{o0})}$, $\widehat{\widetilde{\mathbf{\Gamma}}}_g$, and $\widehat{\widetilde{\mathbf{\Gamma}}}_0$ are respectively the consistent estimates of $\tilde{\psi}_{og}$, $\tilde{\psi}_{o0}$, $\widetilde{(\psi_g + \kappa_{og})}$, $\widetilde{(\psi_0 + \kappa_{o0})}$, $\widetilde{\mathbf{\Gamma}}_g$, and $\widetilde{\mathbf{\Gamma}}_0$ from Procedure 2.1 applied on (2.9), and $\overline{\mathbf{x}}$ and $\overline{\mathbf{x} \otimes \mathbf{x}}$ are consistent estimates of $E(\mathbf{x})$ and $E(\mathbf{x} \otimes \mathbf{x})$ as defined in (2.5).

2.3.2 CF Estimation

To get rid of the endogeneity complications conventional IV faces as mentioned in subsection 2.3.1, CF estimation needs to find a closed form expression for $E(\varepsilon|\mathbf{d}, \mathbf{x}, \mathbf{z})$ and then to add that expression as a control variable (a.k.a. control function terms) back into (2.8). Hence, compared to IV method, CF method is almost always more complex.

To prevent equation-clutter, I left the derivation of finding a closed form expression for $E(\varepsilon|\mathbf{d}, \mathbf{x}, \mathbf{z})$ (and of the estimating equation of CF method) to appendix B. Thus, for derivations, refer to appendix B.

Having said that, (B.10) in appendix B gives me the estimating equation of CF method because I can always write

$$\begin{aligned} y &= \sum_{j=0}^{G} d_j\psi_{oj} + \sum_{j=0}^{G} d_j\mathbf{x}(\psi_j + \kappa_{oj}) + \sum_{j=0}^{G} d_j(\mathbf{x}\otimes\mathbf{x})vec\mathbf{\Gamma}_j + \\ &+ \left(\sum_{j=0}^{G} - \eta_{j,j}d_j log(\Lambda_j)\right) + \sum_{j\neq 0} d_j\eta_{j,0}M_0 + \sum_{j\neq 1} d_j\eta_{j,1}M_1 + \\ &+ \cdots + \sum_{j\neq G} d_j\eta_{j,G}M_G + \sum_{k=1,h=0}^{l,G} p_{k,h}[\sum_{j=0}^{G} d_j x_k E(a_h|d_j=1,\mathbf{x},\mathbf{z})] + \xi, \end{aligned} \tag{2.14}$$

where $E(\xi|\mathbf{d},\mathbf{x},\mathbf{z}) = 0$, $\Lambda_j = exp(\boldsymbol{z}\gamma_j)/\sum_{r=0}^{G} exp(\boldsymbol{z}\gamma_r)$, $M_j = \Lambda_j log(\Lambda_j)/(1-\Lambda_j)$, $E(a_h|d_j = 1,\mathbf{x},\mathbf{z}) = \begin{cases} -log(\Lambda_h) + E(a_h) & , h = j \\ & \text{, and } E(a_h) = 0.5772 \text{ for } h,j = 0,1,\ldots,G. \\ \dfrac{\Lambda_j log(\Lambda_j)}{(1-\Lambda_j)} + E(a_h) & , h \neq j \end{cases}$

So I can prescribe the following three-stage procedure to estimate ATEs:

Procedure 2.2

1. Same as in Procedure 2.1.

2. Run the regression of y_i on d_{0_i}, d_{1_i},..., d_{G_i}, $d_{0_i}\mathbf{x}_i$, $d_{1_i}\mathbf{x}_i$...., $d_{G_i}\mathbf{x}_i$, $d_{0_i}(\mathbf{x}_i\otimes\mathbf{x}_i)$, $d_{1_i}(\mathbf{x}_i\otimes\mathbf{x}_i)$, ..., $d_{G_i}(\mathbf{x}_i\otimes\mathbf{x}_i)$, $-d_{0_i}log(\hat{\Lambda}_{0_i})$, $-d_{1_i}log(\hat{\Lambda}_{1_i})$, ..., $-d_{G_i}log(\hat{\Lambda}_{G_i})$, $d_{1_i}\hat{M}_{0_i}$, $d_{2_i}\hat{M}_{0_i}$, ..., $d_{G_i}\hat{M}_{0_i}$, $d_{0_i}\hat{M}_{1_i}$, $d_{2_i}\hat{M}_{1_i}$, $d_{3_i}\hat{M}_{1_i}$, ..., $d_{G_i}\hat{M}_{1_i}$, ..., $d_{0_i}\hat{M}_{G_i}$, $d_{1_i}\hat{M}_{G_i}$, ..., $d_{G-2_i}\hat{M}_{G_i}$, $d_{G-1_i}\hat{M}_{G_i}$, $\sum_{j=0}^{G} d_{j_i}x_{1_i}\hat{E}(a_0|d_j = 1,\mathbf{x},\mathbf{z})_i$, $\sum_{j=0}^{G} d_{j_i}x_{1_i}\hat{E}(a_1|d_j = 1,\mathbf{x},\mathbf{z})_i$, ..., $\sum_{j=0}^{G} d_{j_i}x_{1_i}\hat{E}(a_G|d_j = 1,\mathbf{x},\mathbf{z})_i$, $\sum_{j=0}^{G} d_{j_i}x_{2_i}\ \hat{E}(a_0|d_j = 1,\mathbf{x},\mathbf{z})_i$, $\sum_{j=0}^{G} d_{j_i}x_{2_i}\hat{E}(a_1|d_j = 1,\mathbf{x},\mathbf{z})_i$, ..., $\sum_{j=0}^{G} d_{j_i}x_{2_i}\hat{E}(a_G|d_j = 1,\mathbf{x},\mathbf{z})_i$, ..., $\sum_{j=0}^{G} d_{j_i}x_{l_i}\hat{E}(a_0|d_j = 1,\mathbf{x},\mathbf{z})_i$,

$\sum_{j=0}^{G} d_{j_i} x_{l_i} \hat{E}(a_1|d_j = 1, \mathbf{x}, \mathbf{z})_i$, ..., and $\sum_{j=0}^{G} d_{j_i} x_{l_i} \hat{E}(a_G|d_j = 1, \mathbf{x}, \mathbf{z})_i$.

3. Same as in Procedure 2.1,

where $\hat{\Lambda}_{g_i} = exp(\boldsymbol{z}_i\hat{\gamma}_g)/\sum_{r=0}^{G} exp(\boldsymbol{z}_i\hat{\gamma}_r)$, $\hat{E}(a_g|d_j = 1, \mathbf{x}, \mathbf{z})_i = \begin{cases} -log(\hat{\Lambda}_{g_i}) + E(a_g) & , g = j \\ \\ \dfrac{\hat{\Lambda}_{j_i} log(\hat{\Lambda}_{j_i})}{(1 - \hat{\Lambda}_{j_i})} + E(a_g) & , g \neq j \end{cases}$,

$\hat{M}_{g_i} = \hat{\Lambda}_{g_i} log(\hat{\Lambda}_{g_i})/(1 - \hat{\Lambda}_{g_i})$, and $E(a_g) = 0.5772$ for $g, j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$.

Notice that, unlike IV method, CF method is robust to two different sources of unobserved heterogeneity in the model described by (2.1), (2.2) and the assumptions (from A.2.1 through A.2.5): one coming from the counterfactual errors u_g and the other from $\mathbf{v}_g$. Even though the treatment variable w is endogenous and correlated with the random vector of slope coefficients $\mathbf{b}_g$ for $g = 0, 1, \ldots, G$, under A.2.1 through A.2.5, CF method yields consistent estimates. Owing to this advantage of CF method over IV method by using the very same instruments $\mathbf{z}$ in CRC (and many other) models, some economists might consider CF method a generalized form of IV method, see, for example, Card (2001, p. 1819) on this.

2.4 Asymptotic Normality Results

The asymptotic theory behind CF method is not much different from the one developed in Chapter 1 because the estimating equation of CF method in (2.14) is still a two step M-estimator with some additional generated regressors. As a result, this two step M-estimator again solves the problem

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} (y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{x}_i, \mathbf{z}_i, \hat{\gamma}), \theta))^2/2, \tag{2.15}$$

where $\hat{\gamma} = (\hat{\gamma}_0', \hat{\gamma}_1', \ldots, \hat{\gamma}_G')'$ is the $(G+1)k \times 1$ vector of $\sqrt{N}-consistent$ and asymptotically normal first stage conditional MLE (CMLE) estimates from the MNL of w_i on $\mathbf{z}_i$ for $i = 1, 2, \ldots, N$. Technically, the first stage estimates does not have to be consistent as long as they converge in *plim*, i.e., $\hat{\gamma} \xrightarrow{p} \gamma^*$ where $\gamma^* \in \Gamma \subset \mathbb{R}^{\mathbf{(G+1)k}}$. However, they are consistent in this setting. CMLE solves the problem

$$\max_{\gamma \in \Gamma} \sum_{i=1}^{N} l_i(\gamma), \tag{2.16}$$

where $\gamma = (\gamma_0', \gamma_1', \ldots, \gamma_G')'$ is the $(G+1)k \times 1$ vector of parameters, and $l_i(\gamma) \equiv log(f(w_i|\mathbf{z}_i;\gamma))$, (i.e., the conditional log likelihood for observation i) is given below

$$log(f(w_i|\mathbf{z}_i,\gamma)) = \sum_{j=0}^{G} 1[w_i = j] log\left(exp(\mathbf{z}_i\gamma_j) / \sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r) \right). \tag{2.17}$$

This CMLE is exactly the same as the one used in Chapter 1. Therefore, to establish that these first stage MLE estimates are $\sqrt{N}-consistent$, I will rely on Theorem 1.1 (Th.1.1) from Chapter 1 which is restated below for readers' convenience and establishes the consistency of CMLE without compactness.

- **Theorem 2.1 (Th.2.1)**: Let $\{(w_i, \mathbf{z}_i) : i = 1, 2, \ldots\}$ be a random sample with $\mathbf{z}_i \in \mathscr{Z} \subset \mathbb{R}^{\mathbf{k}}$, $w_i \in \mathscr{W} \subset \mathbb{R}$. Let $\Gamma \subset \mathbb{R}^{\mathbf{(G+1)k}}$ be the parameter set, and denote the parametric model for the conditional density, $p(\cdot\,|\mathbf{z})$, as $\{f(\cdot\,|\mathbf{z};\gamma) : \mathbf{z} \in \mathscr{Z}, \gamma \in \Gamma\}$. Let $l : \mathscr{W} \times \mathscr{Z} \times \Gamma \to \mathbb{R}$ be a real-valued function. Assume that (a) $f(\cdot\,|\mathbf{z};\gamma)$ is a true density function with respect to the measure $\mu(dw)$ for all $\mathbf{z}$ and γ, so that $\int_{\mathscr{W}} f(w|\mathbf{z})\mu(dw) = 1, \forall \mathbf{z} \in \mathscr{Z}$ holds; (b) for some $\gamma_o \in \Gamma$, $p_o(\cdot\,|\mathbf{z}) = f(\cdot\,|\mathbf{z};\gamma_o)$, $\forall \mathbf{z} \in \mathscr{Z}$, and the true parameter vector γ_o is the unique solution to $\max_{\gamma \in \Gamma} E[l_i(\gamma)]$; (c) γ_o is an element of the interior of a convex parameter space Γ; (d) for each $\gamma \in \Gamma$, $l(\cdot\,,\gamma)$ is a Borel measurable function on $\mathscr{W} \times \mathscr{Z}$; (e) for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, $l(w, \mathbf{z}, \cdot)$ is concave in γ; and (f) $|l(w, \mathbf{z}, \gamma)| \leq b(w, \mathbf{z})$, $\forall \gamma \in \Gamma$, where $b(\cdot, \cdot)$ is a nonnegative function on

$\mathscr{W} \times \mathscr{Z}$ such that $E[b(w, \mathbf{z})] < \infty$. Then there exist a solution to problem in (2.16), the CMLE $\hat{\gamma}$, and $\hat{\gamma} \xrightarrow{p} \gamma_o$.

For the verification of the conditions stated in Th.2.1, see appendix A. I also refer readers to see Theorem 2.7 in Newey and McFadden (1994, p. 2133) for a generic consistency proof of extremum estimators without compactness. To establish that these first stage MLE estimates are asymptotically normal, I will use Theorem 1.2 (Th.1.2) from Chapter 1 which is restated below for readers' convenience.

- **Theorem 2.2 (Th.2.2)**: Let the definitions and conditions of Th.2.1 hold, and define $\mathbf{B}_o^{\mathbf{F}} \equiv Var[\nabla_\gamma' l_i(\gamma_o)]$. Furthermore, assume that (a) γ_o is an element of the interior of a parameter space Γ;—i.e., $\gamma_o \in int(\Gamma)$; (b) for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, $l(w, \mathbf{z}, \cdot)$ is twice continuously differentiable on $int(\Gamma)$; (c) $E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)] = \mathbf{0}$ and $-E[\mathbf{H}_i^{\mathbf{F}}(\gamma_o)] = Var[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)]$, where $\mathbf{s}_i^{\mathbf{F}}(\gamma) \equiv \nabla_\gamma' l_i(\gamma)$ and $\mathbf{H}_i^{\mathbf{F}}(\gamma) \equiv \nabla_\gamma[\nabla_\gamma' l_i(\gamma)]$; (d) the elements of $\nabla_\gamma[\nabla_\gamma' l(w, \mathbf{z}, \gamma)]$ are bounded in absolute value by a function $b(w, \mathbf{z})$, $\forall \gamma \in \Gamma$, where $b(\cdot, \cdot)$ is a nonnegative function on $\mathscr{W} \times \mathscr{Z}$ such that $E[b(w, \mathbf{z})] < \infty$; and (e) $\mathbf{A}_o^{\mathbf{F}} \equiv -E(\nabla_\gamma[\nabla_\gamma' l_i(\gamma_o)])$ is positive definite. Then

$$\sqrt{N}(\hat{\gamma} - \gamma_o) \xrightarrow{d} Normal(\mathbf{0}, (\mathbf{A}_o^{\mathbf{F}})^{-1}\mathbf{B}_o^{\mathbf{F}}(\mathbf{A}_o^{\mathbf{F}})^{-1}). \tag{2.18}$$

Explicitly, the score of the log likelihood for observation i is as follows:

$$\mathbf{s}_i^{\mathbf{F}}(\gamma) \equiv \nabla_\gamma' l_i(\gamma) = \left(\frac{\partial l_i}{\partial \gamma_0}(\gamma), \frac{\partial l_i}{\partial \gamma_1}(\gamma), \ldots, \frac{\partial l_i}{\partial \gamma_{\mathrm{G}}}(\gamma)\right)', \tag{2.19}$$

which is a $(G+1)k \times 1$ vector of partial derivatives of $l_i(\gamma)$ with respect to parameters in γ. The Hessian, $\mathbf{H}_i^{\mathbf{F}}(\gamma) \equiv \nabla_\gamma[\nabla_\gamma' l_i(\gamma)]$, for observation i is the $(G+1)k \times (G+1)k$ matrix of second partial derivatives of $l_i(\gamma)$ with respect to parameters in γ. Thus, using the

definitions in Th.2.2, $\mathbf{A}_o^{\mathbf{F}} \equiv -E[\mathbf{H}_i^{\mathbf{F}}(\gamma_o)]$, and $\mathbf{B}_o^{\mathbf{F}} \equiv Var[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)]$. In addition, $E[\mathbf{s}_i^{\mathbf{F}}(\gamma)] = \mathbf{0}$ and $\mathbf{A}_o^{\mathbf{F}} = \mathbf{B}_o^{\mathbf{F}}$, which are used to simplify the variance expression in (2.18), and see appendix A for how to show these equalities. Using that $E[\mathbf{s}_i^{\mathbf{F}}(\gamma)] = \mathbf{0}$ and $\mathbf{A}_o^{\mathbf{F}} = \mathbf{B}_o^{\mathbf{F}}$, I can rewrite (2.18) as below:

$$\sqrt{N}(\hat{\gamma} - \gamma_o) \quad \xrightarrow{d} \quad Normal(\mathbf{0}, (\mathbf{A}_o^{\mathbf{F}})^{-1}). \tag{2.20}$$

For the verification of the conditions in Th.2.2, see appendix A. Theorem 3.1 in Newey and McFadden (1994, p. 2143) is also helpful for a generic proof of asymptotic normality of extremum estimators. The second stage of CF method is technically OLS with generated regressors as in Chapter 1. For this reason, to establish that second stage estimates are $\sqrt{N} - consistent$, I will use a modified version of Theorem 1.3 (Th.1.3) from Chapter 1 which establishes the consistency of CF method with a compact parameter space.

- **Theorem 2.3 (Th.2.3)**: Let $\mathtt{w} = (y, \mathbb{X}, \mathbf{v})$ be a random vector with $\mathtt{w} \in \mathtt{W} \subset \mathbb{R}^{\mathbf{M+1}}$ and $M = (l(l+1)/2 + 2l + G + 2)(G+1)$. Let $\Theta \subset \mathbb{R}^{\mathbf{M}}$ and $\Gamma \subset \mathbb{R}^{(\mathbf{G+1})\mathbf{k}}$ be the parameter sets. Let $q(\mathtt{w}, \theta, \gamma) : \mathtt{W} \times \Theta \times \Gamma \to \mathbb{R}$ be a real-valued function. Let $\hat{\gamma}$ be an estimator from a preliminary estimation. Assume that (a) $\hat{\gamma} \xrightarrow{p} \gamma^*$ for some $\gamma^* \in \Gamma$; (b) for a given $\gamma^* \in \Gamma$, the true parameter vector θ_o is the unique solution to $\underset{\theta \in \Theta}{min} E[q_i(\theta; \gamma^*)]$; (c) the parameter space $\Theta \times \Gamma$ is compact; (d) for each $(\theta, \gamma) \in \Theta \times \Gamma$, $q(\cdot\,, \theta, \gamma)$ is a Borel measurable function on $\mathtt{W}$; (e) for each $\mathtt{w} \in \mathtt{W}$, $q(\mathtt{w}, \cdot, \cdot)$ is a continuous function on $\Theta \times \Gamma$; and (f)$E[|q(\mathtt{w}_i, \theta; \gamma)|] < \infty \; \forall (\theta, \gamma) \in \Theta \times \Gamma$. Then there exists a solution to problem in (2.15), the CF estimator $\hat{\theta}$, and $\hat{\theta} \xrightarrow{p} \theta_o$.

For the verification of the conditions stated in Th.2.3, one can follow the steps taken in appendix A. In addition, readers can benefit from Wooldridge (1994, p. 2730) for a generic consistency proof of two-step M-estimators with compactness. Before I move into

the asymptotic normality result, I have to introduce some notation. From problem (2.15), we can see that $q(\mathtt{w}, \theta, \gamma)$ in Th.2.3 is as follows:

$$q_i(\theta, \gamma) \equiv q(\mathtt{w}_i, \theta, \gamma) \equiv (y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{x}_i, \mathbf{z}_i, \gamma), \theta))^2/2, \tag{2.21}$$

where $m_i(\mathbf{v}_i(\gamma), \theta) \equiv m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{x}_i, \mathbf{z}_i, \gamma), \theta) \equiv \mathbb{X}_i\delta + \mathbf{v}_i\lambda$ is a real-valued scalar function, $\theta = (\delta', \lambda')'$ is the $M \times 1$ vector of parameters, $\mathbb{X}_i$ is the $1 \times (l(l+1)/2 + l + 1)(G+1)$ vector of regressors in (2.15), and $\mathbf{v}_i$ is the $1 \times (l + G + 1)(G+1)$ vector of generated regressors in (2.15). More explicitly,

$$\begin{aligned}
\mathbb{X}_i &= (\ d_{0_i},\ \cdots,\ d_{G_i},\ d_{0_i}\mathbf{x}_i,\ \cdots,\ d_{G_i}\mathbf{x}_i,\ d_{0_i}(\mathbf{x}_i \otimes \mathbf{x}_i)\ \cdots,\ d_{G_i}(\mathbf{x}_i \otimes \mathbf{x}_i)\ ) \\
\mathbf{v}_i &= (\ -d_{0_i}log(\Lambda_{0_i}),\ \cdots,\ -d_{G_i}log(\Lambda_{G_i}),\ d_{1_i}M_{0_i},\ d_{2_i}M_{0_i},\ \cdots,\ d_{G_i}M_{0_i}, \\
&\quad d_{0_i}M_{1_i},\ d_{2_i}M_{1_i},\ d_{3_i}M_{1_i},\ \cdots,\ d_{G_i}M_{1_i},\ \cdots,\ d_{0_i}M_{G_i},\ d_{1_i}M_{G_i},\ \cdots, \\
&\quad , d_{G-1_i}M_{G_i},\ \sum_{j=0}^{G} d_{j_i}x_{1_i}E(a_0|d_j = 1, \mathbf{x}, \mathbf{z})_i,\ \cdots,\ \sum_{j=0}^{G} d_{j_i}x_{1_i}E(a_G|d_j = 1, \mathbf{x}, \mathbf{z})_i, \\
&\quad \sum_{j=0}^{G} d_{j_i}x_{2_i}E(a_0|d_j = 1, \mathbf{x}, \mathbf{z})_i,\ \cdots,\ \sum_{j=0}^{G} d_{j_i}x_{2_i}E(a_G|d_j = 1, \mathbf{x}, \mathbf{z})_i,\ \cdots, \\
&\quad \sum_{j=0}^{G} d_{j_i}x_{l_i}E(a_0|d_j = 1, \mathbf{x}, \mathbf{z})_i,\ \cdots,\ \sum_{j=0}^{G} d_{j_i}x_{l_i}E(a_G|d_j = 1, \mathbf{x}, \mathbf{z})_i\ ),
\end{aligned} \tag{2.22}$$

where $\Lambda_{g_i} = exp(\mathbf{z}_i\gamma_g)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)$, $E(a_g|d_j = 1, \mathbf{x}, \mathbf{z})_i = \begin{cases} -log(\Lambda_{g_i}) + E(a_g) & , g = j \\ \dfrac{\Lambda_{j_i}log(\Lambda_{j_i})}{(1 - \Lambda_{j_i})} + E(a_g) & , g \neq j \end{cases}$,

$M_{g_i} = \Lambda_{g_i}\ log(\Lambda_{g_i})/(1 - \Lambda_{g_i})$, and $E(a_g) = 0.5772$ for $g, j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$. As one can expect, expressions such as $\hat{\Lambda}_{j_i}$, $\hat{M}_{g_i}$, and $\hat{E}(a_g|d_j = 1, \mathbf{x}, \mathbf{z})_i$ are just consistent estimates of Λ_{g_i}, M_{g_i}, and $E(a_g|d_j = 1, \mathbf{x}, \mathbf{z})_i$ with $\hat{\gamma}_g$ replacing γ_g in Λ_{g_i}, M_{g_i}, and $E(a_g|d_j = 1, \mathbf{x}, \mathbf{z})_i$ respectively. Now, I will state the theorem that is a modified version (in the sense

that dimensions of $\mathbb{X}$ and $\mathbf{v}$ are different) of Theorem 1.4 (Th.1.4) from Chapter 1 and establishes the asymptotic normality of CF method with a compact parameter space.

- **Theorem 2.4 (Th.2.4)**: Let the definitions and conditions of Th.2.3 hold. Furthermore, assume that (a) $\theta_o \in int(\Theta)$ and $\gamma^* \in int(\Gamma)$; (b) $\sqrt{N}(\hat{\gamma} - \gamma^*)$ is bounded in probability —i.e., $\sqrt{N}(\hat{\gamma} - \gamma^*) = O_p(1)$; (c) for each $(\mathtt{w}, \gamma) \in \mathtt{W} \times \Gamma$, $q(\mathtt{w}, \cdot; \gamma)$ is a twice continuously differentiable on $int(\Theta)$; (d) for each $\theta \in \Theta$, $\mathbf{s}(\cdot, \theta; \cdot) \equiv \nabla'_\theta q(\cdot, \theta; \cdot)$ is continuously differentiable on $int(\Gamma)$; (e) for each $(\theta, \gamma) \in \Theta \times \Gamma$, $\mathbf{H}(\cdot, \theta; \gamma) \equiv \nabla_\theta \mathbf{s}(\cdot, \theta; \gamma)$ is a Borel measurable function on $\mathtt{W}$; (f) for each $\mathtt{w} \in \mathtt{W}$, $\mathbf{H}(\mathtt{w}, \cdot; \cdot)$ is continuous on $\Theta \times \Gamma$; (g) $E[\| \mathbf{H}(\mathtt{w}_i, \theta; \gamma) \|] < \infty \; \forall (\theta, \gamma) \in \Theta \times \Gamma$. (h) $\mathbf{A}_o \equiv E[\mathbf{H}(\mathtt{w}_i, \theta_o; \gamma^*)]$ is positive definite; (i) for each $(\theta, \gamma) \in \Theta \times \Gamma$, $\nabla_\gamma \mathbf{s}(\cdot, \theta; \gamma)$ is a Borel measurable function on $\mathtt{W}$; (j) for each $\mathtt{w} \in \mathtt{W}$,$\nabla_\gamma \mathbf{s}(\mathtt{w}, \cdot; \cdot)$ is continuous on $\Theta \times \Gamma$; (k) $E[\| \nabla_\gamma \mathbf{s}(\mathtt{w}_i, \theta; \gamma) \|] < \infty \; \forall (\theta, \gamma) \in \Theta \times \Gamma$; (l) $E[\mathbf{s}_i(\theta_o; \gamma^*)] = \mathbf{0}$, $E[(\mathbf{A}^{\mathbf{F}}_*)^{-1} \mathbf{s}^{\mathbf{F}}_i(\gamma^*)] = \mathbf{0}$, and $E[(\mathbf{A}^{\mathbf{F}}_*)^{-1} \mathbf{s}^{\mathbf{F}}_i(\gamma^*) \mathbf{s}'_i(\theta_o; \gamma^*)] = \mathbf{0}$. Then,

$$\sqrt{N}(\hat{\theta} - \theta_o) \quad \xrightarrow{d} \quad Normal(\mathbf{0}, (\mathbf{A}_o)^{-1} \mathbf{D}_o (\mathbf{A}_o)^{-1}), \tag{2.23}$$

where $\mathbf{D}_o = \mathbf{B}_o + \mathbf{F}_o \mathbf{T}_o + \mathbf{T}'_o \mathbf{F}'_o + \mathbf{F}_o \mathbf{R}^* \mathbf{F}'_o$, $\mathbf{s}_i(\theta_o; \gamma^*) \equiv \nabla'_\theta q(\mathtt{w}_i, \theta_o; \gamma^*)$, $\mathbf{A}_o \equiv E[\nabla_\theta \mathbf{s}_i(\theta_o; \gamma^*)] \equiv E[\mathbf{H}_i(\theta_o; \gamma^*)]$, $\mathbf{B}_o \equiv E[\mathbf{s}_i(\theta_o; \gamma^*) \mathbf{s}'_i(\theta_o; \gamma^*)]$, $\mathbf{F}_o \equiv E[\nabla_\gamma \mathbf{s}_i(\mathtt{w}, \theta_o; \gamma^*)]$, $\mathbf{T}_o \equiv E[\mathbf{r}_i(\gamma^*) \mathbf{s}'_i(\theta_o; \gamma^*)]$, $\mathbf{R}^* \equiv E[\mathbf{r}_i(\gamma^*) \mathbf{r}'_i(\gamma^*)]$, $\mathbf{r}_i(\gamma^*) = (\mathbf{A}^{\mathbf{F}}_*)^{-1} \mathbf{s}^{\mathbf{F}}_i(\gamma^*)$, and $\mathbf{A}^{\mathbf{F}}_* \equiv -E(\nabla_\gamma [\nabla'_\gamma l_i(\gamma^*)])$. For the derivation of asymptotic variance of $\sqrt{N}(\hat{\theta} - \theta_o)$, refer to the subchapter 12.4 in Wooldridge (2010) or subsections 4.3 and 4.4 in Wooldridge (1994). For the verification of the conditions stated in Th.2.4, one can again follow the steps taken in appendix A. Readers can also benefit from Wooldridge (1994, p. 2730) for a generic asymptotic normality proof of two-step M-estimators with compactness. In addition, refer to appendix A for the derivation of the closed forms of the population matrices $\mathbf{A}_o$, $\mathbf{B}_o$, $\mathbf{F}_o$, and $\mathbf{R}^*$ and for seeing that $E[\mathbf{r}_i(\gamma^*)] = \mathbf{0}$, $E[\mathbf{s}_i(\theta_o; \gamma^*)] = \mathbf{0}$, and $\mathbf{T}_o \equiv E[\mathbf{r}_i(\gamma^*) \mathbf{s}'_i(\theta_o; \gamma^*)] = \mathbf{0}$. Since $\mathbf{T}_o = \mathbf{0}$, $\mathbf{D}_o$ in the asymptotic variance of $\sqrt{N}(\hat{\theta} - \theta_o)$ in (2.23) actually simplifies to $\mathbf{B}_o + \mathbf{F}_o \mathbf{R}^* \mathbf{F}'_o$.

Let's construct the following estimators for $\mathbf{A}_o$, $\mathbf{B}_o$, $\mathbf{F}_o$, and $\mathbf{R}^*$ as in Chapter 1 as follows:

$$\hat{\mathbf{A}} = N^{-1}\sum_{i=1}^{N} -\mathbf{H}_i(\hat{\theta};\hat{\gamma}), \tag{2.24}$$

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\mathbf{s}_i(\hat{\theta};\hat{\gamma})\mathbf{s}_i'(\hat{\theta};\hat{\gamma}), \tag{2.25}$$

$$\hat{\mathbf{F}} = N^{-1}\sum_{i=1}^{N}\nabla_\gamma\mathbf{s}_i(\hat{\theta};\hat{\gamma}), \quad and \tag{2.26}$$

$$\hat{\mathbf{R}} = N^{-1}\sum_{i=1}^{N}\mathbf{r}_i(\hat{\gamma})\mathbf{r}_i'(\hat{\gamma}). \tag{2.27}$$

Define $\hat{\mathbf{D}} \equiv \hat{\mathbf{B}} + \hat{\mathbf{F}}\hat{\mathbf{R}}\hat{\mathbf{F}}'$. Then, using the analogy principle and Lemma 1 in Chapter 1, a consistent estimator for $Avar\sqrt{N}(\hat{\theta}-\theta_o)$ is $\hat{Avar}\sqrt{N}(\hat{\theta}-\theta_o) = (\hat{\mathbf{A}})^{-1}\hat{\mathbf{D}}(\hat{\mathbf{A}})^{-1}$. The asymptotic standard errors of CF estimates can be obtained from the matrix $\hat{Avar}(\hat{\theta}) = (\hat{\mathbf{A}})^{-1}\hat{\mathbf{D}}(\hat{\mathbf{A}})^{-1}/N$ as usual or be bootstrapped.

2.4.1 Method of Moments Framework

Following the results from Newey (1984), Newey and McFadden (1994, p. 2132 and 2148), or Heckman, Tobias, and Vytlacil (2003), two-step estimators can be regarded as members of generalized method of moments (GMM) estimators, and the asymptotic theory for these estimators can be derived by stacking moment conditions. GMM estimators take away the burden of deriving the asymptotic variance matrix of a two-step estimator and thus provide an alternative way for inference in CF regression as well. Since the number of moment conditions is the same as the number of parameters to be estimated in my analysis, I technically use method of moments (MoM).

As in Chapter 1, in the first stage of CF method, $\hat{\gamma}$ is the CMLE estimator solving

$$\sum_{i=1}^{N}\mathbf{s}_i^{\mathbf{F}}(\hat{\gamma}) = \mathbf{0}, \tag{2.28}$$

where $\mathbf{s}_i^{\mathbf{F}}(\gamma) = \nabla'_\gamma \sum_{j=0}^{G} 1[w_i = j] log\left(exp(\mathbf{z}_i\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)\right)$. In the second stage, $\hat{\theta}$ is a OLS estimator solving

$$\sum_{i=1}^{N} \mathbf{s}_i(\hat{\theta}; \hat{\gamma}) = \mathbf{0}, \tag{2.29}$$

where $\mathbf{s}_i(\hat{\theta}; \hat{\gamma}) = \nabla'_\theta [y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{x}_i, \mathbf{z}_i, \hat{\gamma}), \hat{\theta})]^2/2$. Newey (1984) proposes stacking the summands in these first order conditions into the unified function

$$g(\theta, \gamma) = \begin{pmatrix} \mathbf{s}^{\mathbf{F}}(\gamma) \\ \mathbf{s}(\theta; \gamma) \end{pmatrix} \tag{2.30}$$

and then applying MoM using the moment conditions $E[g(\theta, \gamma)] = \mathbf{0}$ to obtain consistent estimates for θ and γ, and valid asymptotic variance matrix of $\hat{\theta}$ and $\hat{\gamma}$. Using the GMM results in the appendix of Heckman, Tobias, and Vytlacil (2003), one can also derive the asymptotic distribution theory for the ATE estimators.

2.5 Simulations

In this section, I share some simulation results that compare and contrast the estimation methods (i.e., CF and IV methods in section 2.3) and note what is different and similar in terms of their asymptotic performances, specifically asymptotic efficiency, asymptotic unbiasedness, and consistency. I will change the simulation setup for the model in section 2.2 as I change the distribution of instrument in the latent variable equation or introduce misspecification into the model by ignoring an instrument in the latent variable equation. For the sake of computational simplicity, I adopt a scheme in which there is only one covariate in the counterfactual outcome equation (i.e., $\mathbf{x} = x$) and only one instrument in the latent variable equation (i.e., $\mathbf{z} = z$). When examining the consequences of misspecification, there are two instruments determining the latent treatment variable though. And lastly, the

treatment variable w takes on only three values, and each treatment group comprises at least about 30 percent for each simulation setting.

2.5.1 Data Generating Process

In my simulation analysis, I used four different data generating processes (DGPs): one for the model in section 2.3 with asymmetric instrument, one for the model with symmetric instrument, one for the model with asymmetric instrument and misspecification, and one for the model with symmetric instrument and misspecification. The setup for the DGP of the model in section 2.3 with asymmetric instrument is as follows:

$$w \in \{0, 1, 2\},$$

$$d_g = 1[w = g], \quad g \in \{0, 1, 2\},$$

$$a_g \sim Gumbel(0, 1), \quad g \in \{0, 1, 2\},$$

$$\gamma_0 = 1, \quad \gamma_1 = 5, \quad and, \quad \gamma_2 = 9,$$

$$l_0 = 1, \quad l_1 = 5, \quad and, \quad l_2 = 3,$$

$$\mathbf{z} = z \sim \chi^2(2) - 2,$$

$$w_g^* = l_g + \gamma_g z + a_g, \quad g \in \{0, 1, 2\},$$

$$w = g \quad iff \quad w_g^* \geq w_j^*, \quad \forall j \neq g \quad and \quad g, j \in \{0, 1, 2\},$$

$$e_g \sim N(0, 1), \quad g \in \{0, 1, 2\},$$

$$\eta_{0,0} = 0.05, \quad \eta_{0,1} = 0.10, \quad and \quad \eta_{0,2} = 0.15,$$

$$\eta_{1,0} = 4.05, \quad \eta_{1,1} = 4.10, \quad and \quad \eta_{1,2} = 4.15,$$

$$\eta_{2,0} = 8.05, \quad \eta_{2,1} = 8.10, \quad and \quad \eta_{2,2} = 8.15,$$

$$u_g = \sum_{j=0}^{2} \eta_{g,j} a_j + [-\sum_{j=0}^{2} \eta_{g,j} E(a_j)] + e_g, \quad g \in \{0, 1, 2\},$$

$$\mathbf{x} = x \sim N(0, 1),$$

$$\psi_{o0} = 1, \quad \psi_{o1} = 2, \quad and \quad \psi_{o2} = 3,$$

$$\psi_0 = 4, \quad \psi_1 = 5, \quad and \quad \psi_2 = 6,$$

$$m_g = \psi_{og} + x\psi_g, \quad g \in \{0, 1, 2\},$$

$$\kappa_{o0} = 1, \quad \kappa_{o1} = 4, \quad and \quad \kappa_{o2} = 7,$$

$$\Gamma_0 = 4, \quad \Gamma_1 = 5, \quad and \quad \Gamma_2 = 6,$$

$$p_0 = 1, \quad p_1 = 2, \quad and \quad p_2 = 3,$$

$$e_{v_g} \sim N(0, 1), \quad g \in \{0, 1, 2\},$$

$$v_g = \sum_{j=0}^{2} p_j a_j + e_{v_g}, \quad g \in \{0, 1, 2\},$$

$$b_g = \kappa_{og} + \Gamma_g x + v_g, \quad g \in \{0, 1, 2\},$$

$$y_g = m_g + x b_g + u_g, \quad g \in \{0, 1, 2\},$$

$$and \quad y = d_0 y_0 + d_1 y_1 + d_2 y_2.$$

For the model with symmetric instrument, the DGP setup is almost exactly the same as the one above. However, I make some modifications on both the location parameters and the distribution of instrument appearing in the latent variable equation as follows:

$$l_0 = 1, \quad l_1 = 5.2, \quad and, \quad l_2 = 2,$$

$$\mathbf{z} = z \sim N(0, 2).$$

For the model with asymmetric instrument and misspecification in the latent variable equation, the DGP setup is very similar to the one without misspecification. However, I introduce an additional instrument in the latent variable equation and ignore it from the MNL regression of treatment variable on instruments at the first stage, thus creating misspecification. In line with this, I make the following modifications to the DGP:

$$\mathbf{z} = (z_1, z_2)',$$

$$z_1 \sim \chi^2(2) - 2,$$

$$z_2 \sim \chi^2(2) - 2,$$

$$w_g^* = l_g + \gamma_g z_1 + \vartheta_g z_2 + a_g, \quad g \in \{0, 1, 2\},$$

$$\vartheta_0 = \gamma_1, \quad \vartheta_1 = \gamma_2, \quad and \quad \vartheta_2 = \gamma_0,$$

where z_1 and z_2 are scalar instruments in the choice equation for w_g^* (i.e., the latent variable equation), and ϑ_g is a scalar parameter associated with z_2 in w_g^* for $g = 0, 1, 2$.

Lastly, for the model with symmetric instrument and misspecification in the latent variable equation, the DGP setup is again very similar to the one without misspecification. I make the following modifications to its DGP:

$$\mathbf{z} = (z_1, z_2)',$$

$$z_1 \sim N(0, 2),$$

$$z_2 \sim N(0, 2),$$

$$l_0 = 4.5, \quad l_1 = 4, \quad and, \quad l_2 = 2,$$

$$w_g^* = l_g + \gamma_g z_1 + \vartheta_g z_2 + a_g, \quad g \in \{0, 1, 2\},$$

$$\vartheta_0 = \gamma_1, \quad \vartheta_1 = \gamma_2, \quad and \quad \vartheta_2 = \gamma_0,$$

where z_1, z_2, and ϑ_g are defined just as in the model with asymmetric instrument and misspecification in the latent variable equation.

As in Chapter 1, both γ_g and l_g play a role in determining the percentage of each treatment group in simulations for $g = 0, 1, 2$. Having γ_g's being apart from each other enough is also critical to obtain strong first stage estimates and to ward off identification problems in the first stage estimation. To increase the effect of endogeneity in the model, having $\eta_{g,j}$'s seperated from each other across treatment statuses is also another critical point for $g, j = 0, 1, 2$. Following Wooldridge (2008, p. 106; 2010, p. 947), I also vary the distribution of instrument z in order to see if its distribution can influence IV estimates for ATEs in terms of consistency.

2.5.2 Simulation Results

I present my simulation results in two parts: first, asymptotic efficiency outcomes and second, asymptotic unbiasedness and consistency outcomes. The simulation results reported in Tables B.1 through B.16 aim for comparing CF method with IV method in terms of asymptotic efficiency, asymptotic unbiasedness and consistency. The first eight tables belong

to models without misspecification, whereas the last eight tables include results coming out of models with misspecification.

In Tables B.1 through B.16, I report the Monte Carlo (M.C.) estimates for ψ_{og}, Γ_g, and $ATE_{h,0}$; bias in the M.C. estimate for ATEs; bootstrapped standard errors (BS. SEs) and Monte Carlo standard deviations (M.C. SDs) for ψ_{og}, Γ_g, and $ATE_{h,0}$; and BS. SEs and M.C. SDs for standard errors of ψ_{og}, Γ_g, and $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$. In simulations, I use different sample sizes (i.e., $n = 1000$, $n = 2000$, $n = 5000$ and $n = 10000$) for each DGP setup with the number of M.C. and BS. iterations always equal to 10000. I also used some trimming to remove outliers from my simulation analysis.

As for the notation, in Tables B.1-B.16, $\hat{\psi}_{og}$ is the parameter estimate for ψ_{og}, $\hat{\Gamma}_g$ is the parameter estimate for Γ_g, $\hat{ate}_{h,0}$ is the estimate for $ATE_{h,0}$, and $bias(\hat{ate}_{h,0})$ is the bias in the estimate for $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$. Furthermore, $se(\hat{\psi}_{og})$ is the standard error of parameter estimate for ψ_{og}, $se(\hat{\Gamma}_g)$ is the standard error of parameter estimate for Γ_g, and $se(\hat{ate}_{h,0})$ is the standard error of the estimate for $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$. Since these tables would require a considerable amount of space in the main body of the chapter, I place all simulation tables of this chapter into appendix B.

At this point, it is also important to remember the true values for ψ_{og}, Γ_g, and $ATE_{h,0}$ for $g = 0, 1, 2$ and $h = 1, 2$ since I often refer them throughout this section. As pointed out in section 2.3, since $x \sim N(0, 1)$, the true values are respectively as follows:

$$\psi_{o0} = 1, \quad \psi_{o1} = 2, \quad and, \quad \psi_{o2} = 3,$$

$$\Gamma_0 = 4, \quad \Gamma_1 = 5, \quad and, \quad \Gamma_2 = 6,$$

$$ATE_{1,0} = 2, \quad and \quad ATE_{2,0} = 4.$$

2.5.2.1 Asymptotic Efficiency Outcomes

From an efficiency standpoint, let's first consider the models with no misspecification. In

Table B.1, the simulation results show that BS. SEs and M.C. SDs of the CF estimates for ψ_{og} (Γ_g) are almost always higher (lower) than BS. SEs and M.C. SDs of the counterpart IV estimates, respectively. Similarly, BS. SEs and M.C. SDs of the standard errors of CF estimates for ψ_{og} (Γ_g) are also almost always higher (lower) than than those of the IV estimates. Furthermore, BS. SEs and M.C. SDs of the CF estimates for $ATE_{g,0}$ are always higher than BS. SEs and M.C. SDs of the counterpart IV estimates. For instance, in Table B.1, the BS. SE of the CF parameter estimate for ψ_{o1} (Γ_1) is about 22% (22%) higher (lower) than that of the IV estimate, and the M.C. SD of the CF parameter estimate 59% (32%) higher (lower). Again in Table B.1, the BS. SE of the standard error of CF parameter estimate for ψ_{o1} (Γ_1) is about 57% (43%) higher (lower) than that of the IV estimate, and the M.C. SD of the standard error of CF parameter estimate 135% (57%) higher (lower). Furthermore, the BS. SE of the CF parameter estimate for $ATE_{1,0}$ is about 33% higher than that of the IV estimate, and the M.C. SD of the CF parameter estimate 67% higher. A very similar pattern is observed in Tables B.2 through B.8 as I switch the distribution of instrument z from asymmetric to symmetric and/or increase the sample size but with higher precision in estimates. As a result, when there is no misspecification, the simulation results demonstrate that neither the CF method nor the IV method performs better compared to the other method from the perspective of efficiency: The results suggest that CF method estimates Γ_g and their standard errors more precisely than does IV method; however, IV method estimates ψ_{og}, their standard errors, and ATEs more precisely than does CF method for $g = 0, 1, 2$.

Now let's take a look at the models with misspecification in Tables B.9 through B.16. In Table B.9, the simulation results still provide evidence for that that BS. SEs and M.C. SDs of the CF estimates for ψ_{og} (Γ_g) and $ATE_{g,0}$ are almost always higher (lower) than BS. SEs and M.C. SDs of the counterpart IV estimates, respectively. When it comes to BS. SEs and M.C. SDs of the standard errors of estimates for Γ_g, the simulation results seem to favor CF method: CF method has sharper estimates than does IV method. And from the dot plots

of estimated parameters, it seems like the abundance of outliers in IV estimates play a role in this observation, especially when the sample size is relatively small (i.e., $N = 1000, 2000$) and even after some trimming. As for BS. SEs and M.C. SDs of the standard errors of estimates for ψ_{og}, the simulation results are mixed. To give a few instances, in Table B.9, the BS. SE of the CF parameter estimate for ψ_{o1} ($ATE_{1,0}$) is around 88% (51%) higher than that of the IV estimate, and the M.C. SD of the CF parameter estimate 130% (126%) higher. On the other hand, the BS. SE (the BS. SE of the standard error) of CF parameter estimate for Γ_2 is 22% (77%) lower than that of the IV estimate, and the M.C. SD (the M.C. SD of the standard error) of CF parameter estimate 67% (97%) lower. A very similar pattern is observed in Tables B.10 through B.16 as I switch the distribution of instrument z from asymmetric to symmetric and/or increase the sample size but with higher precision in estimates. For example, the M.C. SD of the CF parameter estimate for ψ_{o0} ($ATE_{2,0}$) is .3526 (1.9578) in Table B.13 which is around 72% (57%) lower than 1.2723 (4.5059), the same CF estimate in Table B.9. As a result, when there is misspecification, the simulation results show that the efficiency results resemble to those when there is no misspecification and no method has a definite efficiency advantage over the other. The results indicate that CF method estimates Γ_g and their standard errors more precisely than does IV method; whereas, IV method estimates ψ_{og} and ATEs more precisely than does CF method for $g = 0, 1, 2$.

2.5.2.2 Asymptotic Unbiasedness and Consistency Outcomes

Using the asymptotic unbiasedness and consistency ideas from Chapter 1, let's first take a look at the results with no misspecification (e.g., those in Tables B.1 through B.8). In the absence of misspecification, the simulation results show that M.C. simulation estimates from CF method for both ψ_{og} and $ATE_{h,0}$ are very close to the true values (even when sample is relatively small), whereas the ones from IV method are not that close at all for $g = 0, 1, 2$ and $h = 1, 2$. For example, in Table B.1, M.C. simulation estimates from IV method for ψ_{o0}, ψ_{o1},

and ψ_{o2} are respectively 1.4465 (about 45% higher than the true value), 1.8806 (around 6% lower than the true value), and 2.8735 (around 4% lower than the true value) and are all off the true values, which causes severe biases in ATE estimates (about 28% lower in estimated $ATE_{1,0}$ and 14% lower in estimated $ATE_{2,0}$) even though M.C. simulation estimates from IV (and CF) method for Γ_0, Γ_1, and Γ_2 are very close to the true values. On the other hand, M.C. simulation estimates from CF method of both ψ_{og} and $ATE_{h,0}$ in Table B.1 are not off the true values at all with almost no biases. As the sample size increases from 1000 in Table B.1 to 10000 in Table B.4, M.C. simulation estimates from IV method do not improve on the biases; however, their BS. SEs and M.C. SDs get closer to zero just as those from CF method. A very similar pattern can also be seen in Tables B.5 through B.8 as I switch the distribution of instrument z from asymmetric to symmetric and/or increase the sample size. As a result, the simulation results indicate that, in the absence of misspecification, CF method is asymptotically unbiased and consistent while IV method is asymptotically biased and inconsistent (except for Γ_g, $g = 0, 1, 2$), which is supportive of the conjecture I made in subsection 2.3.1.

Here, I also need to mark that having unbiased and consistent estimates from IV method for Γ_g is in contrast to my expectations in subsection 2.3.1 since they are the coefficient estimates associated with the interaction terms $d_g(x^2)$ for $g = 0, 1, 2$. I guess one reason behind this result might be that the amount of exogenous variation coming from x^2 overpowers the endogeneity embedded in $d_g(x^2)$, at least in my simulation analysis.

Under the presence of misspecification, the simulation results in Tables B.9 through B.16 indicate that M.C. simulation estimates for ψ_{og} and $ATE_{h,0}$ from both CF method and IV method are off the true values for $g = 0, 1, 2$ and $h = 1, 2$. For instance, in Table B.9, M.C. simulation estimates from CF method for ψ_{o0}, ψ_{o1}, and ψ_{o2} are respectively .9461 (about 5% lower than the true value), 1.5961 (around 20% lower than the true value), and 4.9293 (around 64% higher than the true value) and are all off the true values, which leads to drastic biases in ATE estimates (about 19% lower in estimated $ATE_{1,0}$ and 50% higher in

estimated $ATE_{2,0}$) even though M.C. simulation estimates from both IV and CF methods for Γ_0, Γ_1, and Γ_2 are again very close to the true values as in the case of no misspecification. As noted before, M.C. simulation estimates from IV method are also not close to the true values. However, there is a difference: Biases in the estimates from IV method are lower than those in the estimates of CF method except in the estimates for ψ_{o0} and $ATE_{1,0}$. For example, in Table B.9, M.C. simulation estimates from IV method for ψ_{o2} and $ATE_{2,0}$ are respectively 3.0860 (around 3% higher than the true value) and 3.6704 (around 8% lower than the true value) but these biases are all smaller than their counterparts from CF method, 64% and 50% respectively. As the sample size increases from Table B.10 to Table B.12, M.C. simulation estimates from both IV and CF methods do not improve on the biases; however, their BS. SEs and M.C. SDs get smaller. A very similar pattern can also be seen in Tables B.13 through B.16 as I switch the distribution of instrument z from asymmetric to symmetric and/or increase the sample size. Actually, the simulation results in Tables B.13 through B.16 suggest that IV method performs even better in terms of unbiasedness with more biases only in the estimates for ψ_{o0}. As a result, the simulation results indicate that, under misspecification, CF method is asymptotically more biased compared to IV method, and both methods are inconsistent.

2.6 Conclusion

In this chapter, I introduce an econometric model with a discrete multivalued endogenous treatment variable and CRCs and show how to consistently estimate ATEs by a three step estimation procedure of CF method in such a model where the endogeneity problem is further exacerbated compared to the one in the model without CRCs as in Chapter 1. Moreover, I state that, based off the theorems developed in Chapter 1 and restated in this chapter, the asymptotic distribution of the CF estimates follows a normal distribution, and the CF estimates are $\sqrt{N} - consistent$. I propose a consistent estimator for the asymptotic variance

matrix of CF estimates, which takes into consideration the nonlinear first stage estimation, following the analogy principle as in Chapter 1. For those who do not like to go through multistep estimation, I also express how one can consistently estimate ATEs and obtain valid standard errors for the parameters of interest by using GMM.

In my simulation analysis, I compare CF method and IV method under various specifications with and without misspecification. In the absence of misspecification, the simulation results suggest that CF method is asymptotically unbiased and consistent (but not necessarily more efficient because, for some parameters, IV method provides sharper estimates than CF method). Whereas, IV method is generally asymptotically biased and inconsistent, which is more pronounced when the instrument is asymmetrically distributed. Therefore, the simulation results point that, without misspecification, CF method can consistently estimate ATEs, while IV method cannot. In the presence of misspecification, the simulation results show that both CF and IV methods have biased estimates. However, biases in IV estimates are generally lower than those in the estimates of CF method, which is more obvious when the instrument is symmetrically distributed. With regard to efficiency, the findings from simulations are mixed in the sense that none of the methods outperforms the other one clearly. In addition, especially in the presence of misspecification, the simulation results point that IV method can less precisely estimate the standard errors of standard errors when sample size is relatively small.

All of the research ideas mentioned in the conclusion of Chapter 1 can be explored in a discrete multivalued endogenous treatment model with CRCs, as well. In addition to these research ideas though, it can be worth the time and effort to investigate the ways in which we can extend the model in this chapter to the framework of panel data models. We can also develop tests that measure the existence of CRCs and the degree of endogeneity attached to them as in Heckman, Schmierer, and Urzua (2010). Furthermore, one can also examine the possibility of devising a consistent IV method, i.e., a correction function approach as in Wooldridge (2008), and its large sample properties for the model presented in this chapter.

CHAPTER 3

ESTIMATION FOR MULTIVALUED ENDOGENOUS TREATMENT EFFECT MODELS USING HIGH DIMENSIONAL METHODS: A SIMULATION STUDY

3.1 Introduction

After the expansion of internet usage in early 2000s, digitization accelerated and penetrated all facets of society and science, including the field of economics. As described in Athey and Luca (2019), many technology companies (e.g., Google, Apple, Facebook, Amazon, and Microsoft) have been increasingly employing economists in order to address problems such as individualized marketing and promotions, optimal pricing, auction platform design, and intervention effects. The power of digitization combined with its absorption by technology companies (and now rapidly by other traditional companies as well) also leads to new opportunities for collaborations with academics (e.g., Golub Capital Social Impact Lab at Stanford Graduate School of Business). As a result, in recent years, economists have been making use of big data, which causes the rising popularity of machine learning (ML) techniques in economics and the attempts to improve upon existing econometric approaches by incorporating ML ideas.

The survey article of Donaldson and Storeygard (2016) summarizes several examples of how ML methods are applied in development economics. Mullainathan and Spiess (2017) provide a brief overview of the business-oriented prediction and classification problems (e.g., house valuation, industry classification, and hiring decision) in which ML methods are used. There has been a decent usage of ML methods in policy determination and assignment in economics too, see Athey (2018) for a review. Several econometric theory results for ML methods with regard to estimation and inference have been established in research areas such as treatment effects, structural models of consumer choice, panel data models, and model selection. For instance, in randomized experiments, Chernozhukov *et al.* (2020) advance

generic estimation and inference results for heterogeneous treatment effects that are valid under the usage of various ML techniques such as boosted trees, ensemble methods, neural networks, random forests, and regularized methods. Athey, Tibshirani, and Wager (2019) develop generalized random forests that allow the estimation of any population quantities identified in moment conditions as in generalized method of moments (GMM) method. Farrell, Liang, and Misra (2021) derive new rates of convergence for deep neural networks and use these rates to develop semiparametric inference on parameters of interests. Iskhakov, Rust, and Schjerning (2020) look at how ML can further contribute to structural econometrics. The application of ML methods in panel data models has received some attention from economists, see, for example, Belloni *et al.* (2016); Kock (2016); Chernozhukov, Wuthrich, and Zhu (2019); Semenova *et al.* (2021); and Athey *et al.* (2021). For more on the list of economics research areas in which economists can benefit from the tools originated in the ML literature, see Athey and Imbens (2019).

In recent years, there have been developments in high dimensional regularized models applied to economics. As one of the earliest works in high dimensional regularized models, Belloni and Chernozhukov (2011a) present concepts related to high dimensional sparse econometric models and their estimation using $\ell 1$-regularized and post $\ell 1$-regularized ML methods, particularly least absolute shrinkage and selection operator (LASSO), in linear and nonparametric settings. Technically, the ideas used in this paper go all the way back to Belloni and Chernozhukov (2013)'s 2009 version available in arXiv.org. In these papers, high dimensionality means that the number of parameters to be estimated in an econometric model is more than sample size, and sparsity means that the number of covariates with nonzero coefficients is in reality less than sample size and unknown. In a pioneering follow-up work, Belloni, Chernozhukov, and Hansen (2011a) share inferential (and estimation) results for linear instrumental variables (IV) model with many instruments and partially linear models in high dimensional sparse setting. In another influential paper, Belloni and Chernozhukov (2011b) develop regularized quantile regression in high dimensional sparse

models, show the consistency of their estimator for this model, and evaluate its performance by simulation. Later, Belloni, Chernozhukov, and Kato (2019) also establish new inference methods for the parameters from $\ell 1$-penalized quantile regression in high dimensional sparse settings. For an early overview of estimation techniques and inference in high dimensional models with a focus on model selection and LASSO methods, see Belloni, Chernozhukov, and Hansen (2014b) and Chernozhukov, Hansen, and Spindler (2015a).

In the last decade, there are also some notable work in the casual and debiased estimation of linear high dimensional models. For example, Belloni *et al.* (2012) employ LASSO to select instruments, show the asymptotic properties of the linear IV estimator that is based on post LASSO method dampening shrinkage bias associated with LASSO, and use partialled-out variables for inference when instruments are weak. Belloni, Chernozhukov, and Wei (2016) extend the results in this paper to generalized linear high dimensional regression models with regular sparsity assumption. Belloni, Chernozhukov, and Hansen (2014a) introduce a new estimation method called post double selection and provide uniformly valid post selection inference for treatment effects in sparse high dimensional models. This method is robust to imperfect selection of the controls and allows for non-Gaussian and heteroscedastic disturbances too. Jointly using Neyman-orthogonal scores and crossfitting in Chernozhukov *et al.* (2018), the authors propose residual-on-residual regression method to remove biases associated with regularized ML methods off causal parameters of interest and construct valid confidence intervals for these parameters. For more on debiased estimation and valid post estimation inference results and applications, see Belloni *et al.* (2017); Chernozhukov *et al.* (2017); Chernozhukov, Newey, and Singh (2021); and Athey, Imbens, and Wager (2018). In the last couple of years, there is also some interest in high dimensional moment-based regularized models and high dimensional models with measurement error. For more on these, see, for example, Belloni *et al.* (2018a); Belloni *et al.* (2018b); Caner and Kock (2019); Bach *et al.* (2020); and Belloni, Chernozhukov, and Kaul (2017); Chernozhukov, Wuthrich, and Zhu (2018); Belloni, Kaul, and Rosenbaum (2019); and Chernozhukov *et al.* (2020).

In this chapter, I extend my work from Chapter 1 to a high dimensional sparse model in a particular setting. Imagine that there exists an extra set of high dimensional variables. And a low dimensional (and unknown) subset of these variables has an impact on the outcome (so some of these variables are relevant in the outcome equation); however, all of these high dimensional variables are totally ignorable (or redundant) to the decision to undertake the treatment given some instruments in the selection equation, which is not unheard of in experimental intervention studies. Long known as in Cochran (1957), the addition of distinctly relevant variables into a regression often results in more precise estimation and better prediction. For this reason, the inclusion of high dimensional variables in estimation can potentially improve the efficiency and predictive power of the model in Chapter 1 without jeopardizing its consistency results of interest when the high dimensional variables are orthogonal to (or uncorrelated with) the variables of interest already included. However, the high dimensionality of these extra variables requires the usage of estimation methods that are capable of doing variable selection (especially needed if the sample size is also small) among the extra set of high dimensional variables really influencing the outcome of interest and that produce reliable inference results. Through a simulation analysis, this chapter aims to guide economists in if (and which) LASSO-based inference and variable selection methods would perform better than the control function (CF) estimator from Chapter 1 for discrete multivalued endogenous treatments in a linear scalar outcome high dimensional sparse model with heterogeneous counterfactual errors just described in previous sentences. I also allow for non-Gaussian disturbances in my particular setting. To address endogeneity, I again use the CF approach from Chapter 1: First, a multinomial logit model for the treatment decision in a setting that is not high dimensional is estimated by maximum likelihood to construct CF variables. Second, these CF variables are added into outcome equation in high dimensional setting to be estimated by different ML methods.

For the parameter estimation in the outcome equation with high dimensional variables, I specifically use four different ML methods: LASSO; post partial-out LASSO of Belloni *et*

al. (2012); post double selection LASSO of Belloni, Chernozhukov, and Hansen (2014a); and double/debiased ML LASSO of Chernozhukov *et al.* (2018). In my simulations, I also include the CF method from Chapter 1 as a benchmark with the aim of comparing the finite sample performances of these four LASSO-based ML methods to the CF's. In the comparison, I employ measures such as bias of ATE estimates, standard deviation of ATE estimates, mean absolute prediction error, root mean square error (RMSE), mean number of correctly selected covariates, and mean size of selected set of covariates. To the best of my knowledge, this chapter is the first simulation-based comparative analysis of the LASSO-based methods above in a discrete multivalued endogenous treatment model of linear high dimensional sparse setting with heterogeneous counterfactual errors. The main simulation finding in this setting is that, on top of being on par with the CF method in finite sample bias ground, the LASSO-based methods can surpass the efficiency performance of the CF method in ATE estimation if there exist enough extra predictive variables that are ignorable in treatment selection among a set of high dimensional predictors of outcome.

The rest of this chapter is organized as follows. In section 3.2, I introduce the model. In section 3.3, I summarize the ML methods and the procedure to estimate the parameters of interest. In section 3.4, I share some simulation results. In section 3.5, I conclude. And, in appendix C, I share simulation tables that are hidden from the main body of this chapter.

3.2 The Model

Consider the model from Chapter 1 augmented by the presence of high dimensional variables

$$\begin{aligned} y_g &= \alpha_g + \mathbf{x}\beta_g + \mathbf{h}\delta_g + u_g \\ w_g^* &= \mathbf{z}\gamma_g + a_g, \end{aligned} \tag{3.1}$$

where y_g is the g^{th} counterfactual outcome variable, α_g is the scalar coefficient in the counter-

factual outcome equation for y_g, $\mathbf{x} \equiv (x_1, x_2, \ldots, x_l)$ is the $1 \times l$ vector of exogenous variables in y_g, β_g is the $l \times 1$ vector of slope coefficients associated with $\mathbf{x}$ in y_g, $\mathbf{h} \equiv (h_1, h_2, \ldots, h_{l_h})$ is the $1 \times l_h$ vector of high dimensional exogenous variables in y_g, δ_g is the $l_h \times 1$ vector of slope coefficients associated with $\mathbf{h}$ in y_g, u_g is the counterfactual error in y_g, w_g^* is the latent treatment variable that determines the choice of treatment status among $G+1$ alternative treatment statuses, $\mathbf{z} \equiv (z_1, z_2, \ldots, z_k)$ is the $1 \times k$ vector of instruments that includes a constant term in the choice equation for w_g^*, γ_g is the $k \times 1$ vector of parameters in w_g^*, and a_g is the scalar error term that is independently and identically Gumbel distributed (*i.i.d.*) with location parameter $\mu = 0$ and scale parameter $\beta = 1$ in w_g^* for $g = 0, 1, \ldots, G$. Here, high dimensionality means that l_h is bigger than the sample size N (i.e., $l_h > N$) available in estimation.

As in Chapter 1, let $w \in \{0, 1, \ldots, G\}$ be the observed discrete multivalued endogenous treatment variable whose values are determined by w_g^* that can be regarded as the utility or satisfaction obtained from treatment status g for $g = 0, 1, \ldots, G$. Let the treatment statuses of w be exhaustive and mutually exclusive. Define binary treatment status indicators, $d_g = 1[w = g]$ for $g = 0, 1, \ldots, G$. So the binary treatment status indicator d_g is equal to one if the treatment status is equal to g and zero otherwise. This coupled with the mutual exclusivity of treatment statuses implies that $\sum_{g=0}^{G} d_g = 1$. Define the $1 \times (G+1)$ vector of treatment statuses $\mathbf{d} \equiv (d_0, d_1, \ldots, d_G)$. Let y be the observed outcome. Then, I can write

$$y = d_0 y_0 + d_1 y_1 + \cdots + d_G y_G, \tag{3.2}$$

where y_g is the g^{th} counterfactual outcome for $g = 0, 1, \ldots G$.

After having described the discrete multivalued endogenous treatment model with heterogeneous counterfactual errors and high dimensional variables, I now will list a series of assumptions some of which have also been made in Chapter 1 to further develop the model. First, I assume that the rational economic agents choose the status of treatment from which they receive the most satisfaction out of all possible treatment statuses. That is,

- **Assumption 3.1 (A.3.1)**: One chooses treatment status g, i.e., $w = g$ if and only if $w_g^* \geq w_j^* \ \forall j \neq g$ for $g, j = 0, 1, \ldots, G$.

Second, I assume that identification of the model in (3.1) and (3.2) is aided by the exclusion of some (at least one) variables in the set of instruments $\mathbf{z}$ from the set of exogenous variables in $\mathbf{x}$. The set of exogenous variables in $\mathbf{x}$ can all be included in the set of instruments $\mathbf{z}$.

- **Assumption 3.2 (A.3.2)**: Identification of the model described by (3.1) and (3.2) is strengthened by exclusion of at least one variable in $\mathbf{z}$ from the set of variables in $\mathbf{x}$.

In economics, it is often the case that $\mathbf{z}$ includes all the variables in $\mathbf{x}$. However, this is not an absolute necessity in my model as long as the exclusion restriction above is satisfied. Third, I make an assumption about the irrelevancy of $\mathbf{h}$ given the set of instruments $\mathbf{z}$ in the g^{th} choice equation for $g = 0, 1, \ldots G$.

- **Assumption 3.3 (A.3.3)**: $D(w_g^*|\mathbf{z}, \mathbf{h}) = D(w_g^*|\mathbf{z})$ where $D(\cdot\,|\cdot)$ means conditional distribution. That is, conditional on $\mathbf{z}$, w_g^* (and thereof w) is independent of $\mathbf{h}$.

Conceptually, having variables, such as the ones in $\mathbf{h}$, that appear only in the outcome equation is standard in a randomized controlled trial (RCT) due to treatment selection (or assignment)'s being totally random. For this reason, in RCTs, the propensity score for treatment does not depend on any variables; whereas, the outcome of interest can depend on some variables. Apart from RCTs, here is a framework where A.3.3 is reasonable. Suppose an economist have an experimental intervention with exogenous instruments. In line with

Vella and Verbeek (1999, p. 474) and Heckman (1990, p. 314), the experimental intervention with exogenous instruments means that, using exogenous instruments, unobservables that play a role in the treatment decision and that are correlated with the outcome of interest have been taken into consideration. And further assume that this economist can create so reliable and strong instruments $\mathbf{z}$ via surveys that these instruments include all necessary observed information to make treatment decisions. Well then in this case, it is plausible that the treatment variable w out of this intervention is related only to the instruments $\mathbf{z}$ and is conditionally independent of other outcome-related (and maybe high dimensional) observables $\mathbf{h}$ once $\mathbf{z}$ is controlled for.

It is also important to make the distinction between the set of variables in $\mathbf{x}$ and the set of high dimensional variables $\mathbf{h}$ in the counterfactual outcome equation y_g. $\mathbf{x}$ can be thought as micro-level structural characteristics, and $\mathbf{h}$ as macro-level characteristics. To make the difference between $\mathbf{x}$ and $\mathbf{h}$ clearer, consider the following example from development economics similar to that in Danquah *et al.* (2021). Imagine that one studies how influential household gender wage gap is on women's empowerment, using survey data with limited number of observations (but with a dense/high dimensional set of variables created based on survey answers and outside data sources). In such a study, the dependent variable can be the share of household assets owned by women. The treatment variable can be gender wage gap at household level grouped in low, medium, and high levels of relative difference between average male and female adult household members' earnings. $\mathbf{x}$ can be exogenous covariates such as household and family characteristics that affect the share of household assets owned by women. On the other hand, $\mathbf{h}$ can be high dimensional regressors such as occupation, sector, access to certain infrastructure and services, location, and social norms that impacts the share of household assets owned by women.

As a result, the model in (3.1) and (3.2) together with the assumptions made so far still allows the treatment variable w to follow a multinomial logit model with choice probabilities given as follows:

$$P(w = g|\mathbf{x}, \mathbf{h}, \mathbf{z}) = P(w = g|\mathbf{z}) = exp(\mathbf{z}\gamma_{\mathrm{g}})/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_{\mathrm{r}}), \tag{3.3}$$

for $g = 0, 1, \ldots, G$. See McFadden (1973) for a similar result without the high dimensional variables $\mathbf{h}$. The next assumption is the Dubin and McFadden's linearity assumption combined with the redundancy of $\mathbf{x}$, $\mathbf{h}$, and $\mathbf{z}$ for the expectation of u_g conditional on $\mathbf{a}$.

- **Assumption 3.4 (A.3.4)**: $E(u_g|\mathbf{x}, \mathbf{h}, \mathbf{z}, \mathbf{a}) = E(u_g|\mathbf{a}) = \sum_{j=0}^{G} \eta_{g,j} a_j + [-\sum_{j=0}^{G} \eta_{g,j} E(a_j)]$, where u_g is the counterfactual error in y_g, $\mathbf{x}$ is the $1 \times l$ vector of exogenous variables in y_g, $\mathbf{h}$ is the $1 \times l_h$ vector of high dimensional exogenous variables in y_g, $\mathbf{z}$ is the $1 \times k$ vector of instruments that includes a constant term in the choice equation for w_g^*, $\mathbf{a} \equiv (a_0, a_1, \ldots, a_G)$ is the $1 \times (G+1)$ vector of *i.i.d.* Gumbel distributed errors a_j with location parameter $\mu = 0$ and scale parameter $\beta = 1$ in w_j^*, $\eta_{g,j}$ is the scalar multiple of correlation coefficient between u_g and a_j, and $E(a_j) = 0.5772$ is Euler's constant for $j, g = 0, 1, \ldots, G$.

Using all the assumptions from A.3.1 through A.3.4, I create CF terms as derived in Chapter 1 to deal with endogeneity in w. In section 3.3, I will introduce the ML methods I employ to estimate the parameters of interest in (3.1).

3.3 Estimation

To eliminate the complications of endogeneity in w, I again rely on the idea of CF approach presented in Chapter 1. Referring back to appendix A, it is pretty straightforward

to derive the estimating equation for the parameters of interest in (3.1): Just add $\mathbf{h}$ as another conditioning variable into the equations in that section. Then, the baseline estimating equation for the regressions augmented by the CF terms in this section is as below:

$$\begin{aligned} y &= \sum_{j=0}^{G} d_j \alpha_j + \sum_{k=0}^{G} d_k \mathbf{x} \beta_k + \sum_{m=0}^{G} d_m \mathbf{h} \delta_m + \\ &+ \sum_{g=0}^{G} [-\eta_{g,g} d_g log(\Lambda_g)] + \sum_{g \neq 0} d_g \eta_{g,0} M_0 + \sum_{g \neq 1} d_g \eta_{g,1} M_1 + \cdots + \\ &+ \sum_{g \neq G} d_g \eta_{g,G} M_G + \mathfrak{e}, \end{aligned} \tag{3.4}$$

where $E(\mathfrak{e}|\mathbf{d}, \mathbf{x}, \mathbf{h}, \mathbf{z}) = 0$, $\Lambda_j = exp(\boldsymbol{z}\gamma_j)/\sum_{r=0}^{G} exp(\boldsymbol{z}\gamma_r)$, and $M_j = \Lambda_j log(\Lambda_j)/(1 - \Lambda_j)$ for $j = 0, 1, \ldots, G$. Note that because of A.3.3, the high dimensional variables $\mathbf{h}$ are in fact not needed for consistently estimating the partial effects α_j (and thereof ATEs if $E(\mathbf{x}) = E(\mathbf{h}) = 0$) for $j = 0, 1, \ldots, G$. Therefore, the CF estimator from Chapter 1 applied to (3.4) without the terms $\sum_{m=0}^{G} d_m \mathbf{h} \delta_m$ still purges endogeneity of the model and is still valid for producing consistent partial effect estimates for α_j. Before moving forward with the ML methods and their procedures, let me make the following assumption regarding the sparsity in (3.4).

- **Assumption 3.5 (A.3.5)**: A random sample $\{w_i, \mathbf{z}_i, y_i, \mathbf{x}_i, \mathbf{h}_i\}$ of N observations (i.e., $i = 1, 2, \ldots, N$) is available for the treatment choice and the outcome equations. The number of parameters in both γ_g and β_g is way less than the sample size N (i.e., $k << N$ and $l << N$). Suppose $\mathbf{h}\delta_g = (\mathbf{h}_1 \delta_{g_1} + \mathbf{h}_2 \delta_{g_2})$ where $\mathbf{h}_1$ is the $1 \times l_{h_1}$ vector of exogenous variables in $\mathbf{h}$ associated with nonzero slope coefficients δ_{g_1}, and $\mathbf{h}_2$ is the $1 \times l_{h_2}$ vector of high dimensional (i.e., $l_{h_2} > N$) exogenous variables in $\mathbf{h}$ associated with zero slope coefficients δ_{g_2}. It is not known which variables in $\mathbf{h}$ belong to $\mathbf{h}_1$ or $\mathbf{h}_2$. However, it is true that the number of nonzero slope coefficients in δ_g is way less than the sample size N (i.e., $l_{h_1} << N$, the sparsity condition) such that the number

of coefficients to be estimated in (3.4), say p', is in reality way less than the sample size (i.e., $p' = (l + l_{h_1} + G + 2)(G + 1) << N$)).

A.3.5 implies that, given the assumptions made earlier, the multinomial logit model of w on $\mathbf{z}$ can be consistently estimated using the standard asymptotic theory. On the other hand, since the variables associated with nonzero coefficients in $\mathbf{h}$ is not known, there are potentially $p = (l + l_h + G + 2)(G + 1)$ variables in (3.4). And this number p, already greater than N as assumed by A.3.5, can grow even further as the number of treatment status increases. If one wants to take advantage of the variables in $\mathbf{h}$ while estimating, then the high dimensionality in (3.4) creates a big estimation problem that cannot be handled by low dimensional methods such as the CF estimator from Chapter 1 and that causes the researchers to utilize other estimation methods designed for high dimensional and sparse settings. Due to the sparsity condition made in A.3.5 and the linearity of (3.4) in parameters, LASSO estimation method and its unbiased versions can be used to estimate the parameters of interest in (3.4), e.g., the partial effects α_j for $j = 0, 1, \ldots, G$.

As suggested just above, one can benefit from the existence of the variables in $\mathbf{h}$ using LASSO-based methods. Owing to the relevancy of the high dimensional variables in $\mathbf{h}$ with the outcome variable y and the independence of treatment decision from $\mathbf{h}$, estimating (3.4) by using LASSO-based methods can improve the efficiency and predictive power of the model over that by the CF estimator from Chapter 1 without dangering consistency. For example, in low dimensional settings with randomly assigned treatment, Imbens and Rubin (2015), Lin (2013), and Negi and Wooldridge (2021) all point out efficiency gains out of adding extra variables into regression that are sufficiently predictive of the outcome. Considering all these, with strong instruments $\mathbf{z}$ and $\mathbf{h}_1$ highly predictive of the outcome, it can worthwhile from efficiency and prediction perspectives (and hopefully biaswise as well) to estimate (3.4) by (at least one of) the LASSO-based methods which are more complicated and more time-consuming than the CF estimator from Chapter 1. Now, I can describe the

four LASSO-based methods I will use to estimate the parameters of interest in (3.4) under all assumptions from A.3.1 through A.3.5.

3.3.1 LASSO Estimation

Developed by Tibshirani (1996), LASSO is a regularized regression method similar to ridge regression that economists are familiar with and use when they want to decrease the severity of multicollinearity among covariates at the expense of shrinking their coefficient estimates using a penalty on coefficient sizes. Thus, LASSO and its variants are sometimes also called shrinkage or penalized regression methods. In my model given by (3.1) and (3.2) together with all assumptions from A.3.1 through A.3.5, LASSO estimator solves the following problem

$$\underset{\theta\in\Theta}{min}\left\{\sum_{i=1}^{N}(y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{z}_i, \hat{\gamma}), \theta))^2/2 \;+\; \lambda\sum_{j=1}^{p}|\theta_j|\right\}, \tag{3.5}$$

where $m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{z}_i, \hat{\gamma}), \theta) \equiv \mathbb{X}_i\vartheta + \mathbf{v}_i(\hat{\gamma})\varphi$ is a real-valued scalar function, $\hat{\gamma} = (\hat{\gamma}_0', \hat{\gamma}_1', \ldots, \hat{\gamma}_G')'$ is the $(G+1)k \times 1$ vector of $\sqrt{N} - consistent$ and asymptotically normal first stage conditional maximum likelihood estimates from the multinomial logit (MNL) regression of w_i on $\mathbf{z}_i$ for $i = 1, 2, \ldots, N$, $\theta = (\theta_1, \theta_2, \ldots, \theta_p) = (\vartheta', \varphi')'$ is the $p \times 1$ vector of parameters, $|\cdot|$ is absolute value operator, $\sum_{j=1}^{p}|\theta_j|$ is $\ell 1$ norm (a.k.a. $\ell 1$ LASSO penalty), λ is the Lagrange multiplier (a.k.a. the tuning parameter that determines the strength of the penalty), $\mathbb{X}_i$ is the $1 \times (l_h + l + 1)(G+1)$ vector of regressors, and $\mathbf{v}_i(\hat{\gamma})$ is the $1 \times (G+1)(G+1)$ vector of generated regressors (actually all CF terms).

More clearly,

$$
\begin{aligned}
\mathbb{X}_i &= (\ d_{0_i}, \ d_{1_i}, \ \cdots, \ d_{G_i}, \ d_{0_i}\mathbf{x}_i, \ d_{1_i}\mathbf{x}_i, \ \cdots, \ d_{G_i}\mathbf{x}_i, \ d_{0_i}\mathbf{h}_i, \ d_{1_i}\mathbf{h}_i, \ \cdots, \ d_{G_i}\mathbf{h}_i \ ) \\
\mathbf{v}_i(\hat{\gamma}) &= (\ -d_{0_i}log(\hat{\Lambda}_{0_i}), \ \cdots, \ -d_{G_i}log(\hat{\Lambda}_{G_i}), \ d_{1_i}\hat{M}_{0_i}, \ d_{2_i}\hat{M}_{0_i}, \ \cdots, \ d_{G_i}\hat{M}_{0_i}, \ d_{0_i}\hat{M}_{1_i}, \\
&\quad d_{2_i}\hat{M}_{1_i}, \ d_{3_i}\hat{M}_{1_i}, \ \cdots, \ d_{G_i}\hat{M}_{1_i}, \ \cdots, \ d_{0_i}\hat{M}_{G_i}, \ d_{1_i}\hat{M}_{G_i}, \ \cdots, \\
&\quad , d_{G-1_i}\hat{M}_{G_i}), \qquad (3.6)
\end{aligned}
$$

where $\hat{\Lambda}_{g_i} = exp(\boldsymbol{z}_i\hat{\gamma}_g)/\sum_{r=0}^{G} exp(\boldsymbol{z}_i\hat{\gamma}_r)$ and $\hat{M}_{g_i} = \hat{\Lambda}_{g_i}log(\hat{\Lambda}_{g_i})/(1-\hat{\Lambda}_{g_i})$ for $g = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$.

Notice that LASSO estimator minimizes (3.5) only with respect to θ not with respect to θ and λ together. Indeed, the minimization is done for given values of λ,so the tuning parameter needs to be chosen before the minimization problem. One of the most frequent approaches to choose λ is cross validation which typically runs LASSO (on the training set-some portion of sample) using a candidate λ from a grid of λ values and selects the value of λ to be used in the minimization problem by minimizing, say, the validation set (the remaining portion of sample) prediction error. Also note that, as λ gets larger, $\ell 1$ LASSO penalty gets heavier (meaning more bias in coefficients) and the selected model gets sparser (i.e., more coefficients are set exactly equal to zero). To learn more on shrinkage methods and cross validation, see subchapters 3.4 and 7.10 in Hastie, Tibshirani, and Friedman (2009) respectively. Another method to choose λ comes from the penalty level formula (12) in Belloni, Chernozhukov, and Wang (2011, p.795). The formula is given below:

$$
\lambda = c\sqrt{N}\Phi^{-1}(1-\alpha/2p), \qquad (3.7)
$$

where c=1.1, N is the sample size, Φ^{-1} is the inverse standard normal cumulative distribution function, α=.05, and p is the number of potential variables in regression which is equal to l_h in my case. This method is especially designed for LASSO-based inference methods. It does

not estimate the model several times for other values of λ from a grid of λ values, which makes it much faster than cross validation. Compared to cross validation, LASSO-based inference methods using (3.7) tend to better select the variables that have true nonzero impact on the outcome. For these reason, I also prefer using (3.7) over cross validation in this chapter.

Now, I can prescribe the following LASSO procedure to estimate the partial effects α_j in high dimensional and sparse setting for $j = 0, 1, \ldots, G$:

Procedure 3.1

1. Estimate the predicted probabilities, $\hat{\Lambda}_{j_i} = exp(\mathbf{z}_i\hat{\gamma}_{\mathrm{j}})/\sum_{r=0}^{G} exp(\mathbf{z}_i\hat{\gamma}_{\mathrm{r}})$, from a MNL of w_i on $\mathbf{z}_i$ for $j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$.

2. Run the LASSO regression of y_i on $\mathbb{X}_i$ and $\mathbf{v}_i(\hat{\gamma})$ with only $d_{j_i}\mathbf{h}_i$'s to be selected.

3. Obtain parameter estimates of d_j's from step 2,

where $\mathbb{X}_i$ and $\mathbf{v}_i(\hat{\gamma})$ are defined as in (3.6) for $j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$. Due to the existence of generated regressors in the model, the standard errors and confidence intervals associated with parameter estimates from LASSO unfortunately are not valid. For this reason, I rely on Monte Carlo simulation to draw inference, which is also true for all the other LASSO-based estimations below.

3.3.2 Post Partial-out LASSO Estimation

The main disadvantage of LASSO is that its parameter estimates are biased towards zero after regularization by $\ell 1$ penalty. Therefore, for inference purposes, researchers need to handle the regularization bias. To this end, there have been several LASSO-based methods

proposed by economists, and one of them is called post partial-out LASSO which is based on Belloni et al. (2012). "Post" part simply means the application of linear regression to the model selected by LASSO and alleviates biases in parameter estimates of LASSO due to regularization. "Partial-out" part deals with endogenous covariates and the inclusion of irrelevant variables (or the exclusion of relevant variables).

Belloni et al. (2012) imply that the score equations from the linear regression of an partialled-out outcome on partialled-out covariates are immune to endogeneity and selection mistakes. Here, the partialled-out outcome is the residual after running a linear regression of the outcome on the LASSO-selected covariates, and similarly the partialled-out covariate is the residual after running a linear regression of that covariate on the other LASSO-selected covariates. Because of this partialling-out procedure, the post partial-out LASSO can be called the post residual-on-residual LASSO with references to the conventional partialling-out estimators

Adapting closely from Algorithm 1 given in Chernozhukov, Hansen, and Spindler (2015b), the post partial-out LASSO procedure I can prescribe to estimate the partial effects in high dimensional and sparse setting is as follows:

Procedure 3.2

1. Same as in Procedure 3.1.

2. Run a LASSO regression of d_{j_i} on $d_{0_i}\mathbf{x}_i$, $d_{1_i}\mathbf{x}_i \ldots$, $d_{G_i}\mathbf{x}_i$, $d_{0_i}\mathbf{h}_i$, $d_{1_i}\mathbf{h}_i$, $\ldots$, $d_{G_i}\mathbf{h}_i$, and $\mathbf{v}_i(\hat{\gamma})$, where only $d_{j_i}\mathbf{h}_i'$s are to be selected, and $d_{j_i}\mathbf{x}_i'$s and $\mathbf{v}_i(\hat{\gamma})$ are forced to be included in the selected controls. Let's denote the selected controls by $\mathbf{s}_{j_i}^d$ for $j = 0, 1, \ldots, G$.

3. Run a regression of d_{j_i} on $\mathbf{s}_{j_i}^d$. Let $\hat{d}_{j_i}$ be the residuals from this regression for $j = 0, 1, \ldots, G$.

4. Let $\hat{d}_i = (\hat{d}_{0_i}, \hat{d}_{1_i}, \ldots, \hat{d}_{G_i})$ be the collection of all the residuals from step 3.

5. Run a LASSO regression of y_i on $d_{0_i}\mathbf{x}_i$, $d_{1_i}\mathbf{x}_i$, ..., $d_{G_i}\mathbf{x}_i$, $d_{0_i}\mathbf{h}_i$, $d_{1_i}\mathbf{h}_i$, ..., $d_{G_i}\mathbf{h}_i$, and $\mathbf{v}_i(\hat{\gamma})$, where only $d_{j_i}\mathbf{h}_i'$s are to be selected, and $d_{j_i}\mathbf{x}_i'$s and $\mathbf{v}_i(\hat{\gamma})$ are forced to be included in the selected controls. Let's denote the selected controls by $\mathbf{s}_i^y$.

6. Run a regression of y_i on $\mathbf{s}_i^y$. Let $\hat{y}_i$ be the residuals from this regression.

7. Run a regression of $\hat{y}_i$ on $\hat{d}_i$.

8. Obtain parameter estimates of $\hat{d}$ from step 7,

where $\mathbf{v}_i(\hat{\gamma})$ is defined as in (3.6) for $i = 1, 2, \ldots, N$.

3.3.3 Post Double Selection LASSO Estimation

Another method to lessen the regularization bias on parameters estimated by LASSO is called post double selection LASSO given by Belloni, Chernozhukov, and Hansen (2014a). Technically, the post double selection LASSO is a simplified version of the post partial-out LASSO without partialling out outcome variable and other covariates. However, it is still robust to imperfect selection of the covariates on top of reducing bias on parameter estimates. "Post" part means the same as in the post partial-out LASSO. "Double selection" part, on the other hand, means that covariates are selected for predicting both the variables of interest and the outcome. This double selection of covariates to be included in the final estimating equation is done with the purpose of strengthening the validity of inference results by adding all the important and relevant variables that are correlated with the variables of interest and the outcome.

In line with the estimation steps laid out in Belloni, Chernozhukov, and Hansen (2014a, p. 610), the post double selection LASSO procedure I can prescribe to estimate the partial effects in high dimensional and sparse setting is as follows:

Procedure 3.3

1. Same as in Procedure 3.1.

2. Same as in Procedure 3.2.

3. Run a LASSO regression of y_i on $d_{0_i}\mathbf{x}_i$, $d_{1_i}\mathbf{x}_i$, ..., $d_{G_i}\mathbf{x}_i$, $d_{0_i}\mathbf{h}_i$, $d_{1_i}\mathbf{h}_i$, ..., $d_{G_i}\mathbf{h}_i$, and $\mathbf{v}_i(\hat{\gamma})$, where only $d_{j_i}\mathbf{h}_i'$s are to be selected, and $d_{j_i}\mathbf{x}_i'$s and $\mathbf{v}_i(\hat{\gamma})$ are forced to be included in the selected controls. Let's denote the selected controls by $\mathbf{s}_i^y$.

4. Let $\mathbf{s}_i \equiv \left\{\bigcup_{j=0}^{G} \mathbf{s}_{j_i}^d\right\} \cup \mathbf{s}_i^y$ be the union of all the selected controls from steps 2 and 3.

5. Run a regression of y_i on d_{0_i}, d_{1_i}, ..., d_{G_i}, and $\mathbf{s}_i$.

6. Obtain parameter estimates of d_j's from step 5,

where $\mathbf{v}_i(\hat{\gamma})$ is defined as in (3.6) for $j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$.

3.3.4 Double/Debiased ML LASSO Estimation

Lastly, I briefly go over double/debiased ML LASSO defined in great detail with its theoretical properties in Chernozhukov *et al.* (2018). The double/debiased ML LASSO is the most complicated of all the LASSO estimation methods described in this section and can be seen as the crossfit version of the post partial-out LASSO. To attenuate the effect of regularization bias and overfitting bias on the parameters of interest, the double/debiased ML LASSO uses two important techniques: Neyman-orthogonal moments/scores and crossfitting. For this reason, this new estimation method gets the name "double/debiased ML." Regularization bias is handled by orthogonalization obtained by both partialling out and crossfitting, and crossfitting play a major role in overcoming bias induced by overfitting in the double/debiased ML LASSO.

Chernozhukov *et al.* (2018) defines two different double/debiased ML approaches. In remark 3.1 of that article, the authors' second approach DML2 is recommended over DML1 in many problems. For this reason, following definition 3.2 of double/debiased ML in Chernozhukov *et al.* (2018, p. C23), the double/debiased ML LASSO procedure I can prescribe to estimate the partial effects in high dimensional and sparse setting is as follows:

Procedure 3.4

1. Same as in Procedure 3.1.
2. Divide the sample of N into K equal-sized partitions randomly.
3. Let I_k be the set of observations in partition k and $I_k^C \equiv \{1, 2, \ldots, N\} \setminus I_k$ for $k = 1, 2, \ldots, K$.
4. Run a LASSO regression of d_{j_i} on $d_{0_i}\mathbf{x}_i$, $d_{1_i}\mathbf{x}_i$, $\ldots$, $d_{G_i}\mathbf{x}_i$, $d_{0_i}\mathbf{h}_i$, $d_{1_i}\mathbf{h}_i$, $\ldots$, $d_{G_i}\mathbf{h}_i$, and $\mathbf{v}_i(\hat{\gamma})$, where only $d_{j_i}\mathbf{h}_i'$s are to be selected, and $d_{j_i}\mathbf{x}_i'$s and $\mathbf{v}_i(\hat{\gamma})$ are forced to be included in the selected controls. Let's denote the selected controls by $\mathbf{s}_{j_i}^{d,k}$ for $j = 0, 1, \ldots, G$, $i \in I_k^C$, and $k = 1, 2, \ldots, K$.
5. Run a regression of d_{j_i} on $\mathbf{s}_{j_i}^{d,k}$, and let the parameter estimates from this regression be $\hat{\varsigma}_j^{d,k}$ for $j = 0, 1, \ldots, G$, $i \in I_k^C$, and $k = 1, 2, \ldots, K$.
6. Construct the residuals $\hat{d}_{j_i} = \left(d_{j_i} - \mathbf{s}_{j_i}^{d,k}\hat{\varsigma}_j^{d,k}\right)$ for $j = 0, 1, \ldots, G$, $i \in I_k$, and $k = 1, 2, \ldots, K$.
7. Let $\hat{d}_i = (\hat{d}_{0_i}, \hat{d}_{1_i}, \ldots, \hat{d}_{G_i})$ be the collection of all the residuals from step 6 for $i \in I_k$ and $k = 1, 2, \ldots, K$.
8. Run a LASSO regression of y_i on $d_{0_i}\mathbf{x}_i$, $d_{1_i}\mathbf{x}_i$, $\ldots$, $d_{G_i}\mathbf{x}_i$, $d_{0_i}\mathbf{h}_i$, $d_{1_i}\mathbf{h}_i$, $\ldots$, $d_{G_i}\mathbf{h}_i$, and $\mathbf{v}_i(\hat{\gamma})$, where only $d_{j_i}\mathbf{h}_i'$s are to be selected, and $d_{j_i}\mathbf{x}_i'$s and $\mathbf{v}_i(\hat{\gamma})$ are forced to be included in the selected controls. Let's denote the selected controls by $\mathbf{s}_i^{y,k}$ for $i \in I_k^C$ and $k = 1, 2, \ldots, K$.

9. Run a regression of y_i on $\mathbf{s}_i^{y,k}$, and let the parameter estimates from this regression be $\hat{\varsigma}^{y,k}$ for $i \in I_k^C$ and $k = 1, 2, \ldots, K$.

10. Construct the residuals $\hat{y}_i = \left(y_i - \mathbf{s}_i^{y,k}\hat{\varsigma}^{y,k}\right)$ for $i \in I_k$ and $k = 1, 2, \ldots, K$.

11. Run a regression of $\hat{y}_i$ on $\hat{d}_i$ for $i = 1, 2, \ldots, N$.

12. Obtain parameter estimates of $\hat{d}$ from step 11,

where $\mathbf{v}_i(\hat{\gamma})$ is defined as in (3.6). In a similar fashion portrayed in section 1.4 of Chapter 1, ATEs in this chapter take the following form:

$$\begin{aligned} ATE_{g,0} &= E(y_g - y_0) \\ &= E\left(\alpha_g + \mathbf{x}\beta_g + \mathbf{h}\delta_g + u_g - (\alpha_0 + \mathbf{x}\beta_0 + \mathbf{h}\delta_0 + u_0)\right) \\ &= (\alpha_g - \alpha_0) + (E(\mathbf{x}))(\beta_g - \beta_0) + (E(\mathbf{h}))(\delta_g - \delta_0), \end{aligned} \tag{3.8}$$

where the third equality uses $E(u_g) = 0$ for $g = 1, 2, \ldots, G$. Then, a consistent estimator of $ATE_{g,0}$ is

$$\widehat{ATE_{g,0}} = (\hat{\alpha}_g - \hat{\alpha}_0) + \bar{\mathbf{x}}(\hat{\beta}_g - \hat{\beta}_0) + \bar{\mathbf{h}}(\hat{\delta}_g - \hat{\delta}_0), \tag{3.9}$$

where $\hat{\alpha}_g$, $\hat{\alpha}_0$, $\hat{\beta}_g$, $\hat{\beta}_0$, $\hat{\delta}_g$, $\hat{\delta}_0$, $\bar{\mathbf{x}}= N^{-1}\sum_{n=1}^{N}\mathbf{x}_i$, and $\bar{\mathbf{h}}= N^{-1}\sum_{n=1}^{N}\mathbf{h}_i$ are respectively consistent estimates for α_g, α_0, β_g, β_0, δ_g, δ_0, $E(\mathbf{x})$, and $E(\mathbf{h})$.

One might think that, once the partial effects in high dimensional and sparse setting are estimated using either one of Procedures $3.1 - 3.4$, ATEs can easily be obtained by plugging the parameter estimates $\hat{\alpha}_g$, $\hat{\alpha}_0$, $\hat{\beta}_g$, $\hat{\beta}_0$, $\hat{\delta}_g$, and $\hat{\delta}_0$ into (3.9). However, this is not true due to the fact that Procedures $3.1 - 3.4$ all provide debiased (and consistent) estimates only for α_g for $g = 0, 1, \ldots, G$. The estimates for both β_g and δ_g also need to debiased before using them in (3.9) in order to overcome regularization biases on them owing to estimating high

dimensional parameters by ML methods and to remove overfitting bias on them resulting from using all observations while estimating high dimensional parameters and, afterwards, parameters of interest. Orthogonalization (either done by partialling out the effect of high dimensional variables from the variables of interest to obtain the orthogonalized variables of interest or by doubly selecting high dimensional control variables that are useful for predicting the variables of interest and the outcome variable or by both partialling out and double selection) is used to get rid of regularization bias, and sample splitting to overcome overfitting bias. Since $\mathbf{x}$ is low dimensional, the construction of the Neyman orthogonal scores associated with $\mathbf{x}$ and sample splitting can be done, see, e.g., Example 2.1. in Chernozhukov *et al.* (2018). However, when it comes to estimating δ_g debiasedly, there is a caveat: it is high dimensional. And to the best of my knowledge, there has been no research that provides results regarding consistent estimation of high dimensional parameters δ_g. Reasonably, when consistent estimation of δ_g is not possible, one cannot make inference about these high dimensional parameters either. On the other hand, inference results on low dimensional parameters exist, see, e.g., Belloni, Chernozhukov, and Hansen (2014a, Theorem 2) and Chernozhukov *et al.* (2018, Theorem 3.1 and Corollary 3.1). After all this discussion, there are two cases in which ATEs can still be consistently estimated. First, when $E(\mathbf{h}) = \mathbf{0}$ (or when both $E(\mathbf{x}) = \mathbf{0}$ and $E(\mathbf{h}) = \mathbf{0}$, which is used in my simulations), $ATE_{g,0}$ simplifies to $ATE_{g,0} = (\alpha_g - \alpha_0) + (E(\mathbf{x}))(\beta_g - \beta_0)$ (or $ATE_{g,0} = (\alpha_g - \alpha_0)$ when both $E(\mathbf{x}) = \mathbf{0}$ and $E(\mathbf{h}) = \mathbf{0}$) with one of its consistent estimates given by $\widehat{ATE_{g,0}} = (\hat{\alpha}_g - \hat{\alpha}_0) + \bar{\mathbf{x}}(\hat{\beta}_g - \hat{\beta}_0)$ (or $\widehat{ATE_{g,0}} = (\hat{\alpha}_g - \hat{\alpha}_0)$ when both $E(\mathbf{x}) = \mathbf{0}$ and $E(\mathbf{h}) = \mathbf{0}$) for $g = 1, 2, \ldots, G$. Second, when δ_g is constant across all treatment statuses (e.g., $\delta_g = \delta$ for $g = 0, 1, \ldots, G$), then again $ATE_{g,0} = (\alpha_g - \alpha_0) + (E(\mathbf{x}))(\beta_g - \beta_0)$, and $\widehat{ATE_{g,0}}$ defined above can be used for consistent ATE estimates. Note that consistent estimation of ATEs is possible in these two cases because ATEs do not depend on high dimensional parameters δ_g.

3.4 Simulations

In this section, I report Monte Carlo simulation results that aim to compare and contrast mainly the finite sample performances of the estimation methods from section 3.3 (i.e., LASSO, post partial-out LASSO, post double selection LASSO, double/debiased ML LASSO) and the CF method from Chapter 1. With respect to the measures of performance comparison, I make use of bias of ATE estimates, standard deviation of ATE estimates, mean absolute prediction error (MAPE), root mean square error (RMSE), mean number of correctly selected variables (CSVs), and mean size of selected set of variables (SVs).

My simulation results come from the model introduced in section 3.2. However, I will change the data generating process slightly for this model as I change the correlation structure among the high dimensional variables in $\mathbf{h}$ and the sparsity condition (e.g., the number of variables associated with nonzero parameters in $\mathbf{h}$). For the sake of computational simplicity, I adopt a scheme in which there is only one instrument in the latent variable equation (i.e., $\mathbf{z} = z$), there is only one exogenous low dimensional (and distinct from z) variable in the counterfactual outcome equation (i.e., $\mathbf{x} = x$), and the treatment variable w takes on only three values as always. In all simulations, the treatment is evenly spread among different treatment statuses (i.e., each treatment status has at least about 30 percent of the sample size).

3.4.1 Data Generating Process

In my simulation analysis, I used four different data generating processes (DGPs): one for the model in section 3.3 with correlated variables in $\mathbf{h} = (h_1, h_2, \ldots, h_{695})$ and moderate sparsity (i.e., $l_{h_1} = 4$), one for the model with correlated variables in $\mathbf{h}$ and high sparsity (i.e., $l_{h_1} = 1$),one for the model with uncorrelated variables in $\mathbf{h}$ and moderate sparsity, and one for the model with uncorrelated variables in $\mathbf{h}$ and high sparsity. The setup for the DGP

of the model in section 3.3 with correlated variables in $\mathbf{h}$ and moderate sparsity (i.e., $\mathbf{h}_1$ has 4 variables: h_1, h_2, h_3, and h_4) is as follows:

$$w \in \{0,1,2\},$$

$$d_g = 1[w = g], \quad g \in \{0,1,2\},$$

$$a_g \sim Gumbel(0,1), \quad g \in \{0,1,2\},$$

$$\gamma_0 = 1, \quad \gamma_1 = 5, \quad and, \quad \gamma_2 = 9,$$

$$l_0 = 1, \quad l_1 = 4.5, \quad and, \quad l_2 = 1.5,$$

$$\mathbf{z} = z \sim N(0,2),$$

$$w_g^* = l_g + \gamma_g z + a_g, \quad g \in \{0,1,2\},$$

$$w = g \quad iff \quad w_g^* \geq w_j^*, \quad \forall j \neq g \quad and \quad g,j \in \{0,1,2\},$$

$$e_g \sim N(0,1), \quad g \in \{0,1,2\},$$

$$\eta_{0,0} = 0.05, \quad \eta_{0,1} = 0.10, \quad and \quad \eta_{0,2} = 0.15,$$

$$\eta_{1,0} = 4.05, \quad \eta_{1,1} = 4.10, \quad and \quad \eta_{1,2} = 4.15,$$

$$\eta_{2,0} = 8.05, \quad \eta_{2,1} = 8.10, \quad and \quad \eta_{2,2} = 8.15,$$

$$u_g = \textstyle\sum_{j=0}^{2} \eta_{g,j} a_j + [-\sum_{j=0}^{2} \eta_{g,j} E(a_j)] + e_g, \quad g \in \{0,1,2\},$$

$$\mathbf{x} = x \sim N(0,1),$$

$$\mathbf{h} \sim N(\mathbf{0}, \mathbf{\Sigma}) \ \ with \ \ elements \ \ \mathbf{\Sigma}_{r,c} = (0.5)^{|r-c|}, \quad r,c \in \{1,2,\ldots,695\},$$

$$\alpha_0 = 1, \quad \alpha_1 = 2, \quad and, \quad \alpha_2 = 3,$$

$$\beta_0 = 6, \quad \beta_1 = 7, \quad and, \quad \beta_2 = 8,$$

$$\delta_{0,1} = 1, \quad \delta_{1,1} = 2, \quad and, \quad \delta_{2,1} = 3,$$

$$\delta_{0,2} = 2, \quad \delta_{1,2} = 3, \quad and, \quad \delta_{2,2} = 4,$$

$$\delta_{0,3} = 3, \quad \delta_{1,3} = 4, \quad and, \quad \delta_{2,3} = 5,$$

$$\delta_{0,4} = 4, \quad \delta_{1,4} = 5, \quad and, \quad \delta_{2,4} = 6,$$

$$y_g = \alpha_g + x\beta_g + h_1\delta_{g,1} + h_2\delta_{g,2} + h_3\delta_{g,3} + h_4\delta_{g,4} + u_g, \quad g \in \{0, 1, 2\},$$

$$and \quad y = d_0 y_0 + d_1 y_1 + d_2 y_2.$$

For the model with correlated variables in $\mathbf{h}$ and high sparsity, I reduce the number of variables associated with nonzero parameters in $\mathbf{h}$ from four to one (i.e., h_1 has nonzero coefficient only). This reduction in the dimension of $\mathbf{h}_1$ causes some of the parameters above to be set equal to zero and some of the variables above in $\mathbf{h}_1$ to disappear from y_g. Specifically, $\delta_{g,l}$ will be zero for $l = 2, 3, 4$ and $g = 0, 1, 2$; and h_2, h_3, and h_4 will be removed from y_g. In other words, the DGP of the model with correlated variables in $\mathbf{h}$ and high sparsity will be almost exactly the same as the above DGP with the exception of the following modifications:

$$\delta_{0,2} = 0, \quad \delta_{1,2} = 0, \quad and, \quad \delta_{2,2} = 0,$$

$$\delta_{0,3} = 0, \quad \delta_{1,3} = 0, \quad and, \quad \delta_{2,3} = 0,$$

$$\delta_{0,4} = 0, \quad \delta_{1,4} = 0, \quad and, \quad \delta_{2,4} = 0,$$

$$y_g = \alpha_g + x\beta_g + h_1\delta_{g,1} + u_g, \quad g \in \{0, 1, 2\}.$$

For the model with uncorrelated variables in **h** and moderate sparsity, the DGP setup is almost exactly the same as the one with correlated variables in **h** and moderate sparsity. The only difference is that the variance-covariance matrix $\mathbf{\Sigma}$ now becomes a 695×695 identity matrix, which is also the only difference between the DGPs of the model with uncorrelated variables in **h** and high sparsity and the model with correlated variables in **h** and high sparsity.

3.4.2 Simulation Results

I report my simulation results in Tables C.1 through C.16 in appendix C. My objective out of these simulation examples is to compare the finite performances of several estimation methods used in linear high dimensional sparse settings and of the CF method from Chapter 1 (as a benchmark) in the presence of discrete multivalued endogenous treatment and heterogeneous counterfactual errors. I have the expectation that, from an efficiency perspective (hopefully biaswise and predictionwise too), it is worth running at least one of the LASSO-based methods with the high dimensional variables in **h** which, due to the high dimensional setting, are more involved and more time-consuming than my simpler estimator from Chapter 1. And this expectation has roots in that some of the variables in **h** are predictive of the outcome and that **h** is irrelevant to the process of treatment choice. I present my simulation results in two parts: first, bias and efficiency outcomes and second, prediction and model selection outcomes.

In Tables C.1 through C.16, I report Monte Carlo estimates for $ATE_{h,0}$, bias in the Monte Carlo estimate for ATEs, Monte Carlo standard deviations (SDs) for $ATE_{h,0}$, the mean number of SVs in estimation, the mean number of CSVs in estimation, MAPE, and RMSE for each of the estimation methods described earlier −specifically, LASSO, post partial-out LASSO (PO), post double selection LASSO (DS), double/debiased ML LASSO (XPO), and the CF method from Chapter 1) for $h = 1, 2$. In simulations, I use different small sample

sizes $n = 1000$, $n = 1250$, $n = 1500$ and $n = 2000$ for each DGP setup. For each Monte Carlo experiment, the number of iterations is always equal to 1000 due to the time-costliness of estimating high dimensional models. I also use some trimming to remove outliers from my simulation analysis.

Since the dimension l_h of the variables in $\mathbf{h}$ that potentially (but the researcher does not know for sure which one of these potential variables really have a nonzero effect on the outcome) can have an effect on the outcome is always 695, p (the number of variables included in the second stage estimating equations except the CF method from Chapter 1 that always has 15 variables in estimation) is always 2100. On the other hand, p' (the number of variables that would be included in the second stage estimating equations of the LASSO-based inference methods of section 3.3 if the researcher knew exactly which variables have a nonzero effect on the outcome) changes depending on l_{h_1}, the number of variables in $\mathbf{h}$ that have nonzero effect on the outcome. When $l_{h_1} = 4\,(1)$ in the model, $p' = 27\,(18)$. Since I rely on a CF approach to deal with the endogeneity in the model across all methods, I also know there are 15 variables (all the binary treatment indicators, x interacted with these binary treatment indicators, and the CF terms: these are exactly all the variables that the CF method from Chapter 1 uses) that must be included in the second stage estimating equations. For this reason, the number of variables that are forced to be included in each second stage estimation (denoted by f) is 15, which practically leaves the methods to correctly select only 12 (3) variables when $l_{h_1} = 4\,(1)$ although there always exist 2085 variables to really select. Therefore, when calculating the total number of correctly selected variables reported in Tables C.1-C.16, I consider only these 12 (3) variables when $l_{h_1} = 4\,(1)$. Whereas, when calculating the total number of selected variables in estimation, I consider all the potential variables including f, totally 2100 of them.

As for the notation, in Tables C.1-C.16, $\widehat{ate_{h,0}}$ is the estimate for $ATE_{h,0}$, and $bias(\widehat{ate_{h,0}})$ is the bias in the estimate for $ATE_{h,0}$ for $h = 1, 2$. Moreover, # of SV and # of CSV stand for the number of selected variables (inclusive of f although they are not really selected) in

the second stage and the number of correctly selected variables (not inclusive of f) in the second stage, respectively. As for MAPE and RMSE, they mean mean absolute prediction error in the second stage and root mean square error in the second stage, respectively. As always, since these tables would require a considerable amount of space in the main body of the chapter, I place all simulation tables of this chapter into appendix C.

At this point, it is also important to remember the true values for $ATE_{h,0}$ for $h = 1, 2$ since I often refer them throughout this section. They are respectively as follows: $ATE_{1,0} = 1$ and $ATE_{2,0} = 2$. Besides this, the variables to be correctly selected when $l_{h_1} = 1$ is as follows: $oracle1 = (d_g h_1)$ for $g = 0, 1, 2$. When $l_{h_1} = 4$, this list grows into $oracle4 = (d_g h_m)$ for $g = 0, 1, 2$ and $m = 1, 2, 3, 4$.

3.4.2.1 Bias and Efficiency Outcomes

First, let's consider the results in the two benchmark cases: the one with correlated variables in **h** and l_{h_1} being equal to 4 and the sparser one with exactly the same configurations but l_{h_1} being equal to 1. Starting with the first benchmark case (correlated variables in **h** and l_{h_1} being equal to 4) in Table C.1, the Monte Carlo simulation results show that there are only small estimation biases for $ATE_{h,0}$ in all estimation methods for $h = 1, 2$, and no estimation method has a specific advantage over the other methods in terms of bias. This observation is also in line with my expectation due to **h** totally ignorable in the process of treatment choice and to CF terms taking care of endogeneity in all methods. For example, in Table C.1, the simulation estimates from XPO method for $ATE_{1,0}$ and $ATE_{2,0}$ are respectively 1.0574 (only about 5.7% higher than the true value) and 1.9784 (only around 1.1% lower than the true value). Similarly, the same simulation estimates from the benchmark CF method are respectively .9958 (a mere .4% lower than the true value) and 1.9666 (only about 1.7% lower than the true value). Continuing to explore the first benchmark case in Table C.1, the simulation results indicate that PO is the most efficient method with the

lowest Monte Carlo standard deviations (SDs), and the CF and XPO are the least efficient methods. $ATE_{2,0}$ estimates from LASSO-based inference methods (except XPO) are statistically significant, whereas $ATE_{1,0}$ estimates across all estimation methods are not. For instance, in Table C.1, the SD of the DS parameter estimate for $ATE_{1,0}$ ($ATE_{2,0}$), which is 1.0467 (1.9936), is .8215 (1.0332). Again in line with my expectation, the simulation findings point that the CF method produces less precise ATE estimates than do several LASSO-based inference methods that select from **h**. To give an example, the SD of the PO (LASSO) parameter estimate for $ATE_{1,0}$ is about 34% (29%) lower than that of the same CF parameter estimate. As the sample size increases from 1000 in Table C.1 to 2000 in Table C.4, SDs go down across all methods; however, there is no particular improvement on parameter biases which are still small. Even $ATE_{1,0}$ estimates (except those coming from the CF method) become statistically significant when $N = 2000$. The other patterns observed in Table C.1 are also seen in Tables C.2 through C.4, and all these patterns in Tables C.1 through C.4 outlined in this paragraph constitute the patterns of the first benchmark case.

Second, as I increase the sparsity by moving from l_{h_1} being equal to 4 to 1 in Tables C.5 through C.8, I step into the second benchmark behaviors of the methods. In terms of parameter biases, the simulation findings do not change: All the estimation methods still have finite sample biases (but small) on ATE estimates. As for efficiency, SDs of ATE estimates from the CF method go down compared to the first benchmark results, and all estimation methods (except XPO) start producing results more or less at the same statistical significance levels for each ATE estimate. In a way, SDs of ATE estimates converge to each other. I expect this convergence in SDs, and it is mainly due to running regression with less variables in **h** that really are predictive of the outcome, which implies one should not expect much efficiency gain out of utilizing LASSO-based inference methods over the CF method with only one extra predictive variable in **h**. To give an example, the SD of CF parameter estimate for $ATE_{1,0}$ in Table C.1 goes down from 1.2182 to .8194 in Table C.5 which is not much different from the SD of LASSO (PO) parameter estimate for $ATE_{1,0}$,

.8211 (.8169). On top of these observations, the patterns from the first benchmark case are also traced when $l_{h_1} = 1$, too: PO is still the most efficient method (slightly though due to convergence in SDs), the CF and XPO are still the least efficient methods (only with trifling margins this time), all $ATE_{2,0}$ estimates (except those in XPO) are statistically significant, $ATE_{1,0}$ estimates get statistically significant only when $N = 2000$, and SDs decrease across all methods when N gets bigger with no change on ATE biases which are still small. All these observations in Tables C.5 through C.8 outlined in this paragraph form the patterns of the second benchmark case.

When I alter the correlation structure of the variables in **h** from high correlation to no correlation, the simulation results of course reflect this change as seen in Tables C.9 through C.16. First, let's start summing up the findings in Tables C.9 through C.12 when the variables in **h** are uncorrelated and $l_{h_1} = 4$. As before, the simulations point that all the estimation methods have small biases on ATE estimates when N is small. As to efficiency, the simulation results show that, compared to the first benchmark case, there is very trivial increase in the SDs of ATE estimates from LASSO-based methods and a decrease in the SDs of ATE estimates from the CF method. These changes in SDs are against the fact that the degree of multicollinearity among the uncorrelated covariates in a regression is less than that among the uncorrelated covariates in a regression, and less multicollinearity among the regression covariates means more precision in parameter estimates. Besides these, the patterns of the first benchmark case still apply in this case with uncorrelated variables in **h** and l_{h_1} being equal to 4. This is especially true when it comes to the efficiency gains of LASSO-based methods over the CF method. For instance, the SD of DS parameter estimate $\widehat{ate_{1,0}}$ ($\widehat{ate_{2,0}}$) is about 23% (10%) lower than that of the CF estimate. When Tables C.13 through C.16 are considered with uncorrelated variables in **h** and l_{h_1} being equal to 1, the simulation findings again reveal the existence of small ATE biases resulting from small sample sizes. With regard to efficiency, compared to the case with uncorrelated variables in **h** and l_{h_1} being equal to 4, there is a drop in SDs across all estimation methods, which

was not this apparent in the comparison of the models with uncorrelated variables in **h** but varying l_{h_1}. My explanation for this is that the number of variables included in regressions of Tables C.13 through C.16 are less than that of Tables C.9 through C.12, which can be seen from decreased number of SVs and results in less number of parameters to be estimated and smaller SDs of ATE estimates. For example, the SD of LASSO (XPO) parameter estimate $\widehat{ate_{1,0}}$ reduces by about 12% (6%) from .9350 (.9276) in Table C.9 to .8221 (.8753) in Table C.13. Compared to the second benchmark case, the simulations in Tables C.13 through C.16 provide more or less similar results: the patterns observed in the second benchmark case are seen in Tables C.13 through C.16, as well. Specially, SDs of ATE estimates are close to each other, so efficiency gains out of LASSO-based models are rather limited.

3.4.2.2 Prediction and Model Selection Outcomes

As in bias and efficiency outcomes, the two benchmark cases are still the same: the one with correlated variables in **h** and l_{h_1} being equal to 4 (the first benchmark) and the sparser one with exactly the same configurations but l_{h_1} being equal to 1 (the second benchmark). Let's start interpreting the simulation results from the first benchmark. In Table C.1, the simulation outcomes suggest that XPO followed by DS and PO together has the most number of both SVs and CSVs, and that LASSO selects the least number of variables. Note that since I force all the variables to be included in the CF method, there is actually nothing to select there. For this reason, the number of SVs and CSVs are always left blank for the CF method. With regard to prediction, LASSO has the lowest prediction errors, and after LASSO comes the CF method the second. XPO designed for inference is the worst predictor of outcome variable according to the simulations. For example, in Table C.1, the number of SVs from XPO (LASSO) is around 26.4 (25.9), and the number of CSVs from XPO (DS and PO) is about 11.3 (10.9). In addition, still in Table C.1, the MAPE and RMSE of LASSO (the CF method) are about 7.89 (11.53) and 12.41 (15.74) but the same figures from

XPO are more or less 13.38 and 17.84. One of the most striking features of the simulation results is that DS and PO methods are almost the same: exactly equal numbers of SVs and CSVs, and almost equal MAPE and RMSE values. As the second stage sample size increases from 1000 in Table C.1 to 2000 in Table C.4, the readers continue seeing similar prediction and model selection patterns just with more SVs and CSVs. The LASSO-based methods predict better with larger sample sizes (lower MAPE and RMSE values). On the contrary, the prediction figures of the CF method slightly worsen (higher MAPE and RMSE values). All these patterns in Tables C.1 through C.4 constitute the patterns of the first benchmark case in terms of prediction and model selection.

Based off simulation outcomes in Tables C.5 through C.8, it can be claimed that increased sparsity (l_{h_1} being equal to 1 rather than 4) results in smaller numbers of SVs and CSVs and that, compared to the first benchmark case, prediction improves in all methods. For instance, the total number of SVs and CSVs (the MAPE and RMSE values) from PO in Table C.4 are about 26.78 and 11.77 (13.29 and 17.74); in contrast, those from PO method in Table C.8 are only around 16.88 and 1.88 (10.39 and 14.36). The other prediction and model selection patterns from the first benchmark case are also seen in Tables C.5 through C.8, which collectively forms the second benchmark behaviors of the estimators in reference to prediction and model selection.

When the variables in $\mathbf{h}$ are uncorrelated as in Tables C.9 through C.12 rather than correlated, the most conspicuous observations from the simulations are that, compared to the first benchmark case, less variables are selected (and correctly selected) by all methods and that prediction gets better across all models except LASSO. To give an example, the total number of SVs and CSVs (the MAPE and RMSE values) from DS in Table C.3 are about 26.49 and 11.49 (13.29 and 17.74); in contrast, those from DS in Table C.11 are only around 24.39 and 9.39 (11.92 and 16.11). My explanation for the lesser number of variables selected is that the methods tend not to choose the variables which could have been selected just because of the presence of strong correlation among variables but in reality have zero effect

on the outcome of interest. As far as better predictions are concerned in Tables C.9 through C.12 compared to the first benchmark case, I think this is a natural consequence of estimating sparser models. The other prediction and model selection patterns resemble those from the first benchmark case. In Tables C.13 through C.16 with l_{h_1} set equal to 1 in addition to having uncorrelated variables in $\mathbf{h}$, compared to the previous case with uncorrelated variables in $\mathbf{h}$ and l_{h_1} equal to 4, the simulation outcomes show that there are even lesser numbers of SVs and CSVs across all methods and that prediction becomes better across all methods. However, compared to the second benchmark case, the simulations indicate that both model selection and prediction figures are extremely close to each other, which indirectly implies that how sparse the model is might be more influential than the correlation structure of the variables in $\mathbf{h}$ as far as model selection and prediction are concerned. And the other prediction and model selection patterns are very much like the ones from the first benchmark case.

3.5 Conclusion

In this chapter, I build on my work from Chapter 1 and take the econometric model with a discrete multivalued endogenous treatment variable and heterogeneous counterfactual errors to a linear high dimensional sparse setting where the number of parameters to be estimated is way more than the sample size available for use but the number of variables with nonzero effect on outcome is less than the sample size. Using the CF approach adopted from Chapter 1 to handle the problem of endogeneity in the model, I summarize four LASSO-based methods coupled with detailed procedures to estimate partial effects and ATEs. Three of the LASSO-based methods are XPO of Chernozhukov *et al.* (2018); DS of Belloni, Chernozhukov, and Hansen (2014a); and PO of Belloni *et al.* (2012), which are all developed for providing better inference results. The other LASSO-based method is simply LASSO itself, which is designed for predicting better. To estimate the ATEs, I also use the CF method from Chapter

1; however this method mainly functions as a benchmark in my Monte Carlo simulation analysis.

Using this detailed simulation analysis (the first simulation-based comparative analysis of the LASSO methods mentioned above in my setting), I compare and contrast the finite sample properties, model selection features, and prediction capabilities of the LASSO-based models for discrete multivalued endogenous treatments in a linear scalar outcome high dimensional sparse model with heterogeneous counterfactual errors. In my simulation analysis, I specifically make use of measures such as bias of estimates, standard deviation of estimates, MAPE, RMSE, mean number of CSVs, and mean number of SVs.

Overall, the simulation evidence suggests that none of the methods suffer from huge biases in small samples and that all methods can reliably estimate ATEs in the presence of high dimensional potential variables and under the threat of endogeneity when the finite sample size is 2000. Increased sample sizes, how sparse the model truly is, and less correlation among potential high dimensional variables all seem to have an impact on efficiency but not on finite sample bias. The most important simulation result is that, in the presence of enough extra predictive variables that are ignorable in treatment selection and are from a set of high dimensional predictors of outcome, more complicated LASSO-based methods result in efficiency gains in ATE estimates (more obvious in models with moderate sparsity) over the simpler CF method although both LASSO-based methods and the CF method perform more or less the same as far as finite sample bias is concerned. Among LASSO-based methods, the simulations indicate that PO is often the most efficient method to use.

As far as model selection goes, the simulations show that XPO followed by both DS and PO selects both the most number of potential variables to be used in estimation and correctly selects the most number of variables with true nonzero impact on outcome in estimation. As to prediction, the simulation results suggest that LASSO followed by CF has the best prediction features with the lowest MAPE and RMSE numbers among all the methods compared and that XPO has the least favorable prediction capabilities. Increase

in sample size results in more variables to be included in estimation and to be correctly selected in all methods. On the contrary, increase in sparsity and correlatedness among potential variables cause just the opposite in all methods. Bigger sample sizes and high sparsity also cause better prediction, especially in LASSO-based methods. Moreover, the strength of correlation among potential variables is not as influential as sample size and model sparsity in model selection and prediction. And lastly, the simulations reveal the convergence of DS and PO in terms of their model selection and prediction results.

Concerning further research ideas that can flower out of this chapter, all of the research ideas mentioned in the conclusion of Chapter 1 can of course be explored in the current high dimensional sparse setting, too. Apart from these research ideas though, it would be also very interesting and exciting to theoretically examine the asymptotic properties of XPO, DS, and PO methods in high dimensional sparse models that use generated regressors and the impact of these generated regressors on the asymptotic variance-covariance matrix of these estimators.

APPENDICES

APPENDIX A

APPENDIX FOR CHAPTER 1

A.1 Chapter 1: Derivations in CF Method

To find $E(y|\mathbf{d},\mathbf{x},\mathbf{z})$, note that I can write

$$\begin{aligned}
E(y|\mathbf{d},\mathbf{x},\mathbf{z}) &= E(d_0y_0 + d_1y_1 + \cdots + d_Gy_G|\mathbf{d},\mathbf{x},\mathbf{z}) \\
&= d_0E(y_0|\mathbf{d},\mathbf{x},\mathbf{z}) + d_1E(y_1|\mathbf{d},\mathbf{x},\mathbf{z}) + \cdots + d_GE(y_G|\mathbf{d},\mathbf{x},\mathbf{z}) \\
&= d_0E(\alpha_0 + \mathbf{x}\beta_0 + u_0|\mathbf{d},\mathbf{x},\mathbf{z}) + d_0E(\alpha_1 + \mathbf{x}\beta_1 + u_1|\mathbf{d},\mathbf{x},\mathbf{z}) + \cdots + \\
&\quad + d_GE(\alpha_G + \mathbf{x}\beta_G + u_G|\mathbf{d},\mathbf{x},\mathbf{z}) \\
&= d_0\alpha_0 + d_0\mathbf{x}\beta_0 + d_0E(u_0|\mathbf{d},\mathbf{x},\mathbf{z}) + d_1\alpha_1 + d_1\mathbf{x}\beta_1 + d_1E(u_1|\mathbf{d},\mathbf{x},\mathbf{z}) + \\
&\quad + \ldots + d_G\alpha_G + d_G\mathbf{x}\beta_G + d_GE(u_G|\mathbf{d},\mathbf{x},\mathbf{z}) \\
&= \sum_{j=0}^{G} d_j\alpha_j + \sum_{k=0}^{G} d_k\mathbf{x}\beta_k + \sum_{g=0}^{G} d_gE(u_g|\mathbf{d},\mathbf{x},\mathbf{z}). \qquad \text{(A.1)}
\end{aligned}$$

Next, I need to derive $E(u_g|\mathbf{d},\mathbf{x},\mathbf{z})$. Under A.1.3 and the law of iterated expectations,

$$\begin{aligned}
E(u_g|\mathbf{d},\mathbf{x},\mathbf{z}) &= E(E(u_g|\mathbf{d},\mathbf{x},\mathbf{z},\mathbf{a})|\mathbf{d},\mathbf{x},\mathbf{z}) \\
&= E(E(u_g|\mathbf{x},\mathbf{z},\mathbf{a})|\mathbf{d},\mathbf{x},\mathbf{z}) = E[\sum_{j=0}^{G}\eta_{g,j}(a_j - E(a_j))|\mathbf{d},\mathbf{x},\mathbf{z}] \\
&= \sum_{j=0}^{G}\eta_{g,j}E[(a_j - E(a_j))|\mathbf{d},\mathbf{x},\mathbf{z})]. \qquad \text{(A.2)}
\end{aligned}$$

In equations above, I use that $\mathbf{d}$ is completely determined by $\mathbf{z}$ and $\mathbf{a}$ together. Refer to section 1.2 for seeing this where the main model is described. Hence, the expectation conditional on $\mathbf{d},\mathbf{x},\mathbf{z},\mathbf{a}$ reduces down to the one conditional on $\mathbf{x},\mathbf{z},\mathbf{a}$ only.

Then, using A.1.1, mutual exclusivity of binary treatment indicators, and the fact right below for $g = 0, 1, \ldots, G$:

$$\begin{aligned} E(u_g|\mathbf{d}, \mathbf{x}, \mathbf{z}) &= d_0 E(u_g|d_0 = 1, \mathbf{x}, \mathbf{z}) + d_1 E(u_g|d_1 = 1, \mathbf{x}, \mathbf{z}) + \ldots + \\ &\quad + d_G E(u_g|d_G = 1, \mathbf{x}, \mathbf{z}), \end{aligned} \tag{A.3}$$

I can write $E(y|\mathbf{d}, \mathbf{x}, \mathbf{z})$ as follows:

$$E(y|\mathbf{d}, \mathbf{x}, \mathbf{z}) = \sum_{j=0}^{G} d_j \alpha_j + \sum_{k=0}^{G} d_k \mathbf{x} \beta_k + \sum_{g=0}^{G} d_g \sum_{j=0}^{G} \eta_{g,j} E[(a_j - E(a_j))|d_g = 1, \mathbf{x}, \mathbf{z}]. \tag{A.4}$$

To complete the derivation of the conditional expectation $E(y|\mathbf{d}, \mathbf{x}, \mathbf{z})$, all I need is to find a closed form expression for $\sum_{g=0}^{G} d_g \sum_{j=0}^{G} \eta_{g,j} E[(a_j - E(a_j))|d_g = 1, \mathbf{x}, \mathbf{z}]$. To do so, I will utilize the work of Dubin and McFadden (1984). In their paper, they used the following result:

$$E(a_j - E(a_j)|d_g = 1, \mathbf{x}, \mathbf{z}) = \begin{cases} -log(\Lambda_j) & , j = g \\ \\ \dfrac{\Lambda_g log(\Lambda_g)}{(1 - \Lambda_g)} & , j \neq g \end{cases}, \tag{A.5}$$

where $\Lambda_g = exp(\boldsymbol{z}\gamma_g)/\sum_{r=0}^{G} exp(\boldsymbol{z}\gamma_r)$, i.e., the MNL response probability for $g, j = 0, 1, \ldots, G$.

Using the above result, I have

$$\begin{aligned}
\sum_{g=0}^{G} d_g \sum_{j=0}^{G} \eta_{g,j} E[(a_j - E(a_j))|d_g = 1, \mathbf{x}, \mathbf{z}] &= \sum_{g=0}^{G} d_g \left(\sum_{j=0}^{G} \eta_{g,j} E(a_j - E(a_j)|d_g = 1, \mathbf{x}, \mathbf{z}) \right) \\
&= \sum_{g=0}^{G} d_g \left(-\eta_{g,g} log(\Lambda_g) + \sum_{h \neq g} \eta_{g,h} \frac{\Lambda_h log(\Lambda_h)}{(1-\Lambda_h)} \right) \\
&= \sum_{g=0}^{G} d_g \left(-\eta_{g,g} log(\Lambda_g) + \sum_{h \neq g} \eta_{g,h} M_h \right) \\
&= \sum_{g=0}^{G} -\eta_{g,g} d_g log(\Lambda_g) + \sum_{g=0}^{G} \left(d_g \sum_{h \neq g} \eta_{g,h} M_h \right) \\
&= \left(\sum_{g=0}^{G} -\eta_{g,g} d_g log(\Lambda_g) \right) + \\
&\quad + d_0 \left(\eta_{0,1} M_1 + \eta_{0,2} M_2 + \cdots + \eta_{0,G} M_G \right) + \\
&\quad + d_1 \left(\eta_{1,0} M_0 + \eta_{1,2} M_2 + \eta_{1,3} M_3 + \cdots + \eta_{1,G} M_G \right) \\
&\quad \vdots \\
&\quad + d_G \left(\eta_{G,0} M_0 + \eta_{G,1} M_1 + \cdots + \eta_{G,G-1} M_{G-1} \right) \\
&= \left(\sum_{g=0}^{G} - \eta_{g,g} d_g log(\Lambda_g) \right) + \sum_{g \neq 0} d_g \eta_{g,0} M_0 + \\
&\quad + \sum_{g \neq 1} d_g \eta_{g,1} M_1 + \cdots + \sum_{g \neq G} d_g \eta_{g,G} M_G, \qquad \text{(A.6)}
\end{aligned}$$

where $M_g = \Lambda_g log(\Lambda_g)/(1-\Lambda_g)$ for $g = 0, 1, \ldots, G$.

Finally, by combining (A.4) and (A.6), I can write the expectation of the observed outcome y conditional on the observed variables $(\mathbf{d}, \mathbf{x}, \mathbf{z})$ as follows:

$$\begin{aligned}
E(y|\mathbf{d}, \mathbf{x}, \mathbf{z}) &= \sum_{j=0}^{G} d_j \alpha_j + \sum_{k=0}^{G} d_k \mathbf{x} \beta_k + \\
&\quad + \sum_{g=0}^{G} [-\eta_{g,g} d_g log(\Lambda_g)] + \sum_{g \neq 0} d_g \eta_{g,0} M_0 + \sum_{g \neq 1} d_g \eta_{g,1} M_1 + \cdots + \\
&\quad + \sum_{g \neq G} d_g \eta_{g,G} M_G, \qquad \text{(A.7)}
\end{aligned}$$

where Λ_g and M_g are as in (A.5) and (A.6) for $g = 0, 1, \ldots, G$.

A.2 Chapter 1: Derivations in CF Method-A Special Case

Upon the additional assumption that $\eta_{g,j} = \eta_j$ and the results from the initial model, I can write the conditional expectation $E(y|\mathbf{d}, \mathbf{x}, \mathbf{z})$ as follows:

$$\begin{aligned}
E(y|\mathbf{d}, \mathbf{x}, \mathbf{z}) &= \sum_{l=0}^{G} d_l \alpha_l + \sum_{k=0}^{G} d_k \mathbf{x} \beta_k + \sum_{g=0}^{G} d_g \sum_{j=0}^{G} \eta_{g,j} E[(a_j - E(a_j))|d_g = 1, \mathbf{x}, \mathbf{z}] \\
&= \sum_{l=0}^{G} d_l \alpha_l + \sum_{k=0}^{G} d_k \mathbf{x} \beta_k + \sum_{g=0}^{G} d_g \left(\sum_{j=0}^{G} \eta_j E\left(a_j - E(a_j)|d_g = 1, \mathbf{x}, \mathbf{z}\right) \right) \\
&= \sum_{l=0}^{G} d_l \alpha_l + \sum_{k=0}^{G} d_k \mathbf{x} \beta_k + \sum_{g=0}^{G} d_g \left(-\eta_g log(\Lambda_g) + \sum_{h \neq g}^{G} \eta_h \frac{\Lambda_h log(\Lambda_h)}{(1 - \Lambda_h)} \right) \\
&= \sum_{l=0}^{G} d_l \alpha_l + \sum_{k=0}^{G} d_k \mathbf{x} \beta_k + \sum_{g=0}^{G} d_g \left(-\eta_g log(\Lambda_g) + \sum_{h \neq g}^{G} \eta_h M_h \right), \qquad \text{(A.8)}
\end{aligned}$$

where Λ_g and M_g are as in (A.5) and (A.6) for $g = 0, 1, \ldots, G$.

Note that I can further simplify the last sum of terms in (A.8) as follows:

$$\begin{aligned}
\sum_{g=0}^{G} d_g[-\eta_g log(\Lambda_g) + \sum_{h\neq g}^{G} \eta_h M_h] &= [\sum_{g=0}^{G} -\eta_g d_g log(\Lambda_g)] + \sum_{g=0}^{G}[d_g \sum_{h\neq g}^{G} \eta_h M_h] \\
&= [\sum_{g=0}^{G} -\eta_g d_g log(\Lambda_g)] + d_0\left(\eta_1 M_1 + \eta_2 M_2 + \cdots + \eta_G M_G\right) \\
&\quad + d_1\left(\eta_0 M_0 + \eta_2 M_2 + \eta_3 M_3 + \cdots + \eta_G M_G\right) \\
&\quad \vdots \\
&\quad + d_G\left(\eta_0 M_0 + \eta_1 M_1 + \cdots + \eta_{G-1} M_{G-1}\right) \\
&= [\sum_{g=0}^{G} - \eta_g d_g log(\Lambda_g)] + \sum_{g\neq 0} d_g \eta_0 M_0 + \sum_{g\neq 1} d_g \eta_1 M_1 + \\
&\quad + \cdots + \sum_{g\neq G} d_g \eta_G M_G \\
&= [\sum_{g=0}^{G} - \eta_g d_g log(\Lambda_g)] + ((1-d_0)\eta_0 M_0) + ((1-d_1)\eta_1 M_1) \\
&\quad + \cdots + ((1-d_G)\eta_G M_G) \\
&= [\sum_{g=0}^{G} - \eta_g d_g log(\Lambda_g)] + \sum_{g=0}^{G}(1-d_g)\eta_g M_g \\
&= \sum_{g=0}^{G} \left[(1-d_g)\eta_g M_g - \eta_g d_g log(\Lambda_g)\right] \\
&= \sum_{g=0}^{G} \eta_g \left[(1-d_g) M_g - d_g log(\Lambda_g)\right] \qquad \text{(A.9)}
\end{aligned}$$

Thus, by combining (A.8) and (A.9), I can write the expectation of the observed outcome y conditional on the observed variables $(\mathbf{d}, \mathbf{x}, \mathbf{z})$ as follows:

$$E(y|\mathbf{d}, \mathbf{x}, \mathbf{z}) = \sum_{l=0}^{G} d_l \alpha_l + \sum_{k=0}^{G} d_k \mathbf{x} \beta_k + \sum_{g=0}^{G} \eta_g \left[(1-d_g) M_g - d_g log(\Lambda_g)\right], \qquad \text{(A.10)}$$

where Λ_g and M_g are as in (A.5) and (A.6) for $g = 0, 1, \ldots, G$.

A.3 Chapter 1: Derivations in Asymptotic Normality Results of CF Estimates

Note that

$$\mathbf{s}_i^{\mathbf{F}}(\gamma) \equiv \begin{pmatrix} \frac{\partial l_i}{\partial \gamma_0}(\gamma) \\ \\ \frac{\partial l_i}{\partial \gamma_1}(\gamma) \\ \vdots \\ \frac{\partial l_i}{\partial \gamma_G}(\gamma) \end{pmatrix}$$

$$= \begin{pmatrix} [\sum_{j\neq 0} 1[w_i = j] \frac{exp(\mathbf{z}_i\gamma_j)exp(\mathbf{z}_i\gamma_0)\mathbf{z}_i'}{-(\sum_i)^2\Lambda_{j_i}}] + 1[w_i = 0]\frac{exp(\mathbf{z}_i\gamma_0)\mathbf{z}_i'\sum_i - exp(2\mathbf{z}_i\gamma_0)\mathbf{z}_i'}{(\sum_i)^2\Lambda_{0_i}} \\ \\ [\sum_{j\neq 1} 1[w_i = j] \frac{exp(\mathbf{z}_i\gamma_j)exp(\mathbf{z}_i\gamma_1)\mathbf{z}_i'}{-(\sum_i)^2\Lambda_{j_i}}] + 1[w_i = 1]\frac{exp(\mathbf{z}_i\gamma_1)\mathbf{z}_i'\sum_i - exp(2\mathbf{z}_i\gamma_1)\mathbf{z}_i'}{(\sum_i)^2\Lambda_{1_i}} \\ \vdots \\ [\sum_{j\neq G} 1[w_i = j] \frac{exp(\mathbf{z}_i\gamma_j)exp(\mathbf{z}_i\gamma_G)\mathbf{z}_i'}{-(\sum_i)^2\Lambda_{j_i}}] + 1[w_i = G]\frac{exp(\mathbf{z}_i\gamma_G)\mathbf{z}_i'\sum_i - exp(2\mathbf{z}_i\gamma_G)\mathbf{z}_i'}{(\sum_i)^2\Lambda_{G_i}} \end{pmatrix}$$

$$= \begin{pmatrix} [\sum_{j\neq 0} 1[w_i = j](-\Lambda_{0_i})\mathbf{z}_i'] + 1[w_i = 0](\mathbf{z}_i' - \Lambda_{0_i}\mathbf{z}_i') \\ \\ [\sum_{j\neq 1} 1[w_i = j](-\Lambda_{1_i})\mathbf{z}_i'] + 1[w_i = 1](\mathbf{z}_i' - \Lambda_{1_i}\mathbf{z}_i') \\ \vdots \\ [\sum_{j\neq G} 1[w_i = j](-\Lambda_{G_i})\mathbf{z}_i'] + 1[w_i = G](\mathbf{z}_i' - \Lambda_{G_i}\mathbf{z}_i') \end{pmatrix},$$

where $l_i(\gamma) = \sum_{j=0}^{G} 1[w_i = j] log\left(exp(\mathbf{z}_i\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)\right)$, $\sum_i = \sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)$, and $\Lambda_{j_i} = exp(\mathbf{z}_i\gamma_j)\ /\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)$ for $j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$. From here, I can further

simplify $\mathbf{s}_i^{\mathbf{F}}(\gamma)$ as follows:

$$\mathbf{s}_i^{\mathbf{F}}(\gamma) = \begin{pmatrix} [\sum_{j=0}^{G} 1[w_i = j](-\Lambda_{0_i})\mathbf{z}_i'] + 1[w_i = 0]\mathbf{z}_i' \\ [\sum_{j=1}^{G} 1[w_i = j](-\Lambda_{1_i})\mathbf{z}_i'] + 1[w_i = 1]\mathbf{z}_i' \\ \vdots \\ [\sum_{j=G}^{G} 1[w_i = j](-\Lambda_{G_i})\mathbf{z}_i'] + 1[w_i = G]\mathbf{z}_i' \end{pmatrix}$$

$$= \begin{pmatrix} (-\Lambda_{0_i}\mathbf{z}_i' \sum_{j=0}^{G} 1[w_i = j]) + 1[w_i = 0]\mathbf{z}_i' \\ (-\Lambda_{1_i}\mathbf{z}_i' \sum_{j=0}^{G} 1[w_i = j]) + 1[w_i = 1]\mathbf{z}_i' \\ \vdots \\ (-\Lambda_{G_i}\mathbf{z}_i' \sum_{j=0}^{G} 1[w_i = j]) + 1[w_i = G]\mathbf{z}_i' \end{pmatrix}$$

$$= \begin{pmatrix} 1[w_i = 0] - \Lambda_{0_i} \\ 1[w_i = 1] - \Lambda_{1_i} \\ \vdots \\ 1[w_i = G] - \Lambda_{G_i} \end{pmatrix} \bigotimes \mathbf{z}_i', \tag{A.11}$$

where $\Lambda_{j_i} = exp(\mathbf{z}_i\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)$ for $j = 0, 1, \ldots, G$.

Then, using the law of iterated expectations and that w_i follows a multinomial logit reduced form under $\gamma = \gamma_o$, I have

$$\begin{aligned} E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)] &= E(E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)|\mathbf{z}_i]) \\ &= E\left(\begin{pmatrix} E(1[w_i = 0]|\mathbf{z}_i) - \Lambda_{0_i} \\ E(1[w_i = 1]|\mathbf{z}_i) - \Lambda_{1_i} \\ \vdots \\ E(1[w_i = G]|\mathbf{z}_i) - \Lambda_{G_i} \end{pmatrix} \bigotimes \mathbf{z}_i' \right) \\ &= E\left(\begin{pmatrix} \Lambda_{0_i} - \Lambda_{0_i} \\ \Lambda_{1_i} - \Lambda_{1_i} \\ \vdots \\ \Lambda_{G_i} - \Lambda_{G_i} \end{pmatrix} \bigotimes \mathbf{z}_i' \right) = \mathbf{0}. \end{aligned} \tag{A.12}$$

Having showed that $E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)] = \mathbf{0}$, I will now show $\mathbf{A}_o^{\mathbf{F}} = \mathbf{B}_o^{\mathbf{F}}$. Since $E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)] = \mathbf{0}$, $\mathbf{B}_o^{\mathbf{F}} \equiv Var[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)] = E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)\mathbf{s}_i^{\mathbf{F}'}(\gamma_o)]$. Using (A.11) and the definition of binary treatment indicator d_g for $g = 0, 1, \ldots, G$, I have the symmetric $(G+1)k \times (G+1)k$ matrix

$$\mathbf{s}_i^{\mathbf{F}}(\gamma)\mathbf{s}_i^{\mathbf{F}'}(\gamma) = \begin{pmatrix} (d_{0_i} - \Lambda_{0_i})^2 \mathbf{z}_i'\mathbf{z}_i & (d_{0_i} - \Lambda_{0_i})(d_{1_i} - \Lambda_{1_i})\mathbf{z}_i'\mathbf{z}_i & \cdots & (d_{0_i} - \Lambda_{0_i})(d_{G_i} - \Lambda_{G_i})\mathbf{z}_i'\mathbf{z}_i \\ (d_{1_i} - \Lambda_{1_i})(d_{0_i} - \Lambda_{0_i})\mathbf{z}_i'\mathbf{z}_i & (d_{1_i} - \Lambda_{1_i})^2 \mathbf{z}_i'\mathbf{z}_i & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ (d_{G_i} - \Lambda_{G_i})(d_{0_i} - \Lambda_{0_i})\mathbf{z}_i'\mathbf{z}_i & \cdots & \cdots & (d_{G_i} - \Lambda_{G_i})^2 \mathbf{z}_i'\mathbf{z}_i \end{pmatrix}, \tag{A.13}$$

where $d_{g_i} = 1[w_i = g]$ for $g = 0, 1, \ldots, G$.

Now consider $\mathbf{H}_i^{\mathbf{F}}(\gamma) \equiv \nabla_\gamma \mathbf{s}_i^{\mathbf{F}}(\gamma)$,

$$\nabla_\gamma \mathbf{s}_i^{\mathbf{F}}(\gamma) = \begin{pmatrix} \frac{-e^{(\mathbf{z}_i\gamma_0)}\sum_i \mathbf{z}_i'\mathbf{z}_i + e^{(2\mathbf{z}_i\gamma_0)}\mathbf{z}_i'\mathbf{z}_i}{(\sum_i)^2} & \frac{e^{(\mathbf{z}_i\gamma_0)}e^{(\mathbf{z}_i\gamma_1)}\mathbf{z}_i'\mathbf{z}_i}{(\sum_i)^2} & \cdots & \frac{e^{(\mathbf{z}_i\gamma_0)}e^{(\mathbf{z}_i\gamma_G)}\mathbf{z}_i'\mathbf{z}_i}{(\sum_i)^2} \\ \frac{e^{(\mathbf{z}_i\gamma_1)}e^{(\mathbf{z}_i\gamma_0)}\mathbf{z}_i'\mathbf{z}_i}{(\sum_i)^2} & \frac{-e^{(\mathbf{z}_i\gamma_1)}\sum_i \mathbf{z}_i'\mathbf{z}_i + e^{(2\mathbf{z}_i\gamma_1)}\mathbf{z}_i'\mathbf{z}_i}{(\sum_i)^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{e^{(\mathbf{z}_i\gamma_G)}e^{(\mathbf{z}_i\gamma_0)}\mathbf{z}_i'\mathbf{z}_i}{(\sum_i)^2} & \cdots & \cdots & \frac{-e^{(\mathbf{z}_i\gamma_G)}\sum_i \mathbf{z}_i'\mathbf{z}_i + e^{(2\mathbf{z}_i\gamma_G)}\mathbf{z}_i'\mathbf{z}_i}{(\sum_i)^2} \end{pmatrix}$$

$$= \begin{pmatrix} -\Lambda_{0_i} + (\Lambda_{0_i})^2 & \Lambda_{0_i}\Lambda_{1_i} & \cdots & \Lambda_{0_i}\Lambda_{G_i} \\ \Lambda_{1_i}\Lambda_{0_i} & -\Lambda_{1_i} + (\Lambda_{1_i})^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{G_i}\Lambda_{0_i} & \cdots & \cdots & -\Lambda_{G_i} + (\Lambda_{G_i})^2 \end{pmatrix} \bigotimes \mathbf{z}_i'\mathbf{z}_i, \tag{A.14}$$

where $\sum_i = \sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)$ and $\Lambda_{j_i} = exp(\mathbf{z}_i\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)$ for $j = 0, 1, \ldots, G$ and $i = 1, 2, \ldots, N$. As you can see, $\nabla_\gamma \mathbf{s}_i^{\mathbf{F}}(\gamma)$ is a symmetric $(G+1)k \times (G+1)k$ matrix. Note that, using the law of iterated expectations and mutual exclusivity of binary treatment indicators, under $\gamma = \gamma_o$ I can write

$$\begin{aligned} E[(d_{g_i} - \Lambda_{g_i})^2 \mathbf{z}_i'\mathbf{z}_i | \mathbf{z}_i] &= E[(d_{g_i} - 2d_{g_i}\Lambda_{g_i} + (\Lambda_{g_i})^2)\mathbf{z}_i'\mathbf{z}_i | \mathbf{z}_i] \\ &= [\Lambda_{g_i} - 2(\Lambda_{g_i})^2 + (\Lambda_{g_i})^2]\mathbf{z}_i'\mathbf{z}_i \\ &= [\Lambda_{g_i} - (\Lambda_{g_i})^2]\mathbf{z}_i'\mathbf{z}_i \end{aligned} \tag{A.15}$$

for $g = 0, 1, \ldots, G$ and

$$\begin{aligned} E[(d_{g_i} - \Lambda_{g_i})(d_{h_i} - \Lambda_{h_i})\mathbf{z}_i'\mathbf{z}_i | \mathbf{z}_i] &= E[(-d_{h_i}\Lambda_{g_i} - d_{g_i}\Lambda_{h_i} + \Lambda_{g_i}\Lambda_{h_i})\mathbf{z}_i'\mathbf{z}_i | \mathbf{z}_i] \\ &= [(-\Lambda_{h_i}\Lambda_{g_i} - \Lambda_{g_i}\Lambda_{h_i} + \Lambda_{g_i}\Lambda_{h_i})\mathbf{z}_i'\mathbf{z}_i] \\ &= [-\Lambda_{h_i}\Lambda_{g_i}\mathbf{z}_i'\mathbf{z}_i], \end{aligned} \tag{A.16}$$

where $d_{g_i} = 1[w_i = g]$ and $\Lambda_{j_i} = exp(\mathbf{z}_i\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma_r)$ for $\forall h \neq g$ and $i = 1, 2, \ldots, N$. Using the results in (A.15) and (A.16), (A.13) and (A.14) together imply that $-E[\mathbf{H}_i^{\mathbf{F}}(\gamma_o)|\mathbf{z}_i] =$

$E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)\mathbf{s}_i^{\mathbf{F}'}(\gamma_o)|\mathbf{z}_i]$, which is the conditional information matrix equality (CIME). By taking the expectation of CIME with respect to $\mathbf{z}_i$ and using the law of iterated expectations, I obtain the unconditional information matrix equality (UIME), i.e., $-E[\mathbf{H}_i^{\mathbf{F}}(\gamma_o)] = E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)\mathbf{s}_i^{\mathbf{F}'}(\gamma_o)]$. Hence, $\mathbf{A}_o^{\mathbf{F}} = \mathbf{B}_o^{\mathbf{F}}$.

Now I will show the influence function representation of CMLE estimates, $\hat{\gamma}$. Assuming all the assumptions in Th.1.2 and using a mean value expansion of the first order condition $\sum_{i=1}^N \mathbf{s}_i^{\mathbf{F}}(\hat{\gamma}) = \mathbf{0}$, I can write

$$\sum_{i=1}^N \mathbf{s}_i^{\mathbf{F}}(\hat{\gamma}) = \sum_{i=1}^N \mathbf{s}_i^{\mathbf{F}}(\gamma_o) + \left(\sum_{i=1}^N \mathbf{H}_i^{\mathbf{F}}(\ddot{\gamma})\right)(\hat{\gamma} - \gamma_o), \tag{A.17}$$

where $\ddot{\gamma}$ is the $(G+1)\times 1$ vector of parameter values between $\hat{\gamma}$ and γ_o in $\Gamma \subset \mathbb{R}^{(\mathbf{G+1})\mathbf{k}}$. (A.17) and $\sum_{i=1}^N \mathbf{s}_i^{\mathbf{F}}(\hat{\gamma}) = \mathbf{0}$ together imply that

$$\mathbf{0} = N^{-1/2}\sum_{i=1}^N \mathbf{s}_i^{\mathbf{F}}(\gamma_o) + \left(N^{-1}\sum_{i=1}^N \mathbf{H}_i^{\mathbf{F}}(\ddot{\gamma})\right)\sqrt{N}(\hat{\gamma} - \gamma_o). \tag{A.18}$$

Now lets state a variant of Lemma 12.1 in Wooldridge (2010), which follows from Newey and McFadden (1994).

- **Lemma 1.1 (Lm.1.1)**: Suppose that $\hat{\gamma} \xrightarrow{p} \gamma_o$, and assume that $\mathbf{r}_i(\gamma)$ satisfies the same assumptions on $l_i(\gamma)$ in Th.1.1. Then

$$N^{-1}\sum_{i=1}^N \mathbf{r}_i(\hat{\gamma}) \xrightarrow{p} E[\mathbf{r}_i(\gamma_o)]. \tag{A.19}$$

Assuming $\mathbf{A}_o^{\mathbf{F}} \equiv -E[\mathbf{H}_i^{\mathbf{F}}(\gamma_o)]$ exists and is nonsingular, $N^{-1}\sum_{i=1}^N \mathbf{H}_i^{\mathbf{F}}(\ddot{\gamma})$ is nonsingular with probability approaching 1 and $\left(N^{-1}\sum_{i=1}^N \mathbf{H}_i^{\mathbf{F}}(\ddot{\gamma})\right)^{-1} \xrightarrow{p} (-\mathbf{A}_o^{\mathbf{F}})^{-1}$. Since $E[\mathbf{s}_i^{\mathbf{F}}(\gamma_o)] = \mathbf{0}$, and $\mathbf{s}_i^{\mathbf{F}}(\gamma_o)$ is *iid* random vectors, I can rewrite (A.18) as

$$\sqrt{N}(\hat{\gamma} - \gamma_o) = (-\mathbf{A}_o^{\mathbf{F}})^{-1}\left[-N^{-1/2}\sum_{i=1}^N \mathbf{s}_i^{\mathbf{F}}(\gamma_o)\right] + o_p(1), \tag{A.20}$$

where $o_p(1)$ (little oh p one) means that if y_n, a sequence of random variables, is $o_p(1)$ then $y_n \xrightarrow{p} 0$. From (A.20), I obtain the influence function representation of CMLE estimates as follows:

$$\sqrt{N}(\hat{\gamma} - \gamma_o) = N^{-1/2}\sum_{i=1}^{N}\mathbf{r}_i(\gamma_o) + o_p(1), \tag{A.21}$$

where $\mathbf{r}_i(\gamma_o) \equiv (\mathbf{A}_o^{\mathbf{F}})^{-1}\mathbf{s}_i^{\mathbf{F}}(\gamma_o)$.

To show $E[\mathbf{r}_i(\gamma^*)] = \mathbf{0}$, we need to remember that $E[\mathbf{s}_i^{\mathbf{F}}(\gamma^*)] = \mathbf{0}$ from the first stage CMLE estimation. Then,

$$E[\mathbf{r}_i(\gamma^*)] = E[(\mathbf{A}_*^{\mathbf{F}})^{-1}\mathbf{s}_i^{\mathbf{F}}(\gamma^*)] = (\mathbf{A}_*^{\mathbf{F}})^{-1}E[\mathbf{s}_i^{\mathbf{F}}(\gamma^*)] = \mathbf{0}. \tag{A.22}$$

Now, to show $E[\mathbf{s}_i(\theta_o;\gamma^*)] = \mathbf{0}$, we assume the model for the expectation of y conditional on $\mathbb{X}, \mathbf{v}$ is correctly specified. Then,

$$\begin{aligned} E[\mathbf{s}_i(\theta_o;\gamma^*)] &= E\left(E[\mathbf{s}_i(\theta_o;\gamma^*)|\mathbb{X},\mathbf{v}]\right) \\ &= E\left(E[-\nabla'_\theta m_i(\theta_o;\gamma^*)(y_i - m_i(\theta_o;\gamma^*)|\mathbf{d},\mathbf{x},\mathbf{z}]\right) \\ &= E\left(-\nabla'_\theta m_i(\theta_o;\gamma^*)(E[y_i|\mathbf{d},\mathbf{x},\mathbf{z}] - m_i(\theta_o;\gamma^*)\right) \\ &= \mathbf{0}, \end{aligned} \tag{A.23}$$

where $m_i(\theta_o;\gamma^*) = m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i,\mathbf{x}_i,\mathbf{z}_i,\gamma^*),\theta_o)$ as in (1.17) and $E[y_i|\mathbf{d},\mathbf{x},\mathbf{z}] = m_i(\theta_o;\gamma^*)$, i.e., the conditional mean of y_i is correctly specified. Now, I will derive the closed forms of the expressions appearing in the asymptotic variance matrix of CF estimates. I start with the score function, which is

$$\begin{aligned} \mathbf{s}_i(\theta_o;\gamma^*) &= -\nabla'_\theta m_i(\theta_o;\gamma^*)(y_i - m_i(\theta_o;\gamma^*) \\ &= -\begin{pmatrix} \mathbb{X}'_i \\ \mathbf{v}'_i \end{pmatrix}(y_i - \mathbb{X}_i\delta_o - \mathbf{v}_i\lambda_o). \end{aligned} \tag{A.24}$$

Next, I continue with the expected values of the Hessian, of the outer product of the score, and of the outer product of the influence function as below:

$$\begin{aligned}
\mathbf{A}_o &= E[\nabla_\theta \left(-\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\left(y_i - m_i(\theta_o;\gamma^*)\right) + \left(-\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\nabla_\theta\left(y_i - m_i(\theta_o;\gamma^*)\right)] \\
&= E\left\{E[\nabla_\theta \left(-\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\left(y_i - m_i(\theta_o;\gamma^*)\right)|\mathbf{d},\mathbf{x},\mathbf{z}]\right\} + \\
&\quad + E\left\{E[\left(-\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\nabla_\theta\left(y_i - m_i(\theta_o;\gamma^*)\right)|\mathbf{d},\mathbf{x},\mathbf{z}]\right\} \\
&= E\left\{\nabla_\theta \left(-\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\left(E[y_i|\mathbf{d},\mathbf{x},\mathbf{z}] - m_i(\theta_o;\gamma^*)\right)\right\} + \\
&\quad + E\left\{E[\left(\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\nabla_\theta m_i(\theta_o;\gamma^*)|\mathbf{d},\mathbf{x},\mathbf{z}]\right\} \\
&= E\left\{E[\left(\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\nabla_\theta m_i(\theta_o;\gamma^*)|\mathbf{d},\mathbf{x},\mathbf{z}]\right\} \\
&= E[\nabla'_\theta m_i(\theta_o;\gamma^*)\nabla_\theta m_i(\theta_o;\gamma^*)] \\
&= E[\begin{pmatrix} \mathbb{X}'_i \\ \mathbf{v}'_i \end{pmatrix}\begin{pmatrix} \mathbb{X}_i & \mathbf{v}_i \end{pmatrix}], \qquad \text{(A.25)}
\end{aligned}$$

$$\mathbf{B}_o = E[\begin{pmatrix} \mathbb{X}'_i \\ \mathbf{v}'_i \end{pmatrix}(y_i - \mathbb{X}_i\delta_o - \mathbf{v}_i\lambda_o)^2\begin{pmatrix} \mathbb{X}_i & \mathbf{v}_i \end{pmatrix}], \qquad \text{(A.26)}$$

$$\begin{aligned}
\mathbf{R}^* &= E[(\mathbf{A}^{\mathbf{F}}_*)^{-1}\mathbf{s}^{\mathbf{F}}_i(\gamma^*)\mathbf{s}^{\mathbf{F}'}_i(\gamma^*)(\mathbf{A}^{\mathbf{F}}_*)^{-1}] \\
&= \{E[\mathbf{s}^{\mathbf{F}}_i(\gamma^*)\mathbf{s}^{\mathbf{F}'}_i(\gamma^*)]\}^{-1} \\
&= \begin{bmatrix} E[\begin{pmatrix} (d_{0_i} - \Lambda^*_{0_i})\mathbf{z}'_i \\ \\ (d_{1_i} - \Lambda^*_{1_i})\mathbf{z}'_i \\ \vdots \\ (d_{G_i} - \Lambda^*_{G_i})\mathbf{z}'_i \end{pmatrix} & ((d_{0_i} - \Lambda^*_{0_i})\mathbf{z}_i \quad (d_{1_i} - \Lambda^*_{1_i})\mathbf{z}_i \quad \cdots \quad (d_{G_i} - \Lambda^*_{G_i})\mathbf{z}_i) \ ] \end{bmatrix}^{-1}, \qquad \text{(A.27)}
\end{aligned}$$

where $\Lambda^*_{g_i} = exp(\mathbf{z}_i\gamma^*_g)/\sum_{r=0}^{G} exp(\mathbf{z}_i\gamma^*_r)$.

Next, I continue with the expected value of the expression that represents the effect of the sampling error in γ on the second stage and with the expected value of the product of the score from the second stage estimation and the influence function as follows:

$$\begin{aligned}
\mathbf{F}_o &= E[-\nabla_\gamma \left(\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\left(y_i - m_i(\theta_o;\gamma^*)\right) + \nabla'_\theta m_i(\theta_o;\gamma^*)\nabla_\gamma m_i(\theta_o;\gamma^*)] \\
&= E[-\nabla_\gamma \left(\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\left(y_i - m_i(\theta_o;\gamma^*)\right) + \nabla'_\theta m_i(\theta_o;\gamma^*)\nabla_{\mathbf{v}'} m_i(\theta_o;\gamma^*)\nabla_\gamma \mathbf{v}'_i(\gamma^*)] \\
&= E\left\{E[-\nabla_\gamma \left(\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\left(y_i - m_i(\theta_o;\gamma^*)\right)|\mathbf{d},\mathbf{x},\mathbf{z}]\right\} + \\
&\quad + E\left\{E[\nabla'_\theta m_i(\theta_o;\gamma^*)\nabla_{\mathbf{v}'} m_i(\theta_o;\gamma^*)\nabla_\gamma \mathbf{v}'_i(\gamma^*)|\mathbf{d},\mathbf{x},\mathbf{z}]\right\} \\
&= E\left\{-\nabla_\gamma \left(\nabla'_\theta m_i(\theta_o;\gamma^*)\right)\left(E[y_i|\mathbf{d},\mathbf{x},\mathbf{z}] - m_i(\theta_o;\gamma^*)\right)\right\} + \\
&\quad + E\left\{E[\nabla'_\theta m_i(\theta_o;\gamma^*)\nabla_{\mathbf{v}'} m_i(\theta_o;\gamma^*)\nabla_\gamma \mathbf{v}'_i(\gamma^*)|\mathbf{d},\mathbf{x},\mathbf{z}]\right\} \\
&= E\left\{E[\nabla'_\theta m_i(\theta_o;\gamma^*)\nabla_{\mathbf{v}'} m_i(\theta_o;\gamma^*)\nabla_\gamma \mathbf{v}'_i(\gamma^*)|\mathbf{d},\mathbf{x},\mathbf{z}]\right\} \\
&= E[\nabla'_\theta m_i(\theta_o;\gamma^*)\nabla_{\mathbf{v}'} m_i(\theta_o;\gamma^*)\nabla_\gamma \mathbf{v}'_i(\gamma^*)] \\
&= E[\begin{pmatrix} \mathbb{X}'_i \\ \mathbf{v}'_i \end{pmatrix} \lambda' \nabla_\gamma \mathbf{v}'_i(\gamma^*)] \; and \qquad \text{(A.28)}
\end{aligned}$$

$$\begin{aligned}
\mathbf{T}_o &= E[\mathbf{r}_i(\gamma^*)\mathbf{s}'_i(\theta_o;\gamma^*)] \\
&= E[(\mathbf{A}^{\mathbf{F}}_*)^{-1}\mathbf{s}^{\mathbf{F}}_i(\gamma^*)\{-\nabla_\theta m_i(\theta_o;\gamma^*)(y_i - m_i(\theta_o;\gamma^*)\}] \\
&= E\left(E[(\mathbf{A}^{\mathbf{F}}_*)^{-1}\mathbf{s}^{\mathbf{F}}_i(\gamma^*)\{-\nabla_\theta m_i(\theta_o;\gamma^*)(y_i - m_i(\theta_o;\gamma^*)\}|\mathbf{d},\mathbf{x},\mathbf{z}]\right) \\
&= E\left((\mathbf{A}^{\mathbf{F}}_*)^{-1}\mathbf{s}^{\mathbf{F}}_i(\gamma^*)\{-\nabla_\theta m_i(\theta_o;\gamma^*)(E[y_i|\mathbf{d},\mathbf{x},\mathbf{z}] - m_i(\theta_o;\gamma^*)\}\right) \\
&= \mathbf{0}. \qquad \text{(A.29)}
\end{aligned}$$

A.4 Chapter 1: Verification of the Conditions in Theorems 1.1-1.4

Conditions of Theorem 1.1. Condition (a) holds because under the model in (1.1) and (1.2), A.1.1 and A.1.2 allow the discrete treatment variable w to follow a multinomial logit model with choice probabilities $P(w = g|\mathbf{z};\gamma) = exp(\mathbf{z}\gamma_g)/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)$ for $g = 0, 1, \ldots, G$.

Hence, $\int_{\mathscr{W}} f(w|\mathbf{z})\mu(dw) = \sum_{g=0}^{G} P(w = g|\mathbf{z};\gamma) = 1$, $\forall \mathbf{z} \in \mathscr{Z}$. For condition (b), since I implicitly assume the parametric model $f(\cdot\,|\mathbf{z};\gamma)$ for the conditional density $p(\cdot\,|\mathbf{z})$ is correctly specified, then for some $\gamma_o \in \Gamma$, $p_o(\cdot\,|\mathbf{z}) = f(\cdot\,|\mathbf{z};\gamma_o)$, $\forall \mathbf{z} \in \mathscr{Z}$. Furthermore, assuming $\int_{\mathscr{W}} p_o(w|\mathbf{z})\mu(dw) \geq \int_{\mathscr{W}} f(w|\mathbf{z})\mu(dw)$,then the Kullback-Leibler information criterion in Wooldridge (2010, p. 523) implies that $E[l_i(\gamma_o)|\mathbf{z}_i] \geq E[l_i(\gamma)|\mathbf{z}_i]$, $\forall \gamma \in \Gamma$. Consequently, γ_o is a solution to $\underset{\gamma\in\Gamma}{max} E[l_i(\gamma)]$. And if γ_o is identified, then it is the unique solution to the maximization problem. As for condition (c), without any restrictions, Γ can be an open ball with center $\mathbf{0} \in \mathbb{R}^{(\mathbf{G+1})\mathbf{k}}$ and a very large radius. Since open balls are convex, Γ is an open and convex parameter space. Then, $\gamma_o \in \Gamma$ is in the interior of Γ since it is an open ball. For condition (d), $l(\cdot\,,\gamma) = \sum_{j=0}^{G} 1[w = j] log\left(exp(\mathbf{z}\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)\right)$, which is simply the sum of the products of an indicator function and a logarithmic function. Therefore, if for each $\gamma \in \Gamma$ the logarithmic function is Borel measurable on $\mathscr{Z}$, then $l(\cdot\,,\gamma)$ is a Borel measurable function on $\mathscr{W} \times \mathscr{Z}$. Since, for each $\gamma \in \Gamma$, $exp(\mathbf{z}\gamma_j)$ is an exponential function of $\mathbf{z}\gamma_j$ and $\mathbf{z}\gamma_j$ is linear in $\mathbf{z}$, $exp(\mathbf{z}\gamma_j)$ is continuous at $\mathbf{z}$. As $\mathbf{z}$ is an arbitrary element of $\mathscr{Z}$, $exp(\mathbf{z}\gamma_j)$ is continuous on $\mathscr{Z}$ for $j = 0, 1, \ldots, G$. Then, by Theorem 15.6 of Bartle (1964), the multinomial logistic function $exp(\mathbf{z}\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)$ is continuous on $\mathscr{Z}$. As logarithmic functions are continuous, by Theorem 15.8 of Bartle (1964), $log\left(exp(\mathbf{z}\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)\right)$ is continuous on $\mathscr{Z}$. Hence, by Theorem 13.2 of Billingsley (1995), for each $\gamma \in \Gamma$ $log\left(exp(\mathbf{z}\gamma_j)/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)\right)$ is Borel measurable on $\mathscr{Z}$ and thereof $l(\cdot\,,\gamma)$ is a Borel measurable function on $\mathscr{W} \times \mathscr{Z}$. As for condition (e), it suffices to show that, for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, the Hessian matrix $\mathbf{H^F}(\gamma) \equiv \nabla_\gamma \mathbf{s^F}(\gamma)$ is negative semidefinite or its negative is positive semidefinite, see Theorem 21.5 and Chapter 21 in Simon and Blume (1994) for more on concavity. Note that $-\nabla_\gamma \mathbf{s^F}(\gamma) = -\mathbf{\Lambda} \bigotimes \mathbf{z}'\mathbf{z}$ is the Kronecker product of $-\mathbf{\Lambda}$ and $\mathbf{z}'\mathbf{z}$, where $\mathbf{\Lambda}$ is a $(G+1) \times (G+1)$ symmetric matrix with $-\Lambda_i + \Lambda_i^2$'s on its diagonal and $\Lambda_m\Lambda_n$'s off the diagonal for $i, m, n = 0, 1, \ldots, G$ and $m \neq n$. $-\mathbf{\Lambda}$ is positive semidefinite because , by Equation 30 of Searle (1982), the quadratic form $\mathbf{x}\,(-\mathbf{\Lambda})\,\mathbf{x}' = \sum_{i=0}^{G}(\Lambda_i - \Lambda_i^2)x_i^2 - 2\sum_{i=0}^{G-1}\sum_{j>i}^{G} x_i x_j \Lambda_i \Lambda_j$ for all nonzero row

vector $\mathbf{x}' \in \mathbb{R}^{\mathbf{G+1}}$. After some algebraic manipulation and that $\sum_{i=0}^{G}\Lambda_i = 1$, $\mathbf{x}\left(-\mathbf{\Lambda}\right)\mathbf{x}' = \sum_{i=0}^{G}\Lambda_i\left(\sum_{j\neq i}^{G}\Lambda_j\right)x_i^2 - 2\sum_{i=0}^{G-1}\sum_{j>i}^{G}x_ix_j\Lambda_i\Lambda_j = \sum_{i=0}^{G-1}\sum_{j>i}^{G}\Lambda_i\Lambda_j(x_i^2+x_j^2) - 2x_ix_j\Lambda_i\Lambda_j$. But the summand is the square of $(\sqrt{\Lambda_i\Lambda_j}x_i - \sqrt{\Lambda_i\Lambda_j}x_j)$. Therefore, $\mathbf{x}\left(-\mathbf{\Lambda}\right)\mathbf{x}' \geq 0$, and $-\mathbf{\Lambda}$ is positive semidefinite as claimed. $\mathbf{z}'\mathbf{z}$ is also positive semidefinite because $\mathbf{x}(\mathbf{z}'\mathbf{z})\mathbf{x}' = (\mathbf{z}\mathbf{x}')'\mathbf{z}\mathbf{x}'$ which is equal to the square of $\mathbf{z}\mathbf{x}'$. By Theorem 23.17 of Simon and Blume (1994), all the eigenvalues of both $-\mathbf{\Lambda}$ and $\mathbf{z}'\mathbf{z}$ are nonnegative. Then, by Proposition 2.45 of Dhrymes (2013), all the eigenvalues of $-\nabla_\gamma\mathbf{s}^{\mathbf{F}}(\gamma)$ are nonnegative, too. Hence, by Theorem 23.17 of Simon and Blume (1994) again, $-\nabla_\gamma\mathbf{s}^{\mathbf{F}}(\gamma)$ is positive semidefinite, and thereof, for each $(w,\mathbf{z}) \in \mathscr{W}\times\mathscr{Z}$, $l(w,\mathbf{z},\cdot)$ is concave in γ. As to condition (f), consider the first, second, and mixed partial derivatives of $f(\mathbf{v}) \equiv log\left(exp(v_j)/\sum_{r=0}^{G}exp(v_r)\right)$: $\frac{\partial f(\mathbf{v})}{\partial v_j} = 1 - exp(v_j)/\sum$, $\frac{\partial f(\mathbf{v})}{\partial v_r} = -exp(v_r)/\sum$, $\frac{\partial^2 f(\mathbf{v})}{\partial v_j^2} = (exp(v_j)/\sum)(exp(v_j)/\sum - 1)$, $\frac{\partial^2 f(\mathbf{v})}{\partial v_r^2} = (exp(v_r)/\sum)(exp(v_r)/\sum - 1)$, and $\frac{\partial^2 f(\mathbf{v})}{\partial v_r\partial v_j} = \frac{\partial^2 f(\mathbf{v})}{\partial v_j\partial v_r} = (exp(v_j)/\sum)(exp(v_r)/\sum)$ where $\sum = \sum_{r=0}^{G}exp(v_r)$, $\mathbf{v} = (v_0, v_1, \ldots, v_G)' \in \mathbb{R}^{\mathbf{G+1}}$, $j \neq r$ and $j = 0, 1, \ldots, G$. Note that all these derivatives are less than one in absolute value. By the Taylor's Theorem for functions from $\mathbb{R}^{\mathbf{G+1}}$ to $\mathbb{R}$ (see Theorem 20.16 of Bartle (1964) for more on this) $f(\mathbf{v})$ can be expanded about $\mathbf{0} \in \mathbb{R}^{\mathbf{G+1}}$ such that $f(\mathbf{v}) = log(1/(G+1)) + \sum_{j=0}^{G}\frac{\partial f(\mathbf{0})}{\partial v_j}v_j + (1/2!)\sum_{j=0}^{G}\frac{\partial^2 f(\bar{\mathbf{v}})}{\partial v_j^2}v_j^2 + \sum_{j=0}^{G-1}\sum_{r>j}^{G}\frac{\partial^2 f(\bar{\mathbf{v}})}{\partial v_j\partial v_r}v_jv_r$ where $\bar{\mathbf{v}}$ is a point on the line segment between $\mathbf{v}$ and zero. Then, by the triangle inequality, $|f(\mathbf{v})| \leq |log(1/(G+1))| + \sum_{j=0}^{G}|\frac{\partial f(\mathbf{0})}{\partial v_j}||v_j| + (1/2!)\sum_{j=0}^{G}|\frac{\partial^2 f(\bar{\mathbf{v}})}{\partial v_j^2}||v_j^2| + \sum_{j=0}^{G-1}\sum_{r>j}^{G}|\frac{\partial^2 f(\bar{\mathbf{v}})}{\partial v_j\partial v_r}||v_jv_r|$. Since all the derivatives are less than one in absolute value, $|f(\mathbf{v})| \leq |log(1/(G+1))| + \sum_{j=0}^{G}|v_j| + (1/2!)\sum_{j=0}^{G}|v_j^2| + \sum_{j=0}^{G-1}\sum_{r>j}^{G}|v_jv_r|$, $\forall\mathbf{v} \in \mathbb{R}^{\mathbf{G+1}}$. Now consider $|l(w,\mathbf{z},\gamma)| \leq \sum_{j=0}^{G}|1[w=j]||log\left(exp(\mathbf{z}\gamma_j)/\sum_{r=0}^{G}exp(\mathbf{z}\gamma_r)\right)|$ by the triangle inequality. Since $|1[w=j]| \leq 1$, $|l(w,\mathbf{z},\gamma)| \leq \sum_{j=0}^{G}|log\left(exp(\mathbf{z}\gamma_j)/\sum_{r=0}^{G}exp(\mathbf{z}\gamma_r)\right)|$. Using the upper bound for $|f(\mathbf{v})|$ right above, the Cauchy–Schwarz inequality, and the triangle inequality, $|l(w,\mathbf{z},\gamma)| \leq (G+1)[|log(1/(G+1))| + \sum_{j=0}^{G} \parallel \mathbf{z}' \parallel\parallel \gamma_j \parallel + (1/2!)\sum_{j=0}^{G} \parallel \mathbf{z}' \parallel^2\parallel \gamma_j \parallel^2 + \sum_{j=0}^{G-1}\sum_{r>j}^{G}\{\sum_{i=1}^{k}\sum_{h=1}^{k}|z_iz_h||\gamma_{j_i}||\gamma_{r_h}|\}]$ where $\parallel\cdot\parallel$ is the Euclidean norm. Since $\mathit{\Gamma}$ is an open ball with a fixed (but potentially very large) radius, there exists an $\mathbf{m} \equiv (\mathbf{m}_0', \mathbf{m}_1', \ldots, \mathbf{m}_G')' \in \mathit{\Gamma}^c \subset \mathbb{R}^{\mathbf{(G+1)k}}$ where $\mathit{\Gamma}^c$ is the complement of $\mathit{\Gamma}$, $\mathbf{m}_j \in \mathbb{R}^{\mathbf{k}}$, and

$\| \mathbf{m}_j \| < \infty$, for $j = 0, 1, \ldots, G$ such that for each $\| \gamma_j \| \leq \| \mathbf{m}_j \|$. Furthermore, assuming $E|z_i z_h| < \infty$ for $i, h = 1, 2, \ldots, k$ implies that $E \| \mathbf{z}' \|^2 < \infty$ and $E \| \mathbf{z}' \| < \infty$ by Jensen's inequality. Then, setting $b(w, \mathbf{z}) \equiv (G+1)[|log(1/(G+1))| + \sum_{j=0}^{G} \| \mathbf{z}' \| \| \mathbf{m}_j \| + (1/2!) \sum_{j=0}^{G} \| \mathbf{z}' \|^2 \| \mathbf{m}_j \|^2 + \sum_{j=0}^{G-1} \sum_{r>j}^{G} \{ \sum_{i=1}^{k} \sum_{h=1}^{k} |z_i z_h| |m_{j_i}| |m_{r_h}| \}]$, $|l(w, \mathbf{z}, \gamma)| \leq b(w, \mathbf{z})$, $\forall \gamma \in \Gamma$, where $b(\cdot, \cdot)$ is a nonnegative function on $\mathscr{W} \times \mathscr{Z}$ with $E[b(w, \mathbf{z})] < \infty$.

Conditions of Theorem 1.2. Condition (a) holds because of condition (c) of Th.1.1. For condition (b), since, for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, $exp(\mathbf{z}\gamma_j)$ is an exponential function of $\mathbf{z}\gamma_j$ and $\mathbf{z}\gamma_j$ is linear in γ_j, $exp(\mathbf{z}\gamma_j)$ is differentiable at $\gamma \in int(\Gamma)$. As γ is arbitrary, $exp(\mathbf{z}\gamma_j)$ is differentiable on $int(\Gamma)$ for $j = 0, 1, \ldots, G$. Then, by Theorem 20.8 of Bartle (1964), the multinomial logistic function $exp(\mathbf{z}\gamma_j) / \sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)$ is differentiable on $int(\Gamma)$. As logarithmic functions are differentiable, by Theorem 20.9 of Bartle (1964), $log(exp(\mathbf{z}\gamma_j) / \sum_{r=0}^{G} exp(\mathbf{z}\gamma_r))$ is differentiable on $int(\Gamma)$. But $l(w, \mathbf{z}, \cdot)$ is simply the sum of the products of an indicator function and this logarithmic function; therefore, for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, $l(w, \mathbf{z}, \cdot)$ is differentiable on $int(\Gamma)$. The transpose of the first derivative of $l(w, \mathbf{z}, \cdot)$ is $\mathbf{s}^{\mathbf{F}}(\gamma)$ with elements $(1[w = j] - \Lambda_j) z_h$ where $\Lambda_j = exp(\mathbf{z}\gamma_j) / \sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)$ for $j = 0, 1, \ldots, G$ and $h = 1, 2, \ldots, k$. But Λ_j is differentiable and $1[w = j]$ is an indicator function; hence for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, each element of $\mathbf{s}^{\mathbf{F}}(\gamma)$ is differentiable on $int(\Gamma)$. For this reason, for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, $l(w, \mathbf{z}, \cdot)$ is twice differentiable on $int(\Gamma)$. The second derivative of $l(w, \mathbf{z}, \cdot)$ is $\mathbf{H}^{\mathbf{F}}(\gamma)$ with elements $(-\Lambda_j + (\Lambda_j)^2) z_i z_h$ and $\Lambda_m \Lambda_n z_i z_h$ for $j, m, n = 0, 1, \ldots, G$, $m \neq n$, and $i, h = 1, 2, \ldots, k$. Following the method used in checking condition (d) of Th.1.1, it is obvious that for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, Λ_j and $(\Lambda_j)^2$ are continuous on $int(\Gamma)$. Consequently, each element of $\mathbf{H}^{\mathbf{F}}(\gamma)$, and thereof $\mathbf{H}^{\mathbf{F}}(\gamma)$, is continuous by Theorem 15.6 of Bartle (1964). For this reason, for each $(w, \mathbf{z}) \in \mathscr{W} \times \mathscr{Z}$, $l(w, \mathbf{z}, \cdot)$ is twice continuously differentiable on $int(\Gamma)$. Condition (c) holds as shown in the previous subsection. As for Condition (d), $\nabla_\gamma [\nabla'_\gamma l(w, \mathbf{z}, \gamma)] = \mathbf{\Lambda} \bigotimes \mathbf{z}'\mathbf{z}$, where $\mathbf{\Lambda}$ is a $(G+1) \times (G+1)$ symmetric matrix with $-\Lambda_i + \Lambda_i^2$'s on its diagonal and $\Lambda_m \Lambda_n$'s off the diagonal for $i, m, n = 0, 1, \ldots, G$ and $m \neq n$. By Exercise 6(a) of Laub (2005, p. 149), $\| \mathbf{\Lambda} \bigotimes \mathbf{z}'\mathbf{z} \| = \| \mathbf{\Lambda} \| \| \mathbf{z}'\mathbf{z} \|$, where $\| \cdot \|$ is the square root

of the sum of the squares of matrix elements. Since each element of $\mathbf{\Lambda}$ is less than one in absolute value, $\| \mathbf{\Lambda} \bigotimes \mathbf{z}'\mathbf{z} \| < (G+1) \| \mathbf{z}'\mathbf{z} \|$. Furthermore, assuming $E(z_i z_h)^2 < \infty$ for $i, h = 1, 2, \ldots, k$ implies that $E \| \mathbf{z}'\mathbf{z} \| < \infty$ by Jensen's inequality. Then, setting $b(w, \mathbf{z}) \equiv (G+1) \| \mathbf{z}'\mathbf{z} \|$, the elements of $\nabla_\gamma[\nabla'_\gamma l(w, \mathbf{z}, \gamma)]$ are bounded in absolute value by $b(w, \mathbf{z})$, $\forall \gamma \in \mathit{\Gamma}$, where $b(\cdot, \cdot)$ is a nonnegative function on $\mathscr{W} \times \mathscr{Z}$ such that $E[b(w, \mathbf{z})] < \infty$. As to condition (e), note that by condition (c) $\mathbf{A}_o^{\mathbf{F}} = E(\mathbf{s}_i^{\mathbf{F}}(\gamma_o)\mathbf{s}_i^{\mathbf{F}'}(\gamma_o))$. Then, for all nonzero row vector $\mathbf{x}' \in \mathbb{R}^{(\mathbf{G+1})\mathbf{k}}$, the quadratic form $\mathbf{x}\mathbf{A}_o^{\mathbf{F}}\mathbf{x}' = E(\mathbf{x}\mathbf{s}_i^{\mathbf{F}}(\gamma_o)\mathbf{s}_i^{\mathbf{F}'}(\gamma_o)\mathbf{x}') = E[\sum_{j=0}^{G}\sum_{h=1}^{k}\{(1[w=j] - \Lambda_j)z_h x_{h,j}\}^2] = \sum_{j=0}^{G}\sum_{h=1}^{k}E[(1[w=j] - \Lambda_j)^2 z_h^2 x_{h,j}^2] = \sum_{j=0}^{G}\sum_{h=1}^{k}E[(1[w=j] - \Lambda_j)^2 z_h^2]x_{h,j}^2$. Since there is at least one $x_{h,j} \neq 0$, $\mathbf{x}\mathbf{A}_o^{\mathbf{F}}\mathbf{x}' > 0$ assuming that $E[(1[w=j] - \Lambda_j)^2 z_h^2] > 0$ for at least one j and h corresponding to that $x_{h,j}$. Hence, $\mathbf{A}_o^{\mathbf{F}}$ is positive definite.

Conditions of Theorem 1.3. Condition (a) clearly holds due to Th.1.1 where $\hat{\gamma} \xrightarrow{p} \gamma_o$ and $\gamma_o \in int(\mathit{\Gamma})$. As for condition (b), note that $q(\mathtt{w_i}, \theta; \gamma^*) \equiv (y_i - m(\mathbb{X}_i, \mathbf{v}(\mathbf{d}_i, \mathbf{z}_i, \gamma^*), \theta))^2/2 = [y_i - (\mathbb{X}_i\delta + \mathbf{v}_i\lambda)]^2/2 = [y_i - (\mathbf{h}_i\theta)]^2/2$ where $\mathbf{h}_i = (\mathbb{X}_i, \mathbf{v}_i)$ and $\theta = (\delta', \lambda')'$. Assuming conditional expectation of y_i is correctly specified, i.e., $E(y_i|\mathbf{d}, \mathbf{x}, \mathbf{z}) = (\mathbf{h}_i\theta_{\mathbf{o}})$, the conditional mean identification principle of Hayashi (2000, p. 462-3) suggests that $\underset{\theta \in \mathit{\Theta}}{min} E[q_i(\theta; \gamma^*)]$ occurs uniquely at $\theta_{\mathbf{o}}$ if $\mathbf{h}_i\theta \neq \mathbf{h}_i\theta_{\mathbf{o}}$ for all $\theta \neq \theta_{\mathbf{o}}$. However, this condition is satisfied if and only if $\mathbf{h}_i'\mathbf{h}_i$ is nonsingular. To see why, let $\mathbf{h}_i'\mathbf{h}_i$ be nonsingular and assume that $\exists\theta$ such that $\theta \neq \theta_{\mathbf{o}}$ but $\mathbf{h}_i\theta = \mathbf{h}_i\theta_{\mathbf{o}}$. Then, $\mathbf{h}_i'\mathbf{h}_i\theta = \mathbf{h}_i'\mathbf{h}_i\theta_{\mathbf{o}}$and $(\mathbf{h}_i'\mathbf{h}_i)^{-1}\mathbf{h}_i'\mathbf{h}_i\theta = (\mathbf{h}_i'\mathbf{h}_i)^{-1}\mathbf{h}_i'\mathbf{h}_i\theta_{\mathbf{o}}$, which imply $\theta = \theta_{\mathbf{o}}$. But this is a contradiction, so $\mathbf{h}_i\theta \neq \mathbf{h}_i\theta_{\mathbf{o}}$ for all $\theta \neq \theta_{\mathbf{o}}$ if $\mathbf{h}_i'\mathbf{h}_i$ is nonsingular. As a result, for any given $\gamma^* \in \mathit{\Gamma}$, the true parameter vector θ_o is the unique solution to $\underset{\theta \in \mathit{\Theta}}{min} E[q_i(\theta; \gamma^*)]$. For condition (c), let $\mathit{\Theta}$ be a closed ball with center $\mathbf{0} \in \mathbb{R}^{\mathbf{M}}$ and a very large radius $r_\theta > 0$, i.e., $\mathit{\Theta} \equiv B(\mathbf{0}, r_\theta) = \{\eta \in \mathbb{R}^{\mathbf{M}} : \| \eta \| \leq r_\theta\} \subset \mathbb{R}^{\mathbf{M}}$, and $\mathit{\Gamma}$ be another closed ball with center $\mathbf{0} \in \mathbb{R}^{(\mathbf{G+1})\mathbf{k}}$ and a very large radius $r_\gamma > 0$, i.e., $\mathit{\Gamma} \equiv B(\mathbf{0}, r_\gamma) = \{\zeta \in \mathbb{R}^{(\mathbf{G+1})\mathbf{k}} : \| \zeta \| \leq r_\gamma\} \subset \mathbb{R}^{(\mathbf{G+1})\mathbf{k}}$. Since closed balls include all their boundary points, and the distance between any two points in a closed ball cannot exceed twice its radius, both $\mathit{\Theta}$ and $\mathit{\Gamma}$ are compact parameter spaces by Theorem 9.3 of Bartle (1964). Note that $\mathit{\Theta} \times \mathit{\Gamma}$ is the Cartesian product of two compact spaces $\mathit{\Theta}$ and $\mathit{\Gamma}$. Then, by

Theorem 4.2.17 of Dixmier (1984), the parameter space $\Theta \times \Gamma$ is compact. As to condition (d), $q(\mathtt{w}, \theta; \gamma) = [y - (\mathbb{X}\delta + \mathbf{v}\lambda)]^2/2$, which is simply the square of a linear function in $\mathtt{w}$. For each $\gamma \in \Gamma$, $\mathbb{X}$ is Borel measurable because each element of $\mathbb{X}$ is either a simple binary random variable or the product of a binary random variable and $\mathbf{x}$, and thereof measurable, see the third paragraph of Billingsley (1995, p. 182) and Theorem 3.33 of Davidson (1994) for more on this. In addition, for each $\gamma \in \Gamma$, $\mathbf{v}$ is Borel measurable because each element of $\mathbf{v}$ is either the product of a binary random variable and $log(\Lambda_g)$ or the product of a binary random variable and M_g where $\Lambda_g = exp(\mathbf{z}\gamma_g)/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)$ and $M_g = \Lambda_g log(\Lambda_g)/(1 - \Lambda_g)$ for $g = 0, 1, \ldots, G$. By using the arguments in condition (d) of Th.1.1, it is clear that $log(\Lambda_g)$ and M_g are both measurable, so is $\mathbf{v}$. Since, for each $(\theta, \gamma) \in \Theta \times \Gamma$, $y - (\mathbb{X}\delta + \mathbf{v}\lambda)$ is a linear function of $\mathtt{w}$, $y - (\mathbb{X}\delta + \mathbf{v}\lambda)$ is continuous at $\mathtt{w}$. As $\mathtt{w}$ is an arbitrary element of $\mathtt{W}$, $y - (\mathbb{X}\delta + \mathbf{v}\lambda)$ is continuous on $\mathtt{W}$. As polynomial functions are continuous, by Theorems 15.6 and 15.8 of Bartle (1964), $[y - (\mathbb{X}\delta + \mathbf{v}\lambda)]^2/2$ is continuous on $\mathtt{W}$. Hence, by Theorem 13.2 of Billingsley (1995), for each $(\theta, \gamma) \in \Theta \times \Gamma$, $q(\cdot\,, \theta, \gamma)$ is a Borel measurable function on $\mathtt{W}$. For condition (e), define $t \equiv \mathbb{X}\delta + \mathbf{v}\lambda$, so $q(\mathtt{w}, \cdot; \cdot) = [y - t]^2/2$. Note that, for each $\mathtt{w} \in \mathtt{W}$, $q(\mathtt{w}, \cdot; \cdot)$ is a quadratic function of t which is linear in θ. Since quadratic functions are continuous and t is linear in θ, $q(\mathtt{w}, \cdot; \cdot)$ is continuous at θ. As θ is an arbitrary element of Θ, for each $\mathtt{w} \in \mathtt{W}$, $q(\mathtt{w}, \cdot; \cdot)$ is continuous on $\Theta \times \Gamma$. As for condition (f), note that $q(\mathtt{w}_i, \theta; \gamma) = [y_i - (\mathbb{X}_i\delta + \mathbf{v}_i\lambda)]^2/2 = [y_i - (\mathbf{h}_i\theta)]^2/2$, and $y_i = \mathbf{h}_i\theta_{\mathbf{o}} + \varepsilon_i$ where $E(\varepsilon_i|\mathbf{d}, \mathbf{x}, \mathbf{z}) = 0$. Then, $[y_i - (\mathbf{h}_i\theta)]^2/2 = [\varepsilon_i + \mathbf{h}_i(\theta_{\mathbf{o}} - \theta)]^2/2 = [\varepsilon_i^2 + \varepsilon_i\mathbf{h}_i(\theta_{\mathbf{o}} - \theta) + \{\mathbf{h}_i(\theta_{\mathbf{o}} - \theta)\}^2]/2 = [\varepsilon_i^2 + \varepsilon_i\mathbf{h}_i(\theta_{\mathbf{o}} - \theta) + (\theta_{\mathbf{o}} - \theta)'\mathbf{h}_i'\mathbf{h}_i(\theta_{\mathbf{o}} - \theta)]/2$. Since $|q(\mathtt{w}_i, \theta; \gamma)| = q(\mathtt{w}_i, \theta; \gamma)$ and $E(\varepsilon_i|\mathbf{d}, \mathbf{x}, \mathbf{z}) = 0$, $E[|q(\mathtt{w}_i, \theta; \gamma)|] = E[\{\varepsilon_i^2 + \varepsilon_i\mathbf{h}_i(\theta_{\mathbf{o}} - \theta) + (\theta_{\mathbf{o}} - \theta)'\mathbf{h}_i'\mathbf{h}_i(\theta_{\mathbf{o}} - \theta)\}/2] = \{E(\varepsilon_i^2) + (\theta_{\mathbf{o}} - \theta)'E(\mathbf{h}_i'\mathbf{h}_i)(\theta_{\mathbf{o}} - \theta)\}/2]$. Assuming $E(\mathbf{h}_i'\mathbf{h}_i)$ is positive definite and $E(\varepsilon_i^2)$ is finite, $E[|q(\mathtt{w}_i, \theta; \gamma)|] < \infty \; \forall (\theta, \gamma) \in \Theta \times \Gamma$.

Conditions of Theorem 1.4. Condition (a) holds if θ_o is not boundary point of Θ and γ^* of Γ. However, by Th.1.3, the parameter space $\Theta \times \Gamma$ is compact, so both there is a possibility that θ_o and γ^* are boundary points. Assuming Θ (Γ) is a closed ball with center $\mathbf{0} \in \mathbb{R}^{\mathbf{M}}$ ($\mathbf{0} \in \mathbb{R}^{(\mathbf{G+1})\mathbf{k}}$) and a very large radius $r_\theta > 0$ ($r_\gamma > 0$), this possibility can be very low

though. Without evidence, I accept as true that $\theta_o \in int(\Theta)$ and $\gamma^* \in int(\Gamma)$. As for condition (b), $\sqrt{N}(\hat{\gamma} - \gamma_o) \xrightarrow{d} Normal(\mathbf{0}, (\mathbf{A}_o^{\mathbf{F}})^{-1}\mathbf{B}_o^{\mathbf{F}}(\mathbf{A}_o^{\mathbf{F}})^{-1})$ in Th.1.2. Then, by setting $\gamma^* = \gamma_o$ and by Lemma 4.5 of White (2001), $\sqrt{N}(\hat{\gamma} - \gamma^*) = O_p(1)$. For condition (c), since for each $(\mathtt{w}, \gamma) \in \mathtt{W} \times \Gamma$ $[y - (\mathbb{X}\delta + \mathbf{v}\lambda)]^2/2$ is a quadratic function of $y - (\mathbb{X}\delta + \mathbf{v}\lambda)$ and $y - (\mathbb{X}\delta + \mathbf{v}\lambda)$ is linear in $\theta = (\delta', \lambda')'$, $[y - (\mathbb{X}\delta + \mathbf{v}\lambda)]^2/2$ is differentiable at $\theta \in int(\Theta)$. As θ is arbitrary, $q(\mathtt{w}, \cdot; \gamma) = [y - (\mathbb{X}\delta + \mathbf{v}\lambda)]^2/2$ is differentiable on $int(\Theta)$. The transpose of the first derivative of $q(\mathtt{w}, \cdot; \gamma)$ is $\mathbf{s}(\mathtt{w}, \cdot; \gamma) \equiv \nabla'_\theta q(\mathtt{w}, \cdot; \gamma) = -\begin{pmatrix} \mathbb{X} & \mathbf{v} \end{pmatrix}'(y - \mathbb{X}\delta - \mathbf{v}\lambda)$. But $(y - \mathbb{X}\delta - \mathbf{v}\lambda)$ is differentiable and $-\begin{pmatrix} \mathbb{X} & \mathbf{v} \end{pmatrix}'$ is just constant in θ. Hence for each $(\mathtt{w}, \gamma) \in \mathtt{W} \times \Gamma$, each element of $\mathbf{s}(\mathtt{w}, \cdot; \gamma)$ is differentiable on $int(\Theta)$, and thereof, $q(\mathtt{w}, \cdot; \gamma)$ is twice differentiable on $int(\Theta)$. The second derivative of $q(\mathtt{w}, \cdot; \gamma)$ is $\mathbf{H}(\mathtt{w}, \cdot; \gamma) \equiv \nabla_\theta \mathbf{s}(\mathtt{w}, \cdot; \gamma) = \begin{pmatrix} \mathbb{X}_i & \mathbf{v}_i \end{pmatrix}'\begin{pmatrix} \mathbb{X}_i & \mathbf{v}_i \end{pmatrix}$, which is constant in θ. Consequently, it is obvious that for each $(\mathtt{w}, \gamma) \in \mathtt{W} \times \Gamma$, $\mathbf{H}(\mathtt{w}, \cdot; \gamma)$ is continuous on $int(\Theta)$, and thereof, $q(\mathtt{w}, \cdot; \gamma)$ is twice continuously differentiable on $int(\Theta)$. As to condition (d), note that for each $\theta \in \Theta$, $\mathbf{s}(\cdot, \theta; \cdot) \equiv \nabla'_\theta q(\cdot, \theta; \cdot) = -\begin{pmatrix} \mathbb{X} & \mathbf{v} \end{pmatrix}'(y - \mathbb{X}\delta - \mathbf{v}\lambda)$. Only the variables in $\mathbf{v}$ depend on γ, so as long as $\mathbf{v}$ is differentiable on $int(\Gamma)$, $\mathbf{s}(\cdot, \theta; \cdot)$ is differentiable on $int(\Gamma)$ by Theorem 20.8 of Bartle (1964). $\mathbf{v}$ has two types of variables: $-d_g log(\Lambda_g)$ and $d_h M_g$, where d_h is a binary random variable $\Lambda_g = exp(\mathbf{z}\gamma_g)/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)$, and $M_g = \Lambda_g log(\Lambda_g)/(1 - \Lambda_g)$ for $h, g = 0, 1, \ldots, G$ and $h \neq g$. From condition (b) of Th.1.2, we know that both $\Lambda_g = exp(\mathbf{z}\gamma_g)/\sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)$ and $log(\Lambda_g)$ are differentiable on $int(\Gamma)$– note that $\mathbf{z}$ is taken as given. Then, by Theorem 20.8 of Bartle (1964), M_g is also differentiable on $int(\Gamma)$. Since each element of $\mathbf{v}$ is either $-d_g log(\Lambda_g)$ or $d_h M_g$, for each $\theta \in \Theta$, $\mathbf{v}$ and thereof $\mathbf{s}(\cdot, \theta; \cdot)$, is differentiable on $int(\Gamma)$. The derivative (gradient) of $\mathbf{s}(\cdot, \theta; \cdot)$ with respect to γ is $\nabla_\gamma \mathbf{s}(\cdot, \theta; \cdot) = -\begin{pmatrix} \mathbf{0} \\ \nabla_\gamma \mathbf{v}' \end{pmatrix}(y - \mathbb{X}\delta - \mathbf{v}\lambda) + \begin{pmatrix} \mathbb{X}' \\ \mathbf{v}' \end{pmatrix}\lambda'(\nabla_\gamma \mathbf{v}')$, where $\mathbf{0}$ is the $(l+1)(G+1) \times (G+1)k$ zero vector and $\nabla_\gamma \mathbf{v}'$ is the $(G+1)^2 \times (G+1)k$ gradient vector. $\nabla_\gamma \mathbf{v}'$ is composed of four types of variables: $-d_g(1 - \Lambda_g)\mathbf{z}$, $d_g \Lambda_h \mathbf{z}$, $\{d_g[log(\Lambda_h) + 1 - \Lambda_h]\Lambda_h \mathbf{z}\}/(1 - \Lambda_h)$, and $\{d_g[-M_h - \Lambda_h]\Lambda_i \mathbf{z}\}/(1 - \Lambda_h)$, for $h, g, i = 0, 1, \ldots, G$ and $h \neq g, i$. Using the arguments from condition (d) of Th.1.1, we know that both Λ_g and $log(\Lambda_g)$ are continuous on $int(\Gamma)$.

By Theorems 15.6 and 15.8 of Bartle (1964) and the continuity of logarithmic functions, M_g is also continuous on $int(\Gamma)$. Using the same theorems and $d_g's$ being a binary random variable, each element of both $\nabla_\gamma \mathbf{v}'$ and $\mathbf{v}$ is continuous. Therefore, it is clear that $\nabla_\gamma \mathbf{s}(\cdot,\theta;\cdot)$, is continuous on $int(\Gamma)$. As a result, for each $\theta \in \Theta$, $\mathbf{s}(\cdot,\theta;\cdot) \equiv \nabla'_\theta q(\cdot,\theta;\cdot)$ is continuously differentiable on $int(\Gamma)$. As for condition (e), note that for each $(\theta,\gamma) \in \Theta \times \Gamma$, $\mathbf{H}(\cdot,\theta;\gamma) = \begin{pmatrix} \mathbb{X} & \mathbf{v} \end{pmatrix}' \begin{pmatrix} \mathbb{X} & \mathbf{v} \end{pmatrix}$. From condition (d) of Th.1.3, we know that both $\mathbb{X}$ and $\mathbf{v}$ are Borel measurable functions on $\mathtt{W}$. Since each element of $\mathbf{H}(\cdot,\theta;\gamma)$ is one of the cross product of $\mathbb{X}$ and $\mathbf{v}$, each element of $\mathbf{H}(\cdot,\theta;\gamma)$ is measurable by Theorem 3.33 of Davidson (1994). Hence, for each $(\theta,\gamma) \in \Theta \times \Gamma$, $\mathbf{H}(\cdot,\theta;\gamma)$ is a Borel measurable function on $\mathtt{W}$. For condition (f), from condition (d) of Th.1.4, it is clear that $\mathbf{v}$ is continuous on Γ. In addition, $\mathbb{X}$ is constant in θ and γ, so $\mathbb{X}$is continuous. But each element of $\mathbf{H}(\mathtt{w},\cdot;\cdot)$ is one of the cross product of $\mathbb{X}$ and $\mathbf{v}$, so each element is continuous by Theorem 15.6 of Bartle (1964). For this reason, for each $\mathtt{w} \in \mathtt{W}$, $\mathbf{H}(\mathtt{w},\cdot;\cdot)$ is continuous on $\Theta \times \Gamma$. As to condition (g), $\parallel \mathbf{H}(\mathtt{w}_i,\theta;\gamma) \parallel^2 = \sum_{j=1}^{(l+1)(G+1)} \sum_{h=1}^{(l+1)(G+1)} (\mathbb{X}_{j_i}\mathbb{X}_{h_i})^2 + \sum_{m=1}^{(G+1)^2} \sum_{n=1}^{(G+1)^2} (v_{m_i}v_{n_i})^2 + 2\sum_{j=1}^{(l+1)(G+1)} \sum_{n=1}^{(G+1)^2} (\mathbb{X}_{j_i}v_{n_i})^2$, where $\parallel \cdot \parallel$ is the Euclidean norm, $\mathbb{X}_{j_i}$ is the j^{th} element of $\mathbb{X}_i$, and v_{m_i} is the m^{th} element of $\mathbf{v}_i$. Hence, $E[\parallel \mathbf{H}(\mathtt{w}_i,\theta;\gamma) \parallel^2] = \sum_{j=1}^{(l+1)(G+1)} \sum_{h=1}^{(l+1)(G+1)} E(\mathbb{X}_{j_i}\mathbb{X}_{h_i})^2 + \sum_{m=1}^{(G+1)^2} \sum_{n=1}^{(G+1)^2} E(v_{m_i}v_{n_i})^2 + 2\sum_{j=1}^{(l+1)(G+1)} \sum_{n=1}^{(G+1)^2} E(\mathbb{X}_{j_i}\mathbf{v}_{n_i})^2$. But $\mathbb{X}_i$ contains only binary random variables and their products with $x'_{t_i}s$. Therefore, assuming $E(x_{t_i}x_{r_i})^2 < \infty$ for $t,r = 1,2,\ldots,l$, $E(\mathbb{X}_{j_i}\mathbb{X}_{h_i})^2 < \infty$ for $j,h = 1,2,\ldots,(l+1)(G+1)$. In addition, note that $\mathbf{v}_i$ contains only the product of binary random variables with $log(\Lambda_{h_i})$ and M_{h_i} for $h = 0,1,\ldots,G$ and $i = 1,2,\ldots,N$. But, from condition (f) of Th.1.1, for each $g \in \{0,1,\ldots,G\}$, $|log(\Lambda_{g_i})|^2 \leq [|log(1/(G+1))| + \sum_{j=0}^{G} \parallel \mathbf{z}'_i \parallel \parallel \gamma_j \parallel + (1/2!)\sum_{j=0}^{G} \parallel \mathbf{z}'_i \parallel^2 \parallel \gamma_j \parallel^2 + \sum_{j=0}^{G-1} \sum_{r>j}^{G} \{\sum_{t=1}^{k} \sum_{h=1}^{k} |z_{t_i}z_{h_i}||\gamma_{j_t}||\gamma_{r_h}|\}]^2$. Therefore, assuming $E(z_{t_i}^4 z_{h_i}^4) < \infty$, $E(z_{f_i}^4 z_{t_i}^2 z_{h_i}^2) < \infty$, $E[z_{t_i}^2 z_{h_i}^2 \sqrt{(\sum_{f=1}^{k} z_{f_i}^2)}|z_{l_i}z_{m_i}|] < \infty$, $E[z_{t_i}^2 z_{h_i}^2 z_{f_i}^2 |z_{l_i}z_{m_i}|] < \infty$, and $E[z_{s_i}^2 |z_{t_i}z_{h_i}| \sqrt{(\sum_{f=1}^{k} z_{f_i}^2)}|z_{l_i}z_{m_i}|] < \infty$ for $f,h,l,m,s,t = 1,2,\ldots,k$, $E(log(\Lambda_{g_i}))^4 < \infty$. Furthermore, $M_{h_i} = [\Lambda_{h_i} log(\Lambda_{h_i})]/(1-\Lambda_{h_i})$,and the first derivative of $M_{h_i}$with respect to Λ_{h_i}, $\partial M_{h_i}/\partial \Lambda_{h_i} = (1 - \{\Lambda_{h_i} - log(\Lambda_{h_i})\})/(1-\Lambda_{h_i})^2 < 0$ since $0 < \Lambda_{h_i} < 1$. Note that

$\lim_{\Lambda_{h_i}\to 0} M_{h_i} = 0$ and $\lim_{\Lambda_{h_i}\to 1} M_{h_i} = -1$ by L'Hôpital's rule, so $|M_{h_i}| < 1$, which implies $E(M_{h_i})^4 < \infty$ and $E(M_{h_i}M_{g_i})^2 < \infty$ $\forall h, g \in \{0, 1, \ldots, G\}$. Since binary random variables are mutually exclusive, it follows that $E(v_{m_i}v_{n_i})^2 < \infty$ for $m, n = 1, 2, \ldots, (G+1)^2$. In addition, $|x_{s_i} log(\Lambda_{g_i})|^2 \leq x_{s_i}^2[|log(1/(G+1))| + \sum_{j=0}^{G} \parallel \mathbf{z}_i' \parallel\parallel \gamma_j \parallel + (1/2!)\sum_{j=0}^{G} \parallel \mathbf{z}_i' \parallel^2 \parallel \gamma_j \parallel^2 + \sum_{j=0}^{G-1}\sum_{r>j}^{G}\{\sum_{t=1}^{k}\sum_{h=1}^{k}|z_{t_i}z_{h_i}||\gamma_{j_t}||\gamma_{r_h}|\}]^2$ for $s = 1, 2, \ldots, l$ and $g = 0, 1, \ldots, G$. Therefore, assuming $E[x_{s_i}z_{t_i}z_{h_i}]^2 < \infty$, $E[x_{s_i}^2\sqrt{(\sum_{f=1}^{k} z_{f_i}^2)}|z_{t_i}z_{h_i}|] < \infty$, and $E[x_{s_i}^2 z_{f_i}^2|z_{t_i}z_{h_i}|] < \infty$ for $f, h, t = 1, 2, \ldots, k$, $E(\mathbb{X}_{j_i}\mathbf{v}_{n_i})^2 < \infty$ for $j = 1, 2, \ldots, (l+1)(G+1)$, $n = 1, 2, \ldots, (G+1)^2$, and $i = 1, 2, \ldots, N$. Since $E(\mathbb{X}_{j_i}\mathbb{X}_{h_i})^2 < \infty$, $E(v_{m_i}v_{n_i})^2 < \infty$, and $E(\mathbb{X}_{j_i}\mathbf{v}_{n_i})^2 < \infty$, $E[\parallel \mathbf{H}(\mathtt{w}_i, \theta; \gamma) \parallel] < \infty$ $\forall(\theta, \gamma) \in \Theta \times \Gamma$ by Jensen's inequality. As for condition (h), $\mathbf{A}_o \equiv E[\mathbf{H}(\mathtt{w}_i, \theta_o; \gamma^*)] = E[\mathbf{h}_i'\mathbf{h}_i]$, where $\mathbf{h}_i = (\mathbb{X}_i, \mathbf{v}_i)$. Then, for all nonzero row vector $\mathbf{u}' \in \mathbb{R}^{\mathbf{M}}$, the quadratic form $\mathbf{u}\mathbf{A}_o\mathbf{u}' = E(\mathbf{u}\mathbf{h}_i'\mathbf{h}_i\mathbf{u}') = E[\sum_{j=1}^{M}\{h_{j_i}u_j\}^2] = \sum_{j=1}^{M}E[\{h_{j_i}u_j\}^2] = \sum_{j=1}^{M}E[h_{j_i}^2]u_j^2$. Since there is at least one $u_j \neq 0$, $\mathbf{u}\mathbf{A}_o\mathbf{u}' > 0$ assuming that $E[h_{j_i}^2] > 0$ for that j^{th} variable in $\mathbf{h}_i$. Hence, $\mathbf{A}_o$ is positive definite. For condition (i), $\nabla_\gamma \mathbf{s}(\mathtt{w}, \theta; \gamma) = -\begin{pmatrix} \mathbf{0} \\ \nabla_\gamma \mathbf{v}' \end{pmatrix}(y - \mathbb{X}\delta - \mathbf{v}\lambda) + \begin{pmatrix} \mathbb{X}' \\ \mathbf{v}' \end{pmatrix}\lambda'(\nabla_\gamma \mathbf{v}')$ and $\nabla_\gamma \mathbf{v}'$ is composed of four types of variables: $-d_g(1-\Lambda_g)\mathbf{z}$, $d_g\Lambda_h\mathbf{z}$, $\{d_g[log(\Lambda_h) + 1 - \Lambda_h]\Lambda_h\mathbf{z}\}/(1-\Lambda_h)$, and $\{d_g[-M_h - \Lambda_h]\Lambda_i\mathbf{z}\}/(1-\Lambda_h)$, for $h, g, i = 0, 1, \ldots, G$ and $g, i \neq h$ as in condition (d). Using the arguments from condition (d) of Th.1.1, we know that, for each $(\theta, \gamma) \in \Theta \times \Gamma$, both Λ_g, $log(\Lambda_g)$, and M_g are all continuous on $\mathtt{W}$. Using Theorems 15.6 and 15.8 of Bartle (1964) and $d_g's$ being a binary random variable, each element of both $\nabla_\gamma \mathbf{v}'$ and $\mathbf{v}$ is continuous on $\mathtt{W}$. Therefore, it is clear that, for each $(\theta, \gamma) \in \Theta \times \Gamma$, $\nabla_\gamma \mathbf{s}(\cdot, \theta; \gamma)$, is continuous on $\mathtt{W}$. Then, by Theorem 13.2 of Billingsley (1995), for each $(\theta, \gamma) \in \Theta \times \Gamma$, $\nabla_\gamma \mathbf{s}(\cdot, \theta; \gamma)$, is a Borel measurable function on $\mathtt{W}$. Condition (j) holds because of the very same arguments utilized in condition (i): We just need to replace "for each $(\theta, \gamma) \in \Theta \times \Gamma$" by "for each $\mathtt{w} \in \mathtt{W}$" and "on $\mathtt{W}$" by "on $\Theta \times \Gamma$." As for condition (k), consider the Euclidean norm of both $\begin{pmatrix} \mathbb{X} & \mathbf{v} \end{pmatrix}'$ and $(\nabla_\gamma \mathbf{v}')$. $\parallel \begin{pmatrix} \mathbb{X} & \mathbf{v} \end{pmatrix}' \parallel^2 = \sum_{g=0}^{G} d_g^2 + \sum_{g=0}^{G}\sum_{t=1}^{l} d_g^2 x_t^2 + \sum_{g=0}^{G} d_g^2 log^2(\Lambda_g) + \sum_{h=0}^{G}\sum_{g\neq h} d_g^2 M_h^2 < (G+1) + (G+1)\sum_{t=1}^{l} x_t^2 + \sum_{g=0}^{G} log^2(\Lambda_g) + G\sum_{h=0}^{G} M_h^2 <$

$(G+1)^2+(G+1)\sum_{t=1}^{l}x_t^2+\sum_{g=0}^{G}log^2(\Lambda_g)$, where the first inequality is due to binary variables $d_g's$, and the second inequality is due to that $|M_h|<1$. Before bounding $\parallel(\nabla_\gamma \mathbf{v}')\parallel^2$, define $f(\Lambda_h)\equiv[log(\Lambda_h)+1-\Lambda_h]\Lambda_h/(1-\Lambda_h)$, so $\partial^2 f(\Lambda_h)/\partial\Lambda_h^2=[1/\Lambda_h-\Lambda_h+2log(\Lambda_h)]/(1-\Lambda_h)^3>0$ where $\partial[1/\Lambda_h-\Lambda_h+2log(\Lambda_h)]/\partial\Lambda_h=-(\Lambda_h-1)^2/\Lambda_h^2<0$ and $\lim_{\Lambda_h\to 1}[1/\Lambda_h-\Lambda_h+2log(\Lambda_h)]=0$. Moreover, note that by L'Hôpital's rule $\lim_{\Lambda_h\to 1}f(\Lambda_h)=0$ and $\lim_{\Lambda_h\to 0}f(\Lambda_h)=0$. Since $f(\Lambda_h)$ is convex on $(0,1)$ and $f(1/2)\cong -.1932$, there must exist a real number $\Lambda_{h_o}\in(0,1)$ at which $f(\Lambda_h)$ achieves its finite minimum $c\in\mathbb{R}$, which implies $|f(\Lambda_h)|<c$. Similarly, define $g(\Lambda_h)\equiv(-M_h-\Lambda_h)/(1-\Lambda_h)=-\Lambda_h[log(\Lambda_h)+1-\Lambda_h]/(1-\Lambda_h)^2$. Hence, $\partial g(\Lambda_h)/\partial\Lambda_h=[\Lambda_h log(\Lambda_h)+log(\Lambda_h)-2\Lambda_h+2]/(\Lambda_h-1)^3$. To show $\partial g(\Lambda_h)/\partial\Lambda_h>0$, consider $\partial[\Lambda_h log(\Lambda_h)+log(\Lambda_h)-2\Lambda_h+2]/\partial\Lambda_h=log(\Lambda_h)+1/\Lambda_h-1$ and $\partial^2[\Lambda_h log(\Lambda_h)+log(\Lambda_h)-2\Lambda_h+2]/\partial^2\Lambda_h=1/\Lambda_h(1-1/\Lambda_h)<0$, which implies that $\partial[\Lambda_h log(\Lambda_h)+log(\Lambda_h)-2\Lambda_h+2]/\partial\Lambda_h$ is strictly decreasing. But $\lim_{\Lambda_h\to 1}\partial[\Lambda_h log(\Lambda_h)+log(\Lambda_h)-2\Lambda_h+2]/\partial\Lambda_h=0$, so $\partial[\Lambda_h log(\Lambda_h)+log(\Lambda_h)-2\Lambda_h+2]/\partial\Lambda_h>0$ and thereof $[\Lambda_h log(\Lambda_h)+log(\Lambda_h)-2\Lambda_h+2]$ is strictly increasing. But $\lim_{\Lambda_h\to 1}[\Lambda_h log(\Lambda_h)+log(\Lambda_h)-2\Lambda_h+2]=0$, so $[\Lambda_h log(\Lambda_h)+log(\Lambda_h)-2\Lambda_h+2]<0$. This last strict inequality suggests that $\partial g(\Lambda_h)/\partial\Lambda_h>0$, and $g(\Lambda_h)$ is strictly increasing. But, by L'Hôpital's rule, $\lim_{\Lambda_h\to 1}g(\Lambda_h)=1/2$ and $\lim_{\Lambda_h\to 0}g(\Lambda_h)=0$. As a result, $|g(\Lambda_h)|<1/2$. Now consider $\parallel(\nabla_\gamma\mathbf{v}')\parallel^2=\sum_{g=0}^{G}\sum_{t=1}^{k}d_g^2(1-\Lambda_g)^2z_t^2+\sum_{h=0}^{G}\sum_{g\neq h}\sum_{t=1}^{k}d_g^2\Lambda_h^2z_t^2+\sum_{h=0}^{G}\sum_{g\neq h}\sum_{t=1}^{k}d_g^2\{[log(\Lambda_h)+1-\Lambda_h]\Lambda_h/(1-\Lambda_h)\}^2z_t^2+\sum_{h=0}^{G}\sum_{g\neq h}\sum_{i\neq h}\sum_{t=1}^{k}d_g^2[(-M_h-\Lambda_h)/(1-\Lambda_h)]^2\Lambda_i^2z_t^2<(G+1)^2\sum_{t=1}^{k}z_t^2+c^2(G+1)G\sum_{t=1}^{k}z_t^2+1/4(G+1)G^2\sum_{t=1}^{k}z_t^2=(G+1)[(G+1)+c^2G+G^2/4]\sum_{t=1}^{k}z_t^2$, where the inequality is due to that d_g is binary, $|\Lambda_g|<1$, $|f(\Lambda_h)|<c$, and $|g(\Lambda_h)|<1/2$. Now consider $\parallel\nabla_\gamma\mathbf{s}(\mathtt{w}_i,\theta;\gamma)\parallel\leq\parallel\begin{pmatrix}\mathbf{0}\\ \nabla_\gamma\mathbf{v}'\end{pmatrix}\parallel|(y-\mathbb{X}\delta-\mathbf{v}\lambda)|+\parallel\begin{pmatrix}\mathbb{X}'\\ \mathbf{v}'\end{pmatrix}\parallel\parallel\lambda'\parallel\parallel(\nabla_\gamma\mathbf{v}')\parallel$ by the triangle inequality and Definition 7.45 and Example 7.46 of Laub (2005). By using the bounds for $\parallel\begin{pmatrix}\mathbb{X}&\mathbf{v}\end{pmatrix}'\parallel^2$ and $\parallel(\nabla_\gamma\mathbf{v}')\parallel^2$, $\parallel\nabla_\gamma\mathbf{s}(\mathtt{w}_i,\theta;\gamma)\parallel<\parallel\begin{pmatrix}\mathbf{0}\\ \nabla_\gamma\mathbf{v}'\end{pmatrix}\parallel|(y-\mathbb{X}\delta-\mathbf{v}\lambda)|+\parallel\lambda'\parallel\{[(G+1)^2+(G+1)\sum_{t=1}^{l}x_t^2+\sum_{g=0}^{G}log^2(\Lambda_g)][(G+1)\{(G+1)+c^2G+G^2/4\}\sum_{t=1}^{k}z_t^2]\}^{1/2}$.

Also note that $E(y - \mathbb{X}\delta - \mathbf{v}\lambda|\mathbf{d}, \mathbf{x}, \mathbf{z}) = 0$, so $E(\| \begin{pmatrix} \mathbf{0} \\ \nabla_\gamma \mathbf{v}' \end{pmatrix} \| \, |(y - \mathbb{X}\delta - \mathbf{v}\lambda)|) = 0$ by iterated law of expectations. Then, $E[\| \nabla_\gamma \mathbf{s}(\mathtt{w}_i, \theta; \gamma) \|] < \| \lambda' \| E\{[(G+1)^2 + (G+1)\sum_{t=1}^{l} x_t^2 + \sum_{g=0}^{G} log^2(\Lambda_g)][(G+1)\{(G+1) + c^2 G + G^2/4\}\sum_{t=1}^{k} z_t^2]\}^{1/2} \leq \| \lambda' \| \{E([(G+1)^2 + (G+1)\sum_{t=1}^{l} x_t^2 + \sum_{g=0}^{G} log^2(\Lambda_g)][(G+1)\{(G+1) + c^2 G + G^2/4\}\sum_{t=1}^{k} z_t^2])\}^{1/2}$, where the last inequality is due to Jensen's inequality. Assuming $E[x_s z_t]^2 < \infty$, $E[z_l z_m z_t]^2 < \infty$, $E[z_l^2 \sqrt{(\sum_{f=1}^{k} z_f^2)} |z_m z_t|] < \infty$, and $E[z_l^2 z_f^2 |z_m z_t|] < \infty$ for $s = 1, 2, \ldots, l$, and $f, l, m, t = 1, 2, \ldots, k$, then $E[\| \nabla_\gamma \mathbf{s}(\mathtt{w}_i, \theta; \gamma) \|] < \infty \; \forall (\theta, \gamma) \in \Theta \times \Gamma$. Condition (l) holds as shown in the previous subsection.

A.5 Chapter 1: Proof of Theorem 1.5

As for (a), the LM statistic comes from the limiting distribution of $N^{-1/2}\sum_{i=1}^{N} \mathbf{s}_i(\tilde{\theta}; \hat{\gamma})$ under H_0. By Theorem 4.36 (Mean Value Theorem) of White (2001); Lm.1.1 in appendix A with $\mathbf{r}_i(\theta) \equiv \mathbf{H}_i(\theta; \hat{\gamma})$ and $l_i(\theta) = q_i(\theta; \gamma)$ in Th.1.3, and assumptions in Th.1.3; and the conditions of Th.1.4, $N^{-1/2}\sum_{i=1}^{N} \mathbf{s}_i(\tilde{\theta}; \hat{\gamma})$ can be expanded around θ_o as $N^{-1/2}\sum_{i=1}^{N} \mathbf{s}_i(\tilde{\theta}; \hat{\gamma}) = N^{-1/2}\sum_{i=1}^{N} \mathbf{s}_i(\theta_o; \hat{\gamma}) + \mathbf{A}_o\sqrt{N}(\tilde{\theta} - \theta_o) + o_p(1)$ and $N^{-1/2}\sum_{i=1}^{N} \mathbf{s}_i(\theta_o; \hat{\gamma})$ around γ_o as $N^{-1/2}\sum_{i=1}^{N} \mathbf{s}_i(\theta_o; \hat{\gamma}) = N^{-1/2}\sum_{i=1}^{N} \mathbf{s}_i(\theta_o; \gamma_o) + \mathbf{F}_o\sqrt{N}(\hat{\gamma} - \gamma_o) + o_p(1)$. Note that both $\mathbf{A}_o$ and $\mathbf{F}_o$ are bounded in probability by conditions (g) and (k) of Th.1.4. We also know that a first-order representation for $\sqrt{N}(\hat{\gamma} - \gamma_o)$ is available as in (A.21)– i.e., $\sqrt{N}(\hat{\gamma} - \gamma_o) = N^{-1/2}\sum_{i=1}^{N} \mathbf{r}_i(\gamma_o) + o_p(1)$ where $\mathbf{r}_i(\gamma_o) \equiv (\mathbf{A}_o^{\mathbf{F}})^{-1}\mathbf{s}_i^{\mathbf{F}}(\gamma_o)$ as in Th.1.2. Hence, after some algebra and by Lemma 4.6 of White (2001), $N^{-1/2}\sum_{i=1}^{N} \mathbf{s}_i(\theta_o; \hat{\gamma}) = N^{-1/2}\sum_{i=1}^{N} [\mathbf{s}_i(\theta_o; \gamma_o) + \mathbf{F}_o\mathbf{r}_i(\gamma_o)] + o_p(1)$. Then, by Lemma 3.2 of Wooldridge (2010), $N^{-1/2}\sum_{i=1}^{N} \mathbf{s}_i(\tilde{\theta}; \hat{\gamma}) = N^{-1/2}\sum_{i=1}^{N} [\mathbf{s}_i(\theta_o; \gamma_o) + \mathbf{F}_o\mathbf{r}_i(\gamma_o)] + \mathbf{A}_o\sqrt{N}(\tilde{\theta} - \theta_o) + o_p(1)$. Note that the constrained estimation in (1.35) implies $\mathbf{c}(\tilde{\theta}) \equiv 0$ and, under H_0, $\mathbf{c}(\theta_o) = 0$. By taking a mean value expansion of $\mathbf{c}(\tilde{\theta})$ around θ_o, $\mathbf{c}(\tilde{\theta}) = \mathbf{c}(\theta_o) + \mathbf{C}(\ddot{\theta})(\tilde{\theta} - \theta_o)$ where $\mathbf{C}(\theta) \equiv \nabla_\theta \mathbf{c}(\theta)$ is the $Q \times M$ gradient of $\mathbf{c}(\theta)$ with rank Q, and $\ddot{\theta}$ lies on the segment connecting $\tilde{\theta}$ and θ_o. In addition, Th.1.2, Th.1.4, and A.1.4 all

together enable $\sqrt{N}(\tilde{\alpha}-\alpha_o)$ to be bounded in probability and $\tilde{\alpha}$ to be consistent for α_o in the constrained estimation. Since $\theta=\mathbf{d}(\alpha)$, $\sqrt{N}(\tilde{\theta}-\theta_o)=\mathcal{D}(\ddot{\alpha})\sqrt{N}(\tilde{\alpha}-\alpha_o)$ by a mean value expansion of $\mathbf{d}(\tilde{\alpha})$ around α_o, where $\ddot{\alpha}$ lies on the segment connecting $\tilde{\alpha}$ and α_o. On the other hand, $\mathcal{D}(\ddot{\alpha})\sqrt{N}(\tilde{\alpha}-\alpha_o)=\mathcal{D}(\alpha_o)\sqrt{N}(\tilde{\alpha}-\alpha_o)+[\mathcal{D}(\ddot{\alpha})-\mathcal{D}(\alpha_o)]\sqrt{N}(\tilde{\alpha}-\alpha_o)=\mathcal{D}(\alpha_o)\sqrt{N}(\tilde{\alpha}-\alpha_o)+o_p(1)$ due to conditions (e) and (f) of A.1.4, $\tilde{\alpha}'s$ being consistent for α_o, $\sqrt{N}(\tilde{\alpha}-\alpha_o)'s$ being bounded in probability and Lemma 3.2 of Wooldridge (2010). Hence, $\sqrt{N}(\tilde{\theta}-\theta_o)=\mathcal{D}(\alpha_o)\sqrt{N}(\tilde{\alpha}-\alpha_o)+o_p(1)$, and $\sqrt{N}(\tilde{\theta}-\theta_o)$ is bounded in probability because of $\sqrt{N}(\tilde{\alpha}-\alpha_o)'s$ and $\mathcal{D}(\alpha_o)'s$ being bounded in probability and Lemma 3.1 of Wooldridge (2010). Moreover, by conditions (c) of A.1.4 and $\tilde{\alpha}'s$ being consistent for α_o, $\mathbf{C}(\ddot{\theta})\xrightarrow{p}\mathbf{C}(\theta_o)$. Then, under H_0, $\mathbf{0}\equiv\sqrt{N}\mathbf{c}(\tilde{\theta})=\sqrt{N}\mathbf{c}(\theta_o)+\mathbf{C}(\ddot{\theta})\sqrt{N}(\tilde{\theta}-\theta_o)=\mathbf{C}(\ddot{\theta})\sqrt{N}(\tilde{\theta}-\theta_o)=\mathbf{C}(\theta_o)\sqrt{N}(\tilde{\theta}-\theta_o)+[\mathbf{C}(\ddot{\theta})-\mathbf{C}(\theta_o)]\sqrt{N}(\tilde{\theta}-\theta_o)$, where the last term vanishes in probability due to that $\mathbf{C}(\ddot{\theta})\xrightarrow{p}\mathbf{C}(\theta_o)$, $\sqrt{N}(\tilde{\theta}-\theta_o)'s$ being bounded in probability, and Lemma 4.6 of White (2001). Therefore, $\mathbf{0}=\mathbf{C}(\theta_o)\sqrt{N}(\tilde{\theta}-\theta_o)+o_p(1)$. Since $\mathbf{0}$ is also $o_p(1)$, $\mathbf{C}(\theta_o)\sqrt{N}(\tilde{\theta}-\theta_o)=o_p(1)$. Going back to $N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i(\tilde{\theta};\hat{\gamma})$ and multiplying it by $\mathbf{C}_o\mathbf{A}_o^{-1}$, we have $\mathbf{C}_o\mathbf{A}_o^{-1}N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i(\tilde{\theta};\hat{\gamma})=\mathbf{C}_o\mathbf{A}_o^{-1}N^{-1/2}\sum_{i=1}^{N}[\mathbf{s}_i(\theta_o;\gamma_o)+\mathbf{F}_o\mathbf{r}_i(\gamma_o)]+o_p(1)$ by Lemma 3.2 of Wooldridge (2010). Note that Th.1.4 implies $E[\mathbf{s}_i(\theta_o;\gamma_o)+\mathbf{F}_o\mathbf{r}_i(\gamma_o)]=\mathbf{0}$ and $N^{-1/2}\sum_{i=1}^{N}[\mathbf{s}_i(\theta_o;\gamma_o)+\mathbf{F}_o\mathbf{r}_i(\gamma_o)]\xrightarrow{d}N(\mathbf{0},\mathbf{D}_o)$ where $\mathbf{D}_o$ is as in Th.1.4. Consequently, $\mathbf{C}_o\mathbf{A}_o^{-1}N^{-1/2}\sum_{i=1}^{N}[\mathbf{s}_i(\theta_o;\gamma_o)+\mathbf{F}_o\mathbf{r}_i(\gamma_o)]\xrightarrow{d}N(\mathbf{0},\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o')$ by Example 4.12 of White (2001). Then, by Lemma 4.7 (Asymptotic Equivalence Lemma) of White (2001), $\mathbf{C}_o\mathbf{A}_o^{-1}N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i(\tilde{\theta};\hat{\gamma})\xrightarrow{d}N(\mathbf{0},\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o')$. Now take a look at the rank of $\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o'$ which is a $Q\times Q$ matrix. Assuming that $\mathbf{D}_o$ is positive definite (so it has full rank– see Table 6.1 of Searle (1982, p.172) for more) and using Theorem A.1.3 of Greene (2012, p.1038) and the first lemma of Searle (1982, p.206), $\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}$ has full rank and is positive definite. I will prove that $\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o'$ has full rank by contradiction. Suppose $\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o'$ does not have full rank, i.e., $\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o'\mathbf{x}'=\mathbf{0}$ for some nonzero row vector $\mathbf{x}'\in\mathbb{R}^{\mathbf{Q}}$. Since $\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o'$ is symmetric, $\mathbf{x}\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o'\mathbf{x}'=0$, too. But $\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}$ is positive definite so $\mathbf{C}_o'\mathbf{x}'$ must be equal to $\mathbf{0}$, which implies

that $\mathbf{x}' = \mathbf{0}$ due to $\mathbf{C}_o's$ having full rank by A.1.4. However, this is a contradiction and then $\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o'$ has full rank, and thereof, is invertible. Hence, by spectral decomposition (see Searle (1982, p.308) for more on this) and Example 4.12 of White (2001), $[\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o']^{-1/2}\mathbf{C}_o\mathbf{A}_o^{-1}N^{-1/2}\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma}) \xrightarrow{d} N(\mathbf{0},\mathbf{I}_Q)$. Therefore, by Corollary 4.28 of White (2001), $\left(\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma})\right)' \mathbf{A}_o^{-1}\mathbf{C}_o'[\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}\mathbf{C}_o']^{-1}\mathbf{C}_o\mathbf{A}_o^{-1}\left(\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma})\right)/N \xrightarrow{d} \chi^2_Q$. As to (b), $LM_N \equiv \left(\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma})\right)' \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{C}}'[\tilde{\mathbf{C}}\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{D}}\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{C}}']^{-1}\tilde{\mathbf{C}}\tilde{\mathbf{A}}^{-1}\left(\sum_{i=1}^N \mathbf{s}_i(\tilde{\theta};\hat{\gamma})\right)/N$ as in Th.1.5. Under the assumptions made in Th.1.5 and A.1.4., Lm.1.1 in appendix A with appropriate adjustments implies that $\tilde{\mathbf{A}}^{-1} \xrightarrow{p} \mathbf{A}_o^{-1}$, $\tilde{\mathbf{C}} \xrightarrow{p} \mathbf{C}_o$, and $\tilde{\mathbf{D}} \xrightarrow{p} \mathbf{D}_o$. Then, it follows from Lemma 4.7 of White (2001) that $LM_N \xrightarrow{d} \chi^2_Q$.

A.6 Chapter 1: Ignorability of Generated Instruments in IV Estimation with $\eta_{g,j} = \eta_j$

In simple terms, generated instruments are instruments that are some functions of both some exogenous variables and some first-stage estimator of parameters. The effect of the first stage estimation on the second stage estimation has been studied fairly well, see, for instance, Newey and McFadden (1994). In the framework of GMM estimation, they show that ignoring the first stage estimation can cause inconsistent asymptotic variances for the second step estimator. Wooldridge (2010) specifically gives some sufficient conditions under which one can ignore the impact of generated instruments on the standard errors of IV estimator.

Now, let us define the population counterparts of the generated instruments used in Procedure 1.1

$$\hat{\mathbf{c}} \equiv \mathbf{f}(\mathbf{z},\hat{\gamma}) \equiv (\hat{\Lambda}_0, \cdots, \hat{\Lambda}_G, \hat{\Lambda}_0\mathbf{x}, \cdots, \hat{\Lambda}_G\mathbf{x}),$$

where $\hat{\mathbf{c}}$ is the $1\times(l+1)(G+1)$ vector of generated instruments, $\mathbf{f}$ is a known function, $\mathbf{z}$ is the $1\times k$ vector of instruments in the choice equation for w_g^*, $\hat{\gamma} = (\hat{\gamma}_0', \hat{\gamma}_1', \ldots, \hat{\gamma}_G')'$ is the $(G+1)k\times 1$

vector of $\sqrt{N}-consistent$ and asymptotically normal first stage conditional MLE (CMLE) estimates from the MNL of w_i on $\mathbf{z}_i$ for $i = 1, 2, \ldots, N$, $\hat{\Lambda}_j = exp(\mathbf{z}\hat{\gamma}_j)/\sum_{r=0}^{G} exp(\mathbf{z}\hat{\gamma}_r)$ for $j = 0, 1, \ldots, G$, and $\mathbf{x}$ is the $1 \times l$ vector of exogenous variables in y_g. According to Wooldridge (2010, p. 125), $\sqrt{N}-consistent$ $\hat{\gamma}$ and $E[\nabla_\gamma \mathbf{f}'(\mathbf{z}, \gamma)u] = \mathbf{0}$, where by A.1.3′ $u_g = u$ for $g = 0, 1, \ldots, G$ (homogeneous counterfactual errors) and thereof $u' = u$ in (15), are the two sufficient conditions of ignoring the first step estimation for inference. $\hat{\gamma}$ used in $\hat{\mathbf{c}}$ comes from the CMLE and is already $\sqrt{N}-consistent$, so I only need to see whether $E[\nabla_\gamma \mathbf{f}'(\mathbf{z}, \gamma)u]$ is equal to zero or not. Note that, since $\mathbf{z}$ contains instruments and all the variables in $\mathbf{x}$ are exogenous, I implicitly assume that $E[u|\mathbf{x}, \mathbf{z}] = 0$.

Consider the expected value of the gradient of the product of the function $\mathbf{f}'$ times u, i.e., $E[\nabla_\gamma \mathbf{f}'(\mathbf{z}, \gamma)u] = E(\ \nabla_\gamma \Lambda_0 u \ \cdots \ \nabla_\gamma \Lambda_G u \ \ \nabla_\gamma \Lambda_0 \mathbf{x} u \ \cdots \ \nabla_\gamma \Lambda_G \mathbf{x} u\ )'$. And this last expectation is equal to the following matrix

$$\begin{pmatrix}
E(\frac{e^{(\mathbf{z}\gamma_0)}\mathbf{z}' \sum -e^{(2\mathbf{z}\gamma_0)}\mathbf{z}'}{(\sum)^2} u) & E(\frac{-e^{(\mathbf{z}\gamma_0)}e^{(\mathbf{z}\gamma_1)}\mathbf{z}'}{(\sum)^2} u) & \cdots & E(\frac{-e^{(\mathbf{z}\gamma_0)}e^{(\mathbf{z}\gamma_G)}\mathbf{z}'}{(\sum)^2} u) \\
E(\frac{-e^{(\mathbf{z}\gamma_1)}e^{(\mathbf{z}\gamma_0)}\mathbf{z}'}{(\sum)^2} u) & E(\frac{e^{(\mathbf{z}\gamma_1)}\mathbf{z}' \sum -e^{(2\mathbf{z}\gamma_1)}\mathbf{z}'}{(\sum)^2} u) & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
E(\frac{-e^{(\mathbf{z}\gamma_G)}e^{(\mathbf{z}\gamma_0)}\mathbf{z}'}{(\sum)^2} u) & \cdots & \cdots & E(\frac{e^{(\mathbf{z}\gamma_G)}\mathbf{z}' \sum -e^{(2\mathbf{z}\gamma_G)}\mathbf{z}'}{(\sum)^2} u) \\
E(\mathbf{x}' \otimes \frac{e^{(\mathbf{z}\gamma_0)}\mathbf{z}' \sum -e^{(2\mathbf{z}\gamma_0)}\mathbf{z}'}{(\sum)^2} u) & E(\mathbf{x}' \otimes \frac{-e^{(\mathbf{z}\gamma_0)}e^{(\mathbf{z}\gamma_1)}\mathbf{z}'}{(\sum)^2} u) & \cdots & E(\mathbf{x}' \otimes \frac{-e^{(\mathbf{z}\gamma_0)}e^{(\mathbf{z}\gamma_G)}\mathbf{z}'}{(\sum)^2} u) \\
E(\mathbf{x}' \otimes \frac{-e^{(\mathbf{z}\gamma_1)}e^{(\mathbf{z}\gamma_0)}\mathbf{z}'}{(\sum)^2} u) & E(\mathbf{x}' \otimes \frac{e^{(\mathbf{z}\gamma_1)}\mathbf{z}' \sum -e^{(2\mathbf{z}\gamma_1)}\mathbf{z}'}{(\sum)^2} u) & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
E(\mathbf{x}' \otimes \frac{-e^{(\mathbf{z}\gamma_G)}e^{(\mathbf{z}\gamma_0)}\mathbf{z}'}{(\sum)^2} u) & \cdots & \cdots & E(\mathbf{x}' \otimes \frac{e^{(\mathbf{z}\gamma_G)}\mathbf{z}' \sum -e^{(2\mathbf{z}\gamma_G)}\mathbf{z}'}{(\sum)^2} u)
\end{pmatrix},$$

where $\sum = \sum_{r=0}^{G} exp(\mathbf{z}\gamma_r)$.

Let's take a closer look at the elements of the matrix above.

$$\begin{aligned} E(\frac{e^{(\mathbf{z}\gamma_j)}\mathbf{z}'\sum -e^{(2\mathbf{z}\gamma_j)}\mathbf{z}'}{(\sum)^2}u) &= E(E(\frac{e^{(\mathbf{z}\gamma_j)}\mathbf{z}'\sum -e^{(2\mathbf{z}\gamma_j)}\mathbf{z}'}{(\sum)^2}u|\mathbf{x},\mathbf{z})) \\ &= E(\frac{e^{(\mathbf{z}\gamma_j)}\mathbf{z}'\sum -e^{(2\mathbf{z}\gamma_j)}\mathbf{z}'}{(\sum)^2}E\left(u|\mathbf{x},\mathbf{z}\right)) \\ &= \mathbf{0} \end{aligned}$$

$$\begin{aligned} E(\frac{-e^{(\mathbf{z}\gamma_j)}e^{(\mathbf{z}\gamma_h)}\mathbf{z}'}{(\sum)^2}u) &= E(E(\frac{-e^{(\mathbf{z}\gamma_j)}e^{(\mathbf{z}\gamma_h)}\mathbf{z}'}{(\sum)^2}u|\mathbf{x},\mathbf{z})) \\ &= E(\frac{-e^{(\mathbf{z}\gamma_j)}e^{(\mathbf{z}\gamma_h)}\mathbf{z}'}{(\sum)^2}E\left(u|\mathbf{x},\mathbf{z}\right)) \\ &= \mathbf{0} \end{aligned}$$

$$\begin{aligned} E(\mathbf{x}'\otimes\frac{e^{(\mathbf{z}\gamma_j)}\mathbf{z}'\sum -e^{(2\mathbf{z}\gamma_j)}\mathbf{z}'}{(\sum)^2}u) &= E(E(\mathbf{x}'\otimes\frac{e^{(\mathbf{z}\gamma_j)}\mathbf{z}'\sum -e^{(2\mathbf{z}\gamma_j)}\mathbf{z}'}{(\sum)^2}u|\mathbf{x},\mathbf{z})) \\ &= E(\mathbf{x}'\otimes\frac{e^{(\mathbf{z}\gamma_j)}\mathbf{z}'\sum -e^{(2\mathbf{z}\gamma_j)}\mathbf{z}'}{(\sum)^2}E\left(u|\mathbf{x},\mathbf{z}\right)) \\ &= \mathbf{0} \end{aligned}$$

$$\begin{aligned} E(\mathbf{x}'\otimes\frac{-e^{(\mathbf{z}\gamma_j)}e^{(\mathbf{z}\gamma_h)}\mathbf{z}'}{(\sum)^2}u) &= E(E(\mathbf{x}'\otimes[\frac{-e^{(\mathbf{z}\gamma_j)}e^{(\mathbf{z}\gamma_h)}\mathbf{z}'}{(\sum)^2}]u|\mathbf{x},\mathbf{z})) \\ &= E(\mathbf{x}'\otimes\frac{-e^{(\mathbf{z}\gamma_j)}e^{(\mathbf{z}\gamma_h)}\mathbf{z}'}{(\sum)^2}E\left(u|\mathbf{x},\mathbf{z}\right)) \\ &= \mathbf{0} \end{aligned}$$

for $j,h = 0,1,\ldots,G$ and $\forall h \neq j$. Since each individual expectation is zero, the second sufficiency condition ($E[\nabla_\gamma \mathbf{f}'(\mathbf{z},\gamma)u] = \mathbf{0}$) is also satisfied. Hence, the impact of generated instruments in Procedure 1.1 on inference can indeed be ignored when the counterfactual errors are homogeneous.

A.7 Chapter 1: Tables-Simulations

Table A.1: Model without Correlated Random Coefficients but with Asymmetric Instrument, N=1000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0042	.2749	.2106	.2543	1.3394	.2898	.1892	.2206
$\hat{\alpha}_1$	1.9654	1.8970	.5777	.8593	1.9078	.6193	.5033	.5339
$\hat{\alpha}_2$	2.9928	1.1659	.9383	1.0154	2.9018	.9465	.7928	.8391
$\widehat{ate_{10}}$	.9612	1.9154	.6158	.8990	.5684	.8102	.6030	.6567
$\widehat{ate_{20}}$	1.9885	1.1990	.9625	1.0479	1.5623	.9544	.7986	.8472
$bias(\widehat{ate_{10}})$	-.0387				-.4315			
$bias(\widehat{ate_{20}})$	-.0114				-.4376			
$se(\hat{\alpha}_0)$			.0136	.0298			.0171	.0235
$se(\hat{\alpha}_1)$			.0809	.2563			.0414	.0467
$se(\hat{\alpha}_2)$			.0555	.0726			.0409	.0493
$se(\widehat{ate_{10}})$			.0794	.2549			.0499	.0562
$se(\widehat{ate_{20}})$			.0541	.0711			.0385	.0461

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.2: Model without Correlated Random Coefficients but with Asymmetric Instrument, N=2000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0050	.1693	.1915	.1727	1.3416	.1652	.1640	.1555
$\hat{\alpha}_1$	1.9887	1.0914	.6707	.5983	1.9041	.3701	.3906	.3738
$\hat{\alpha}_2$	3.0056	.6876	.7519	.7149	2.9052	.5850	.6161	.5917
$\widehat{ate_{10}}$	.9837	1.1045	.6967	.6213	.5624	.4597	.4846	.4605
$\widehat{ate_{20}}$	2.0006	.7082	.7757	.7345	1.5635	.5967	.6184	.5969
$bias(\widehat{ate_{10}})$	-.0162				-.4375			
$bias(\widehat{ate_{20}})$	.0006				-.4364			
$se(\hat{\alpha}_0)$			.0106	.0148			.0105	.0117
$se(\hat{\alpha}_1)$			.0945	.1298			.0238	.0231
$se(\hat{\alpha}_2)$			.0367	.0360			.0247	.0245
$se(\widehat{ate_{10}})$			.0934	.1291			.0295	.0278
$se(\widehat{ate_{20}})$			.0356	.0353			.0229	.0229

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.3: Model without Correlated Random Coefficients but with Asymmetric Instrument, N=5000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	.9997	.0925	.1139	.1066	1.3371	.0902	.1061	.0973
$\hat{\alpha}_1$	1.9890	.6551	.4360	.3783	1.9094	.2601	.2631	.2357
$\hat{\alpha}_2$	2.9955	.4536	.4936	.4539	2.8963	.3803	.3954	.3759
$\widehat{ate_{10}}$	.9892	.6617	.4515	.3937	.5722	.3132	.3260	.2904
$\widehat{ate_{20}}$	1.9957	.4630	.5054	.4661	1.5591	.3828	.3954	.3784
$bias(\widehat{ate_{10}})$	-.0107				-.4277			
$bias(\widehat{ate_{20}})$	-.0042				-.4408			
$se(\hat{\alpha}_0)$			.0045	.0059			.0052	.0047
$se(\hat{\alpha}_1)$			.0274	.0540			.0134	.0093
$se(\hat{\alpha}_2)$			.0161	.0143			.0114	.0097
$se(\widehat{ate_{10}})$			.0271	.0537			.0163	.0112
$se(\widehat{ate_{20}})$			.0157	.0141			.0105	.0091

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.4: Model without Correlated Random Coefficients but with Asymmetric Instrument, N=10000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	.9998	.0745	.0790	.0755	1.3396	.0711	.0705	.0699
$\hat{\alpha}_1$	1.9913	.5145	.2552	.2637	1.9095	.1820	.1672	.1664
$\hat{\alpha}_2$	2.9999	.3137	.3313	.3152	2.9017	.2631	.2723	.2634
$\widehat{ate_{10}}$	.9914	.5199	.2663	.2735	.5698	.2243	.2068	.2058
$\widehat{ate_{20}}$	2.0001	.3224	.3406	.3228	1.5621	.2665	.2737	.2661
$bias(\widehat{ate_{10}})$	-.0085				-.4301			
$bias(\widehat{ate_{20}})$	.0001				-.4378			
$se(\hat{\alpha}_0)$			.0019	.0029			.0026	.0023
$se(\hat{\alpha}_1)$			.0111	.0268			.0048	.0046
$se(\hat{\alpha}_2)$			.0067	.0071			.0048	.0048
$se(\widehat{ate_{10}})$			.0110	.0267			.0058	.0055
$se(\widehat{ate_{20}})$			.0065	.0070			.0045	.0045

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.5: Model without Correlated Random Coefficients but with Symmetric Instrument, N=1000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0005	.1586	.1141	.1228	.9727	.1387	.1128	.1199
$\hat{\alpha}_1$	1.9931	.3323	.2339	.2663	2.3728	.3190	.3289	.3189
$\hat{\alpha}_2$	2.9722	1.0310	.8305	.7607	2.6951	.8062	.6787	.6518
$\widehat{ate_{10}}$	.9925	.3683	.2600	.2934	1.4001	.3949	.3847	.3801
$\widehat{ate_{20}}$	1.9716	1.0431	.8383	.7699	1.7223	.8069	.6788	.6560
$bias(\widehat{ate_{10}})$	-.0074				.4001			
$bias(\widehat{ate_{20}})$	-.0283				-.2776			
$se(\hat{\alpha}_0)$			.0083	.0109			.0071	.0089
$se(\hat{\alpha}_1)$			.0184	.0467			.0429	.0464
$se(\hat{\alpha}_2)$			.0669	.0660			.0442	.0398
$se(\widehat{ate_{10}})$			.0167	.0434			.0466	.0501
$se(\widehat{ate_{20}})$			.0662	.0651			.0428	.0382

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.6: Model without Correlated Random Coefficients but with Symmetric Instrument, N=2000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0004	.1137	.0888	.0858	.9717	.0981	.0880	.0843
$\hat{\alpha}_1$	2.0001	.2493	.1767	.1886	2.3704	.2407	.2146	.2250
$\hat{\alpha}_2$	2.9930	.7036	.5215	.5334	2.7053	.5367	.4519	.4593
$\widehat{ate_{10}}$	.9997	.2742	.1970	.2072	1.3986	.2895	.2629	.2676
$\widehat{ate_{20}}$	1.9926	.7127	.5293	.5408	1.7335	.5362	.4542	.4613
$bias(\widehat{ate_{10}})$	-.0002				.3986			
$bias(\widehat{ate_{20}})$	-.0073				-.2664			
$se(\hat{\alpha}_0)$			.0050	.0054			.0040	.0045
$se(\hat{\alpha}_1)$			.0147	.0241			.0214	.0233
$se(\hat{\alpha}_2)$			.0275	.0329			.0207	.0199
$se(\widehat{ate_{10}})$			.0133	.0224			.0235	.0251
$se(\widehat{ate_{20}})$			.0270	.0325			.0199	.0191

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.7: Model without Correlated Random Coefficients but with Symmetric Instrument, N=5000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	.9991	.0676	.0558	.0543	.9721	.0596	.0539	.0530
$\hat{\alpha}_1$	2.0000	.1540	.1244	.1172	2.3682	.1457	.1402	.1422
$\hat{\alpha}_2$	2.9978	.4258	.3223	.3351	2.7100	.3362	.2804	.2865
$\widehat{ate_{10}}$	1.0009	.1682	.1361	.1295	1.3961	.1752	.1682	.1696
$\widehat{ate_{20}}$	1.9987	.4311	.3270	.3390	1.7379	.3366	.2821	.2882
$bias(\widehat{ate_{10}})$	.0009				.3961			
$bias(\widehat{ate_{20}})$	-.0012				-.2620			
$se(\hat{\alpha}_0)$			.0019	.0021			.0015	.0019
$se(\hat{\alpha}_1)$			.0084	.0098			.0098	.0094
$se(\hat{\alpha}_2)$			.0109	.0130			.0077	.0079
$se(\widehat{ate_{10}})$			.0077	.0091			.0108	.0102
$se(\widehat{ate_{20}})$			.0108	.0128			.0074	.0076

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.8: Model without Correlated Random Coefficients but with Symmetric Instrument, N=10000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0002	.0482	.0383	.0377	.9722	.0426	.0372	.0374
$\hat{\alpha}_1$	2.0001	.1088	.0812	.0823	2.3691	.1074	.0959	.1004
$\hat{\alpha}_2$	3.0011	.3007	.2337	.2378	2.7116	.2401	.1990	.2024
$\widehat{ate_{10}}$	.9999	.1191	.0892	.0900	1.3968	.1289	.1142	.1190
$\widehat{ate_{20}}$	2.0009	.3046	.2367	.2403	1.7393	.2397	.2009	.2036
$bias(\widehat{ate_{10}})$	-.0001				.3968			
$bias(\widehat{ate_{20}})$	.0009				-.2606			
$se(\hat{\alpha}_0)$			.0009	.0010			.0007	.0009
$se(\hat{\alpha}_1)$			.0027	.0050			.0048	.0047
$se(\hat{\alpha}_2)$			.0059	.0065			.0039	.0040
$se(\widehat{ate_{10}})$			.0025	.0046			.0050	.0050
$se(\widehat{ate_{20}})$			.0058	.0064			.0037	.0038

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.9: Model without Correlated Random Coefficients but with Asymmetric Instrument, $\eta_{g,j} = \eta_j$, N=1000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0070	.5628	.5115	.5102	1.0089	.5760	.5334	.5343
$\hat{\alpha}_1$	2.0070	.6340	.5617	.5897	2.0088	.7174	.5878	.6417
$\hat{\alpha}_2$	3.0049	.5437	.5885	.5584	3.0042	.5483	.5948	.5619
$\widehat{ate_{10}}$	1.0000	.9553	.8776	.8795	.9999	1.0526	.9223	.9477
$\widehat{ate_{20}}$	1.9979	.7651	.7600	.7391	1.9953	.7720	.7769	.7592
$bias(\widehat{ate_{10}})$	.0000				-.0001			
$bias(\widehat{ate_{20}})$	-.0020				-.0046			
$se(\hat{\alpha}_0)$			.0285	.0290			.0325	.0326
$se(\hat{\alpha}_1)$			.0281	.0358			.0372	.0444
$se(\hat{\alpha}_2)$			.0413	.0345			.0422	.0350
$se(\widehat{ate_{10}})$			.0384	.0450			.0497	.0537
$se(\widehat{ate_{20}})$			.0339	.0303			.0351	.0317

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.10: Model without Correlated Random Coefficients but with Asymmetric Instrument, $\eta_{g,j} = \eta_j$, N=2000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0012	.3857	.3539	.3606	1.0015	.3966	.3679	.3778
$\hat{\alpha}_1$	1.9970	.3965	.4265	.4170	1.9983	.4486	.4706	.4529
$\hat{\alpha}_2$	3.0074	.4024	.3917	.3948	3.0073	.4057	.3925	.3979
$\widehat{ate_{10}}$	.9957	.6271	.6157	.6217	.9968	.6872	.6689	.6694
$\widehat{ate_{20}}$	2.0062	.5426	.5191	.5220	2.0057	.5547	.5299	.5369
$bias(\widehat{ate_{10}})$	-.0042				-.0031			
$bias(\widehat{ate_{20}})$	.0062				.0057			
$se(\hat{\alpha}_0)$			.0122	.0143			.0141	.0160
$se(\hat{\alpha}_1)$			.0147	.0177			.0214	.0217
$se(\hat{\alpha}_2)$			.0148	.0173			.0153	.0176
$se(\widehat{ate_{10}})$			.0178	.0220			.0247	.0260
$se(\widehat{ate_{20}})$			.0128	.0148			.0135	.0154

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.11: Model without Correlated Random Coefficients but with Asymmetric Instrument, $\eta_{g,j} = \eta_j$, N=5000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0023	.2267	.2362	.2277	1.0023	.2322	.2485	.2387
$\hat{\alpha}_1$	1.9962	.2625	.2539	.2636	1.9967	.2851	.2781	.2862
$\hat{\alpha}_2$	3.0002	.2572	.2407	.2493	3.0004	.2601	.2443	.2516
$\widehat{ate_{10}}$	.9938	.3867	.3921	.3926	.9944	.4158	.4228	.4228
$\widehat{ate_{20}}$	1.9979	.3351	.3290	.3295	1.9981	.3405	.3391	.3393
$bias(\widehat{ate_{10}})$	-.0061				-.0055			
$bias(\widehat{ate_{20}})$	-.0020				-.0018			
$se(\hat{\alpha}_0)$			.0073	.0058			.0082	.0064
$se(\hat{\alpha}_1)$			.0054	.0072			.0079	.0087
$se(\hat{\alpha}_2)$			.0056	.0069			.0059	.0070
$se(\widehat{ate_{10}})$			.0086	.0089			.0111	.0104
$se(\widehat{ate_{20}})$			.0061	.0059			.0066	.0061

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.12: Model without Correlated Random Coefficients but with Asymmetric Instrument, $\eta_{g,j} = \eta_j$, N=10000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0028	.1613	.1559	.1610	1.0028	.1655	.1655	.1689
$\hat{\alpha}_1$	1.9997	.1806	.1899	.1863	1.9998	.1959	.2085	.2022
$\hat{\alpha}_2$	3.0013	.1754	.1749	.1763	3.0015	.1774	.1763	.1780
$\widehat{ate_{10}}$	.9969	.2702	.2775	.2776	.9970	.2892	.3033	.2989
$\widehat{ate_{20}}$	1.9985	.2323	.2307	.2330	1.9986	.2375	.2379	.2400
$bias(\widehat{ate_{10}})$	-.0030				-.0029			
$bias(\widehat{ate_{20}})$	-.0014				-.0013			
$se(\hat{\alpha}_0)$			.0024	.0028			.0031	.0032
$se(\hat{\alpha}_1)$			.0031	.0036			.0044	.0043
$se(\hat{\alpha}_2)$			.0034	.0034			.0035	.0034
$se(\widehat{ate_{10}})$			.0036	.0044			.0052	.0051
$se(\widehat{ate_{20}})$			.0028	.0029			.0031	.0030

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.13: Model without Correlated Random Coefficients but with Symmetric Instrument, $\eta_{g,j} = \eta_j$, N=1000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0045	.5044	.4942	.4976	1.0045	.5108	.4992	.5033
$\hat{\alpha}_1$	1.9959	.4345	.4461	.4584	1.9979	.4646	.4593	.5124
$\hat{\alpha}_2$	3.0060	.4603	.4646	.4625	3.0047	.4632	.4666	.4675
$\widehat{ate_{10}}$	.9913	.7033	.7182	.7240	.9934	.7343	.7294	.7755
$\widehat{ate_{20}}$	2.0015	.6754	.6679	.6706	2.0002	.6816	.6728	.6788
$bias(\widehat{ate_{10}})$	-.0086				-.0065			
$bias(\widehat{ate_{20}})$	.0015				.0002			
$se(\hat{\alpha}_0)$			.0267	.0327			.0280	.0334
$se(\hat{\alpha}_1)$			.0249	.0264			.0271	.0362
$se(\hat{\alpha}_2)$			.0275	.0284			.0277	.0285
$se(\widehat{ate_{10}})$			.0316	.0356			.0315	.0412
$se(\widehat{ate_{20}})$			.0261	.0297			.0262	.0298

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.14: Model without Correlated Random Coefficients but with Symmetric Instrument, $\eta_{g,j} = \eta_j$, N=2000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	.9997	.3357	.3363	.3515	1.0001	.3385	.3415	.3562
$\hat{\alpha}_1$	2.0009	.3175	.3457	.3235	2.0028	.3449	.3886	.3618
$\hat{\alpha}_2$	2.9995	.3211	.3197	.3266	2.9982	.3231	.3258	.3307
$\widehat{ate_{10}}$	1.0011	.4914	.5093	.5111	1.0027	.5211	.5560	.5481
$\widehat{ate_{20}}$	1.9997	.4594	.4590	.4733	1.9980	.4621	.4655	.4800
$bias(\widehat{ate_{10}})$	.0011				.0027			
$bias(\widehat{ate_{20}})$	-.0002				-.0019			
$se(\hat{\alpha}_0)$			.0134	.0164			.0139	.0168
$se(\hat{\alpha}_1)$			.0115	.0131			.0195	.0181
$se(\hat{\alpha}_2)$			.0112	.0141			.0117	.0141
$se(\widehat{ate_{10}})$			.0155	.0179			.0203	.0206
$se(\widehat{ate_{20}})$			.0117	.0145			.0120	.0146

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.15: Model without Correlated Random Coefficients but with Symmetric Instrument, $\eta_{g,j} = \eta_j$, N=5000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	.9956	.2154	.2129	.2221	.9950	.2190	.2156	.2254
$\hat{\alpha}_1$	1.9966	.2070	.2074	.2046	1.9984	.2319	.2313	.2287
$\hat{\alpha}_2$	3.0023	.2052	.2018	.2065	3.0016	.2067	.2029	.2093
$\widehat{ate_{10}}$	1.0009	.3188	.3165	.3229	1.0034	.3435	.3404	.3466
$\widehat{ate_{20}}$	2.0066	.2938	.2883	.2991	2.0065	.2971	.2907	.3036
$bias(\widehat{ate_{10}})$	.0009				.0034			
$bias(\widehat{ate_{20}})$	.0066				.0065			
$se(\hat{\alpha}_0)$			.0050	.0064			.0053	.0066
$se(\hat{\alpha}_1)$			.0056	.0052			.0085	.0070
$se(\hat{\alpha}_2)$			.0048	.0055			.0049	.0055
$se(\widehat{ate_{10}})$			.0058	.0070			.0083	.0080
$se(\widehat{ate_{20}})$			.0047	.0057			.0048	.0057

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.16: Model without Correlated Random Coefficients but with Symmetric Instrument, $\eta_{g,j} = \eta_j$, N=10000, and I=10000

	CF Approach				IV Approach			
	Estimate	A. SE	BS. SE	M.C. SD	Estimate	U. SE	BS. SE	M.C. SD
$\hat{\alpha}_0$	1.0015	.1617	.1532	.1570	1.0015	.1644	.1555	.1594
$\hat{\alpha}_1$	1.9954	.1402	.1470	.1446	1.9957	.1579	.1680	.1617
$\hat{\alpha}_2$	3.0019	.1474	.1521	.1460	3.0018	.1478	.1550	.1480
$\widehat{ate_{10}}$	.9938	.2275	.2257	.2282	.9941	.2449	.2452	.2450
$\widehat{ate_{20}}$	2.0004	.2167	.2117	.2114	2.0003	.2185	.2160	.2147
$bias(\widehat{ate_{10}})$	-.0061				-.0058			
$bias(\widehat{ate_{20}})$	.0004				.0003			
$se(\hat{\alpha}_0)$			.0034	.0032			.0035	.0033
$se(\hat{\alpha}_1)$			.0023	.0026			.0042	.0035
$se(\hat{\alpha}_2)$			.0032	.0027			.0032	.0028
$se(\widehat{ate_{10}})$			.0034	.0035			.0045	.0040
$se(\widehat{ate_{20}})$			.0031	.0028			.0032	.0029

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
A. SE= Analytical Standard Error. U. SE= Uncorrected Standard Error.
BS. SE=Bootstrapped Standard Error. M.C. SD= Monte Carlo Standard Deviation.

Table A.17: Model without Correlated Random Coefficients but with Misspecification and Asymmetric Instrument, N=1000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\alpha}_0$	.9310	.6576	.7161	1.3118	.4491	.3495
$\hat{\alpha}_1$	1.5886	.9572	1.2296	1.6833	.4969	.4129
$\hat{\alpha}_2$	4.4840	2.2880	1.5690	3.0697	1.4534	1.1045
$\widehat{ate_{10}}$	.6453	1.0073	1.2998	.3705	.7117	.5793
$\widehat{ate_{20}}$	3.5647	2.7309	2.2873	1.7553	.7834	.6469
$bias(\widehat{ate_{10}})$	-.3546			-.6294		
$bias(\widehat{ate_{20}})$	1.5647			-.2446		
$se(\hat{\alpha}_0)$		.0547	.1621		.1437	.0854
$se(\hat{\alpha}_1)$		.0575	.1642		.1364	.0856
$se(\hat{\alpha}_2)$		.3402	.2581		.3788	.1869
$se(\widehat{ate_{10}})$		.0551	.1691		.2005	.1230
$se(\widehat{ate_{20}})$		.2927	.3345		.1670	.0705

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table A.18: Model without Correlated Random Coefficients but with Misspecification and Asymmetric Instrument, N=2000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\alpha}_0$	.9210	.4811	.4993	1.3125	.2614	.2417
$\hat{\alpha}_1$	1.6553	.9694	.8787	1.6767	.2865	.2868
$\hat{\alpha}_2$	4.5738	1.2292	1.0925	3.0702	.7646	.7612
$\widehat{ate_{10}}$	.7374	.9876	.9243	.3619	.4100	.3994
$\widehat{ate_{20}}$	3.6540	1.6400	1.5774	1.7568	.4567	.4433
$bias(\widehat{ate_{10}})$	-.2625			-.6380		
$bias(\widehat{ate_{20}})$	1.6540			-.2431		
$se(\hat{\alpha}_0)$		.0312	.0832		.0415	.0379
$se(\hat{\alpha}_1)$		.0451	.0814		.0372	.0385
$se(\hat{\alpha}_2)$		.1229	.1328		.0798	.0822
$se(\widehat{ate_{10}})$		.0437	.0837		.0548	.0551
$se(\widehat{ate_{20}})$		.1040	.1704		.0331	.0296

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table A.19: Model without Correlated Random Coefficients but with Misspecification and Asymmetric Instrument, N=5000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\alpha}_0$	.9144	.2755	.3160	1.3100	.1561	.1496
$\hat{\alpha}_1$	1.6573	.5443	.5585	1.6850	.1895	.1782
$\hat{\alpha}_2$	4.6245	.6799	.6815	3.0713	.4877	.4732
$\widehat{ate_{10}}$	.7486	.5673	.5826	.3737	.2625	.2503
$\widehat{ate_{20}}$	3.7099	.9235	.9838	1.7626	.2949	.2795
$bias(\widehat{ate_{10}})$	-.2513			-.6262		
$bias(\widehat{ate_{20}})$	1.7099			-.2373		
$se(\hat{\alpha}_0)$		.0452	.0354		.0162	.0140
$se(\hat{\alpha}_1)$		.0181	.0326		.0156	.0144
$se(\hat{\alpha}_2)$		.0535	.0565		.0322	.0307
$se(\widehat{ate_{10}})$		.0180	.0333		.0227	.0204
$se(\widehat{ate_{20}})$		.0725	.0713		.0138	.0110

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table A.20: Model without Correlated Random Coefficients but with Misspecification and Asymmetric Instrument, N=10000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\alpha}_0$	.9127	.2037	.2218	1.3055	.1051	.1061
$\hat{\alpha}_1$	1.6801	.4012	.3936	1.6925	.1219	.1256
$\hat{\alpha}_2$	4.6405	.4644	.4835	3.0690	.3257	.3336
$\widehat{ate_{10}}$	.7718	.4153	.4057	.3875	.1707	.1758
$\widehat{ate_{20}}$	3.7286	.6497	.6953	1.7626	.1948	.1957
$bias(\widehat{ate_{10}})$	-.2281			-.6124		
$bias(\widehat{ate_{20}})$	1.7286			-.2373		
$se(\hat{\alpha}_0)$		.0146	.0185		.0070	.0068
$se(\hat{\alpha}_1)$		.0097	.0160		.0067	.0070
$se(\hat{\alpha}_2)$		.0249	.0294		.0141	.0150
$se(\widehat{ate_{10}})$		.0094	.0163		.0098	.0100
$se(\widehat{ate_{20}})$		.0281	.0372		.0049	.0054

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

A.8 Chapter 1: Tables-Empirical Analysis

Table A.21: Variables Description and Summary Statistics, N=38779

Variable	Mean	S.D	Min	Max	Description
hearnings	2.18	.50	1.34	4.62	Log of hourly earnings in 1989
ep	2.01	.62	1	3	English proficiency (not well, well, and very well)
oarrival	2.64	.69	0	3	Age at arrival (US born, 0 to 11 years old,12 to 17 years old, and 18 or older)
educ	11.18	3.73	0	20	Number of years of education
exp	19.14	11.89	0	58	Number of years of potential experience
gender	.34	.47	0	1	Dummy = 1 if female
managerial	.16	.36	0	1	Dummy = 1 if in managerial or professional specialty occupations
technical	.25	.43	0	1	Dummy = 1 if in technical, sales or administrative support occupations
service	.13	.34	0	1	Dummy = 1 if in service occupations
repair	.16	.36	0	1	Dummy = 1 if in precision production, craft or repair occupations
operators	.24	.43	0	1	Dummy = 1 if operators, fabricators or laborers
military	.003	.05	0	1	Dummy = 1 if in military
usborn	.01	.10	0	1	Dummy = 1 if born in the USA
spanish	.11	.32	0	1	Dummy = 1 if from Spain
mexican	.56	.49	0	1	Dummy = 1 if from Mexico
camerica	.06	.23	0	1	Dummy = 1 if from Central America
samerica	.05	.22	0	1	Dummy = 1 if from South America
puerto	.07	.27	0	1	Dummy = 1 if from Puerto Rico
cuban	.05	.22	0	1	Dummy = 1 if from Cuba
empu	.02	.15	0	1	Dummy = 1 if unemployed
empnl	.02	.15	0	1	Dummy = 1 if not in labor force
classp	.78	.41	0	1	Dummy = 1 if working in private for profit company
classnp	.04	.20	0	1	Dummy = 1 if working in private for nonprofit organization
classg	.15	.36	0	1	Dummy = 1 if working in government
age	36.09	10.84	16	64	Years of age

Note: N=Sample Size. S.D=Standard Deviation.

Table A.22: English Proficiency, Earnings, and Other Characteristics

	Treatment			
	1	2	3	Total
Number of observations	7248	23655	7876	38779
Percentage of observations	18.67	61.02	20.31	100
Average log wage (in dollars)	1.91	2.21	2.34	2.18
S.D	.40	.49	.52	.50
Average education (in years)	7.80	11.68	12.79	11.18
Average experience (in years)	23.35	18.92	15.92	19.14
Females (in percentages)	27.73	34.76	37.79	34.06
Managerial positions (in percentages)	3.61	17.50	23.69	16.16
Operators (in percentages)	41.34	22.39	16.86	24.81
Mexican (in percentages)	62.23	55.64	55.57	56.86
Born outside USA (in percentages)	99.36	98.62	99.24	98.89
Unemployment rate	3.14	2.39	1.94	2.44
Age	36.63	36.41	34.64	36.09

Note: S.D=Standard Deviation.

Table A.23: Multinomial Logit Regressions of English Proficiency, N=38779

	Model					
English Proficiency	1	2	3	3a	4	4a
Well	-.20***	-.20***	-.25***	-	-.25***	-
Very Well	.71***	.71***	.60***	-	.60***	-
Arrival Age	Yes	Yes	Yes	-	Yes	-
Education	-	-	Yes	Yes	Yes	Yes
Gender	-	-	-	-	Yes	Yes
χ^2	1452	1452	8905	7553	8927	7577
Change in χ^2 (in percentages)	-	-	+513	-15	+0.2	-15

Note: English proficiency has 3 levels: Not well, well, and very well with not well being the base outcome in the all regressions. All numbers in the rows associated with treatment levels indicate the estimated parameters on Hispanic workers' arrival age in the USA. The arrival age has four groups: US born, arrived as a child (0 to 11 years old), arrived as a teenager (12-17 years old) , and arrived as an adult (18 or older). The predicted probabilities from these first stage regressions are used as instruments for the binary endogenous English proficiency indicators in the associated IV regressions of Table A.25. The *** besides a number indicates statistical significance at the 1% level. χ^2 is the likelihood ratio chi square test statistic of goodness of fit. N=Sample Size.

Table A.24: Multinomial Logit Regressions of English Proficiency (Continuing), N=38779

	Model							
English Proficiency	5	5a	6	6a	7	7a	8	8a
Well	-.27***	-	-.26***	-	-.26***	-	-.27***	-
Very Well	.59***	-	.51***	-	.51***	-	.50***	-
Arrival Age	Yes	-	Yes	-	Yes	-	Yes	-
Education	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ancestry	-	-	Yes	Yes	Yes	Yes	Yes	Yes
Employment Status	-	-	-	-	Yes	Yes	Yes	Yes
Worker Class	-	-	-	-	-	-	Yes	Yes
χ^2	10098	8747	12092	11038	12096	11044	12470	11401
Change in χ^2 (in percentages)	+13.1	-13.4	+19.7	-8.7	+0.03	-8.7	+3.1	-8.6

Note: English proficiency has 3 levels: Not well, well, and very well with not well being the base outcome in the all regressions. All numbers in the rows associated with treatment levels indicate the estimated parameters on Hispanic workers' arrival age in the USA. The arrival age has four groups: US born, arrived as a child (0 to 11 years old), arrived as a teenager (12-17 years old), and arrived as an adult (18 or older). The predicted probabilities from these first stage regressions are used as instruments for the binary endogenous English proficiency indicators in the associated IV regressions of Table A.26. The *** besides a number indicates statistical significance at the 1% level. χ^2 is the likelihood ratio chi square test statistic of goodness of fit. N=Sample Size.

Table A.25: Average Treatment Effects (ATEs) of English Proficiency on Log Hourly Wages, N=38779

Estimation Method	ATEs	Model 1	Model 2	Model 3	Model 4
CF	$ep_2 - ep_1$	3.59	-13.63	.30	.39
		(10.28)	(10.33)	(.34)	(.34)
	$ep_3 - ep_1$	5.43	5.85	.83**	1.09***
		(11.86)	(11.89)	(.40)	(.40)
IV	$ep_2 - ep_1$	.85***	.77***	.48***	.56***
		(.05)	(.07)	(.05)	(.05)
	$ep_3 - ep_1$	.95***	.33**	.08	.18***
		(.05)	(.16)	(.05)	(.05)
OLS	$ep_2 - ep_1$	.29***	.32***	.14***	.15***
		(.005)	(.005)	(.006)	(.006)
	$ep_3 - ep_1$	.43***	.48***	.26***	.27***
		(.007)	(.007)	(.007)	(.007)
Experience		-	Yes	Yes	Yes
Education		-	-	Yes	Yes
Gender		-	-	-	Yes

Note: The dependent variable in all regressions is log hourly wages. English proficiency has 3 levels: Not well, well, and very well with binary indicators ep_1, ep_2, and ep_3 respectively. CF is control function estimation with control function terms. IV is instrumental variables estimation augmented with interactions of binary endogenous English proficiency indicators and exogenous outcome variables. All exogenous regressors are demeaned in all regressions. The numbers in parentheses are the standard errors of ATE estimates. The standard errors in CF regressions come from the analytical formula. The standard errors in IV regressions are bootstrapped. The *** besides a number indicates statistical significance at the 1% level and the ** at the 5% level. N=Sample Size.

Table A.26: Average Treatment Effects (ATEs) of English Proficiency on Log Hourly Wages (Continuing), N=38779

Estimation Method	ATEs	Model 5	Model 6	Model 7	Model 8
CF	$ep_2 - ep_1$	.34	.21	.18	.30
		(.31)	(.23)	(.20)	(.21)
	$ep_3 - ep_1$	.67*	.42*	.38*	.79***
		(.36)	(.24)	(.22)	(.24)
IV	$ep_2 - ep_1$	.27***	.25	.24	.36
		(.08)	(2.63)	(2.87)	(1.32)
	$ep_3 - ep_1$	-.04	-.43	-.47	-.37
		(.08)	(4.77)	(6.49)	(5.12)
OLS	$ep_2 - ep_1$	.12***	.12***	.12***	.12***
		(.005)	(.005)	(.005)	(.005)
	$ep_3 - ep_1$	.23***	.22***	.22***	.22***
		(.007)	(.007)	(.007)	(.007)
Experience		Yes	Yes	Yes	Yes
Education		Yes	Yes	Yes	Yes
Gender		Yes	Yes	Yes	Yes
Occupation		Yes	Yes	Yes	Yes
Ancestry		-	Yes	Yes	Yes
Employment Status		-	-	Yes	Yes
Worker Class		-	-	-	Yes

Note: The dependent variable in all regressions is log hourly wages. English proficiency has 3 levels: Not well, well, and very well with binary indicators ep_1, ep_2, and ep_3 respectively. CF is control function estimation with control function terms. IV is instrumental variables estimation augmented with interactions of binary endogenous English proficiency indicators and exogenous outcome variables. All exogenous regressors are demeaned in all regressions. The numbers in parentheses are the standard errors of ATE estimates. The standard errors in CF regressions come from the analytical formula. The standard errors in IV regressions are bootstrapped. The *** besides a number indicates statistical significance at the 1% level and the * at the 10% level. N=Sample Size.

Table A.27: Average Treatment Effects (ATEs) of English Proficiency on Log Hourly Wages with Bounds, N=38779

ATEs/Estimation Method	CF	IV	OLS	Lower Bound	Upper Bound
$ep_2 - ep_1$	.30	.36	.12***	0	.32
	(.21)	(1.32)	(.005)	[0, .33]	
$ep_3 - ep_1$	.79***	-.37	.22***	0	.43
	(.24)	(5.12)	(.007)	[0, .44]	

Note: The dependent variable in all regressions is log hourly wages. English proficiency has 3 levels: Not well, well, and very well with binary indicators ep_1, ep_2, and ep_3 respectively. CF is control function estimation with control function terms. IV is instrumental variables estimation augmented with interactions of binary endogenous English proficiency indicators and exogenous outcome variables. Bounds are calculated based on the combined MTS and MTR assumptions of Manski and Pepper (2000). The regressors and instruments used in CF, IV, and OLS regressions are potential experience, education, gender, occupation, ancestry, employment status, worker class, and the predicted probabilities, respectively. All exogenous regressors are demeaned in CF, IV, and OLS regressions. The numbers in parentheses are the standard errors of ATE estimates. The standard errors in CF regressions come from the analytical formula. The standard errors in IV regressions are bootstrapped. The numbers in brackets are the lower and upper limits of the 95% bootstrap confidence intervals. The *** besides a number indicates statistical significance at the 1% level. N=Sample Size.

Table A.28: Average Treatment Effects (ATEs) of English Proficiency on Log Hourly Wages with Bounds, Only Male, N=25568

ATEs/Estimation Method	CF	IV	OLS	Lower Bound	Upper Bound
$ep_2 - ep_1$	.33	.12	.13***	0	.33
	(.25)	(1.17)	(.007)	[0, .35]	
$ep_3 - ep_1$	.85***	-4.37	.24***	0	.47
	(.30)	(3.50)	(.009)	[0, .49]	

Note: The dependent variable in all regressions is log hourly wages. English proficiency has 3 levels: Not well, well, and very well with binary indicators ep_1, ep_2, and ep_3 respectively. CF is control function estimation with control function terms. IV is instrumental variables estimation augmented with interactions of binary endogenous English proficiency indicators and exogenous outcome variables. Bounds are calculated based on the combined MTS and MTR assumptions of Manski and Pepper (2000). The regressors and instruments used in CF, IV, and OLS regressions are potential experience, education, gender, occupation, ancestry, employment status, worker class, and the predicted probabilities, respectively. All exogenous regressors are demeaned in CF, IV, and OLS regressions. The numbers in parentheses are the standard errors of ATE estimates. The standard errors in CF regressions come from the analytical formula. The standard errors in IV regressions are bootstrapped. The numbers in brackets are the lower and upper limits of the 95% bootstrap confidence intervals. The *** besides a number indicates statistical significance at the 1% level. N=Sample Size.

Table A.29: Average Treatment Effects (ATEs) of English Proficiency on Log Hourly Wages with Bounds, Only Female, N=13211

ATEs/Estimation Method	CF	IV	OLS	Lower Bound	Upper Bound
$ep_2 - ep_1$	.20	.06	.08***	0	.32
	(.37)	(3.16)	(.01)	[0, .34]	
$ep_3 - ep_1$	.46**	-.36	.16***	0	.40
	(.22)	(2.71)	(.01)	[0, .42]	

Note: The dependent variable in all regressions is log hourly wages. English proficiency has 3 levels: Not well, well, and very well with binary indicators ep_1, ep_2, and ep_3 respectively. CF is control function estimation with control function terms. IV is instrumental variables estimation augmented with interactions of binary endogenous English proficiency indicators and exogenous outcome variables. Bounds are calculated based on the combined MTS and MTR assumptions of Manski and Pepper (2000). The regressors and instruments used in CF, IV, and OLS regressions are potential experience, education, gender, occupation, ancestry, employment status, worker class, and the predicted probabilities, respectively. All exogenous regressors are demeaned in CF, IV, and OLS regressions. The numbers in parentheses are the standard errors of ATE estimates. The standard errors in CF regressions come from the analytical formula. The standard errors in IV regressions are bootstrapped. The numbers in brackets are the lower and upper limits of the 95% bootstrap confidence intervals. The *** besides a number indicates statistical significance at the 1% level and the ** at the 5% level. N=Sample Size.

Table A.30: Average Treatment Effects (ATEs) of English Proficiency on Log Hourly Wages with Bounds, Only Operators, N=9622

ATEs/Estimation Method	CF	IV	OLS	Lower Bound	Upper Bound
$ep_2 - ep_1$	.68	.28	.13***	0	.22
	(.43)	(10.41)	(.009)	[0, .24]	
$ep_3 - ep_1$	1.37***	1.87	.22***	0	.30
	(.56)	(13.64)	(.01)	[0, .33]	

Note: The dependent variable in all regressions is log hourly wages. English proficiency has 3 levels: Not well, well, and very well with binary indicators ep_1, ep_2, and ep_3 respectively. CF is control function estimation with control function terms. IV is instrumental variables estimation augmented with interactions of binary endogenous English proficiency indicators and exogenous outcome variables. Bounds are calculated based on the combined MTS and MTR assumptions of Manski and Pepper (2000). The regressors and instruments used in CF, IV, and OLS regressions are potential experience, education, gender, occupation, ancestry, employment status, worker class, and the predicted probabilities, respectively. All exogenous regressors are demeaned in CF, IV, and OLS regressions. The numbers in parentheses are the standard errors of ATE estimates. The standard errors in CF regressions come from the analytical formula. The standard errors in IV regressions are bootstrapped. The numbers in brackets are the lower and upper limits of the 95% bootstrap confidence intervals. The *** besides a number indicates statistical significance at the 1% level. N=Sample Size.

Table A.31: Average Treatment Effects (ATEs) of English Proficiency on Log Hourly Wages with Bounds, Only Repair, N=6209

ATEs/Estimation Method	CF	IV	OLS	Lower Bound	Upper Bound
$ep_2 - ep_1$	.54	.14	.16***	0	.27
	(.52)	(4.39)	(.01)	[0, .30]	
$ep_3 - ep_1$	1.14*	-.31	.28***	0	.38
	(.63)	(6.30)	(.01)	[0, .42]	

Note: The dependent variable in all regressions is log hourly wages. English proficiency has 3 levels: Not well, well, and very well with binary indicators ep_1, ep_2, and ep_3 respectively. CF is control function estimation with control function terms. IV is instrumental variables estimation augmented with interactions of binary endogenous English proficiency indicators and exogenous outcome variables. Bounds are calculated based on the combined MTS and MTR assumptions of Manski and Pepper (2000). The regressors and instruments used in CF, IV, and OLS regressions are potential experience, education, gender, occupation, ancestry, employment status, worker class, and the predicted probabilities, respectively. All exogenous regressors are demeaned in CF, IV, and OLS regressions. The numbers in parentheses are the standard errors of ATE estimates. The standard errors in CF regressions come from the analytical formula. The standard errors in IV regressions are bootstrapped. The numbers in brackets are the lower and upper limits of the 95% bootstrap confidence intervals. The *** besides a number indicates statistical significance at the 1% level and the * at the 10% level. N=Sample Size.

Table A.32: Average Treatment Effects (ATEs) of English Proficiency on Log Hourly Wages with Bounds, Only Service, N=5417

ATEs/Estimation Method	CF	IV	OLS	Lower Bound	Upper Bound
$ep_2 - ep_1$	.21	.42	.09***	0	.23
	(.20)	(7.15)	(.01)	[0, .25]	
$ep_3 - ep_1$	.47**	-.19	.18***	0	.34
	(.23)	(4.53)	(.01)	[0, .38]	

Note: The dependent variable in all regressions is log hourly wages. English proficiency has 3 levels: Not well, well, and very well with binary indicators ep_1, ep_2, and ep_3 respectively. CF is control function estimation with control function terms. IV is instrumental variables estimation augmented with interactions of binary endogenous English proficiency indicators and exogenous outcome variables. Bounds are calculated based on the combined MTS and MTR assumptions of Manski and Pepper (2000). The regressors and instruments used in CF, IV, and OLS regressions are potential experience, education, gender, occupation, ancestry, employment status, worker class, and the predicted probabilities, respectively. All exogenous regressors are demeaned in CF, IV, and OLS regressions. The numbers in parentheses are the standard errors of ATE estimates. The standard errors in CF regressions come from the analytical formula. The standard errors in IV regressions are bootstrapped. The numbers in brackets are the lower and upper limits of the 95% bootstrap confidence intervals. The *** besides a number indicates statistical significance at the 1% level and the ** at the 5% level. N=Sample Size.

APPENDIX B

APPENDIX FOR CHAPTER 2

B.1 Chapter 2: Derivations in CF Method

The expectation of the observed outcome y conditional on the observed variables $(\mathbf{d}, \mathbf{x}, \mathbf{z})$, i.e., $E(y|\mathbf{d}, \mathbf{x}, \mathbf{z})$, is obtained in a couple of steps as shown below. Note I can write

$$
\begin{aligned}
E(y|\mathbf{d}, \mathbf{x}, \mathbf{z}) &= E(d_0 y_0 + d_1 y_1 + \cdots + d_G y_G|\mathbf{d}, \mathbf{x}, \mathbf{z}) \\
&= d_0 E(y_0|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_1 E(y_1|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \cdots + d_G E(y_G|\mathbf{d}, \mathbf{x}, \mathbf{z}) \\
&= d_0 E(m_0 + \mathbf{x}\mathbf{b}_0 + u_0|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_1 E(m_1 + \mathbf{x}\mathbf{b}_1 + u_1|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \cdots + \\
&\quad + d_G E(m_G + \mathbf{x}\mathbf{b}_G + u_G|\mathbf{d}, \mathbf{x}, \mathbf{z}) \\
&= d_0 E(m_0|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_0 \mathbf{x} E(\mathbf{b}_0|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_0 E(u_0|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \\
&\quad + d_1 E(m_1|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_1 \mathbf{x} E(\mathbf{b}_1|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_1 E(u_1|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \\
&\quad + \ldots + d_G E(m_G|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_G \mathbf{x} E(\mathbf{b}_G|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_G E(u_G|\mathbf{d}, \mathbf{x}, \mathbf{z}) \\
&= d_0(\psi_{o0} + \mathbf{x}\psi_0) + d_0 \mathbf{x} E(\kappa_{o0} + \mathbf{\Gamma}_0 \mathbf{x}' + \mathbf{v}_0|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_0 E(u_0|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \\
&\quad + d_1(\psi_{o1} + \mathbf{x}\psi_1) + d_1 \mathbf{x} E(\kappa_{o1} + \mathbf{\Gamma}_1 \mathbf{x}' + \mathbf{v}_1|\mathbf{d}, \mathbf{x}, \mathbf{z}) + d_1 E(u_1|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \\
&\quad + \cdots + d_G(\psi_{oG} + \mathbf{x}\psi_G) + d_G \mathbf{x} E(\kappa_{oG} + \mathbf{\Gamma}_G \mathbf{x}' + \mathbf{v}_G|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \\
&\quad + d_G E(u_G|\mathbf{d}, \mathbf{x}, \mathbf{z}) \\
&= d_0(\psi_{o0} + \mathbf{x}\psi_0) + d_0 \mathbf{x}(\kappa_{o0} + \mathbf{\Gamma}_0 \mathbf{x}' + E(\mathbf{v}_0|\mathbf{d}, \mathbf{x}, \mathbf{z})) + d_0 E(u_0|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \\
&\quad + d_1(\psi_{o1} + \mathbf{x}\psi_1) + d_1 \mathbf{x}(\kappa_{o1} + \mathbf{\Gamma}_1 \mathbf{x}' + E(\mathbf{v}_1|\mathbf{d}, \mathbf{x}, \mathbf{z})) + d_1 E(u_1|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \\
&\quad + \cdots + d_G(\psi_{oG} + \mathbf{x}\psi_G) + d_G \mathbf{x}(\kappa_{oG} + \mathbf{\Gamma}_G \mathbf{x}' + E(\mathbf{v}_G|\mathbf{d}, \mathbf{x}, \mathbf{z})) + \\
&\quad + d_G E(u_G|\mathbf{d}, \mathbf{x}, \mathbf{z}) \\
&= \sum_{j=0}^{G} d_j \psi_{oj} + \sum_{j=0}^{G} d_j \mathbf{x}\psi_j + \sum_{j=0}^{G} d_j \mathbf{x}\kappa_{oj} + \sum_{j=0}^{G} d_j \mathbf{x}\mathbf{\Gamma}_j \mathbf{x}' + \\
&\quad + \sum_{j=0}^{G} d_j \mathbf{x} E(\mathbf{v}_j|\mathbf{d}, \mathbf{x}, \mathbf{z}) + \sum_{j=0}^{G} d_j E(u_j|\mathbf{d}, \mathbf{x}, \mathbf{z}). \qquad \text{(B.1)}
\end{aligned}
$$

Next, I need to derive $E(u_j|\mathbf{d},\mathbf{x},\mathbf{z})$ and $E(\mathbf{v}_j|\mathbf{d},\mathbf{x},\mathbf{z})$. Under A.2.3, A.2.5 and the law of iterated expectations,

$$\begin{aligned} E(u_j|\mathbf{d},\mathbf{x},\mathbf{z}) &= E(E(u_j|\mathbf{d},\mathbf{x},\mathbf{z},\mathbf{a})|\mathbf{d},\mathbf{x},\mathbf{z}) \\ &= E(E(u_j|\mathbf{x},\mathbf{z},\mathbf{a})|\mathbf{d},\mathbf{x},\mathbf{z}) = E[\sum_{g=0}^{G}\eta_{j,g}(a_g - E(a_g))|\mathbf{d},\mathbf{x},\mathbf{z}] \\ &= \sum_{g=0}^{G}\eta_{j,g}E[(a_g - E(a_g))|\mathbf{d},\mathbf{x},\mathbf{z})]. \end{aligned} \tag{B.2}$$

$$\begin{aligned} E(\mathbf{v}_j|\mathbf{d},\mathbf{x},\mathbf{z}) &= E(E(\mathbf{v}_j|\mathbf{d},\mathbf{x},\mathbf{z},\mathbf{a})|\mathbf{d},\mathbf{x},\mathbf{z}) \\ &= E(E(\mathbf{v}_j|\mathbf{x},\mathbf{z},\mathbf{a})|\mathbf{d},\mathbf{x},\mathbf{z}) = E(\mathbf{P}\mathbf{a}'|\mathbf{d},\mathbf{x},\mathbf{z}) = \mathbf{P}E(\mathbf{a}'|\mathbf{d},\mathbf{x},\mathbf{z}) \\ &= \mathbf{P}E(\mathbf{a}'|\mathbf{d},\mathbf{x},\mathbf{z}). \end{aligned} \tag{B.3}$$

In equations above, I use that $\mathbf{d}$ is completely determined by $\mathbf{z}$ and $\mathbf{a}$ together. Refer to section 2.2 for seeing this where the main model is described. Hence, the expectation conditional on $\mathbf{d},\mathbf{x},\mathbf{z},\mathbf{a}$ reduces down to the one conditional on $\mathbf{x},\mathbf{z},\mathbf{a}$ only. Using A.2.1, mutual exclusivity of binary treatment indicators, and the facts right below for $g = 0, 1, \ldots, G$:

$$\begin{aligned} E(\mathbf{a}'|\mathbf{d},\mathbf{x},\mathbf{z}) &= d_0E(\mathbf{a}'|d_0 = 1,\mathbf{x},\mathbf{z}) + d_1E(\mathbf{a}'|d_1 = 1,\mathbf{x},\mathbf{z}) + \ldots + \\ &\quad + d_GE(\mathbf{a}'|d_G = 1,\mathbf{x},\mathbf{z}) \end{aligned} \tag{B.4}$$

$$\begin{aligned} E(u_g|\mathbf{d},\mathbf{x},\mathbf{z}) &= d_0E(u_g|d_0 = 1,\mathbf{x},\mathbf{z}) + d_1E(u_g|d_1 = 1,\mathbf{x},\mathbf{z}) + \ldots + \\ &\quad + d_GE(u_g|d_G = 1,\mathbf{x},\mathbf{z}), \end{aligned} \tag{B.5}$$

I can write $E(y|\mathbf{d},\mathbf{x},\mathbf{z})$ as follows:

$$\begin{aligned} E(y|\mathbf{d},\mathbf{x},\mathbf{z}) &= \sum_{j=0}^{G} d_j\psi_{oj} + \sum_{j=0}^{G} d_j\mathbf{x}\psi_j + \sum_{j=0}^{G} d_j\mathbf{x}\kappa_{oj} + \sum_{j=0}^{G} d_j\mathbf{x}\mathbf{\Gamma}_j\mathbf{x}' + \\ &\quad + \sum_{j=0}^{G} d_j\mathbf{x}\mathbf{P}E(\mathbf{a}'|d_j = 1,\mathbf{x},\mathbf{z}) + \\ &\quad + \sum_{j=0}^{G} d_j\sum_{g=0}^{G}\eta_{j,g}E[(a_g - E(a_g))|d_j = 1,\mathbf{x},\mathbf{z}]. \end{aligned} \tag{B.6}$$

To complete the derivation of the conditional expectation $E(y|\mathbf{d},\mathbf{x},\mathbf{z})$, all I need is to find closed form expressions for both $\sum_{j=0}^{G} d_j \mathbf{x}\mathbf{P}E(\mathbf{a}'|d_j=1,\mathbf{x},\mathbf{z})$ and $\sum_{j=0}^{G} d_j \sum_{g=0}^{G} \eta_{j,g} E[(a_g - E(a_g))|d_j=1,\mathbf{x},\mathbf{z}]$. I will start with the latter expression. At this point, it is crucial to remember a result from the work of Dubin and McFadden (1984). In their paper, they used the following result:

$$E(a_g - E(a_g)|d_j=1,\mathbf{x},\mathbf{z}) = \begin{cases} -log(\Lambda_g) & ,g=j \\ \dfrac{\Lambda_j log(\Lambda_j)}{(1-\Lambda_j)} & ,g\neq j \end{cases}, \tag{B.7}$$

where $\Lambda_g = exp(\boldsymbol{z}\gamma_g)/\sum_{r=0}^{G} exp(\boldsymbol{z}\gamma_r)$, i.e., the MNL response probability for $g,j=0,1,\ldots,G$.

Using the above result, I have

$$\begin{aligned}
\sum_{j=0}^{G} d_j \sum_{g=0}^{G} \eta_{j,g} E[(a_g - E(a_g))|d_j=1,\mathbf{x},\mathbf{z}] &= \sum_{j=0}^{G} d_j \left(\sum_{g=0}^{G} \eta_{j,g} E(a_g - E(a_g)|d_j=1,\mathbf{x},\mathbf{z}) \right) \\
&= \sum_{j=0}^{G} d_j \left(-\eta_{j,j} log(\Lambda_j) + \sum_{h\neq j} \eta_{j,h} \frac{\Lambda_h log(\Lambda_h)}{(1-\Lambda_h)} \right) \\
&= \sum_{j=0}^{G} d_j \left(-\eta_{j,j} log(\Lambda_j) + \sum_{h\neq j} \eta_{j,h} M_h \right) \\
&= \sum_{j=0}^{G} -\eta_{j,j} d_j log(\Lambda_j) + \sum_{j=0}^{G} \left(d_j \sum_{h\neq j} \eta_{j,h} M_h \right) \\
&= \left(\sum_{j=0}^{G} -\eta_{j,j} d_j log(\Lambda_j) \right) \\
&\quad + d_0 \left(\eta_{0,1} M_1 + \eta_{0,2} M_2 + \cdots + \eta_{0,G} M_G \right) \\
&\quad + d_1 \left(\eta_{1,0} M_0 + \eta_{1,2} M_2 + \eta_{1,3} M_3 + \cdots + \eta_{1,G} M_G \right) \\
&\quad \vdots \\
&\quad + d_G \left(\eta_{G,0} M_0 + \eta_{G,1} M_1 + \cdots + \eta_{G,G-1} M_{G-1} \right)
\end{aligned}$$

Then, I have the following equality:

$$\sum_{j=0}^{G} d_j \sum_{g=0}^{G} \eta_{j,g} E[(a_g - E(a_g))|d_j = 1, \mathbf{x}, \mathbf{z}] = \left(\sum_{j=0}^{G} - \eta_{j,j} d_j log(\Lambda_j) \right) + \sum_{j \neq 0} d_j \eta_{j,0} M_0 + \sum_{j \neq 1} d_j \eta_{j,1} M_1 + \cdots + \sum_{j \neq G} d_j \eta_{j,G} M_G, \quad \text{(B.8)}$$

where $M_g = \Lambda_g log(\Lambda_g)/(1 - \Lambda_g)$ for $g = 0, 1, \ldots, G$. Now, I will derive the former expression that CF method hinges on, $\sum_{j=0}^{G} d_j \mathbf{x} \mathbf{P} E(\mathbf{a}'|d_j = 1, \mathbf{x}, \mathbf{z})$. Let the $l \times (G+1)$ matrix of parameters, $\mathbf{P}$, have the following form:

$$\mathbf{P} = \begin{bmatrix} p_{1,0} & p_{1,1} & \cdots & p_{1,G} \\ p_{2,0} & p_{2,1} & \cdots & p_{2,G} \\ \vdots & \vdots & \ddots & \vdots \\ p_{l,0} & p_{l,1} & \cdots & p_{l,G} \end{bmatrix}_{lx(G+1).}$$

Thus, I have

$$\sum_{j=0}^{G} d_j \mathbf{x} \mathbf{P} E(\mathbf{a}'|d_j = 1, \mathbf{x}, \mathbf{z}) = \sum_{j=0}^{G} d_j \mathbf{x} \begin{pmatrix} \sum_{h=0}^{G} p_{1,h} E(a_h|d_j = 1, \mathbf{x}, \mathbf{z}) \\ \sum_{h=0}^{G} p_{2,h} E(a_h|d_j = 1, \mathbf{x}, \mathbf{z}) \\ \vdots \\ \sum_{h=0}^{G} p_{l,h} E(a_h|d_j = 1, \mathbf{x}, \mathbf{z}) \end{pmatrix}$$

$$= \sum_{j=0}^{G} [\sum_{h=0}^{G} p_{1,h} d_j x_1 E(a_h|d_j = 1, \mathbf{x}, \mathbf{z}) + \sum_{h=0}^{G} p_{2,h} d_j x_2 E(a_h|d_j = 1, \mathbf{x}, \mathbf{z}) + \cdots + \sum_{h=0}^{G} p_{l,h} d_j x_l E(a_h|d_j = 1, \mathbf{x}, \mathbf{z})]$$

Then, I have the following equality:

$$
\begin{aligned}
\sum_{j=0}^{G} d_j \mathbf{x}\mathbf{P} E(\mathbf{a}'|d_j = 1, \mathbf{x}, \mathbf{z}) &= \sum_{j=0}^{G}[(p_{1,0} d_j x_1 E(a_0|d_j = 1, \mathbf{x}, \mathbf{z}) + \cdots + \\
&\quad + p_{1,G} d_j x_1 E(a_G|d_j = 1, \mathbf{x}, \mathbf{z})) + \\
&\quad + (p_{2,0} d_j x_2 E(a_0|d_j = 1, \mathbf{x}, \mathbf{z}) + \cdots + \\
&\quad + p_{2,G} d_j x_2 E(a_G|d_j = 1, \mathbf{x}, \mathbf{z})) + \\
&\quad \vdots \\
&\quad + (p_{l,0} d_j x_l E(a_0|d_j = 1, \mathbf{x}, \mathbf{z}) + \cdots + \\
&\quad + p_{l,G} d_j x_l E(a_G|d_j = 1, \mathbf{x}, \mathbf{z}))] \\
&= p_{1,0} \sum_{j=0}^{G} d_j x_1 E(a_0|d_j = 1, \mathbf{x}, \mathbf{z}) + \cdots + \\
&\quad + p_{1,G} \sum_{j=0}^{G} d_j x_1 E(a_G|d_j = 1, \mathbf{x}, \mathbf{z}) + \\
&\quad + p_{2,0} \sum_{j=0}^{G} d_j x_2 E(a_0|d_j = 1, \mathbf{x}, \mathbf{z}) + \cdots + \\
&\quad + p_{2,G} \sum_{j=0}^{G} d_j x_2 E(a_G|d_j = 1, \mathbf{x}, \mathbf{z}) + \\
&\quad \vdots \\
&\quad + p_{l,0} \sum_{j=0}^{G} d_j x_l E(a_0|d_j = 1, \mathbf{x}, \mathbf{z}) + \cdots + \\
&\quad + p_{l,G} \sum_{j=0}^{G} d_j x_l E(a_G|d_j = 1, \mathbf{x}, \mathbf{z}) \\
&= \sum_{k=1,h=0}^{l,G} p_{k,h} \left(\sum_{j=0}^{G} d_j x_k E(a_h|d_j = 1, \mathbf{x}, \mathbf{z}) \right), \qquad \text{(B.9)}
\end{aligned}
$$

where $E(a_h|d_j = 1, \mathbf{x}, \mathbf{z})$ is as in (2.14) for $k = 1, 2, \ldots, l$ and $h, j = 0, 1, \ldots, G$.

By combining equations (B.6), (B.8), and (B.9), I can write the expectation of the observed outcome y conditional on the observed variables $(\mathbf{d}, \mathbf{x}, \mathbf{z})$ as follows:

$$
\begin{aligned}
E(y|\mathbf{d},\mathbf{x},\mathbf{z}) &= \sum_{j=0}^{G} d_j\psi_{oj} + \sum_{j=0}^{G} d_j\mathbf{x}\psi_j + \sum_{j=0}^{G} d_j\mathbf{x}\kappa_{oj} + \sum_{j=0}^{G} d_j\mathbf{x}\mathbf{\Gamma}_j\mathbf{x}' + \\
&\quad + \left(\sum_{j=0}^{G} -\eta_{j,j}d_j log(\Lambda_j)\right) + \sum_{j\neq 0} d_j\eta_{j,0}M_0 + \sum_{j\neq 1} d_j\eta_{j,1}M_1 + \\
&\quad + \cdots + \sum_{j\neq G} d_j\eta_{j,G}M_G + \\
&\quad + \sum_{k=1,h=0}^{l,G} p_{k,h}\left(\sum_{j=0}^{G} d_j x_k E(a_h|d_j=1,\mathbf{x},\mathbf{z})\right) \\
&= \sum_{j=0}^{G} d_j\psi_{oj} + \sum_{j=0}^{G} d_j\mathbf{x}(\psi_j+\kappa_{oj}) + \sum_{j=0}^{G} d_j(\mathbf{x}\otimes\mathbf{x})vec\mathbf{\Gamma}_j + \\
&\quad + \left(\sum_{j=0}^{G} -\eta_{j,j}d_j log(\Lambda_j)\right) + \sum_{j\neq 0} d_j\eta_{j,0}M_0 + \sum_{j\neq 1} d_j\eta_{j,1}M_1 + \\
&\quad + \cdots + \sum_{j\neq G} d_j\eta_{j,G}M_G + \\
&\quad + \sum_{k=1,h=0}^{l,G} p_{k,h}\left(\sum_{j=0}^{G} d_j x_k E(a_h|d_j=1,\mathbf{x},\mathbf{z})\right), \qquad \text{(B.10)}
\end{aligned}
$$

where Λ_j, M_j, and $E(a_h|d_j=1,\mathbf{x},\mathbf{z})$ are as in (2.14) for $h,j=0,1,\ldots,G$; $vec(\cdot)$ is the column vectorization operator; $vec(ABC)=(C'\otimes A)vec(B)$ for conformable matrices A, B, and C; and $vec(D)=D$ if and only if D is a one by one square matrix.

B.2 Chapter 2: Tables-Simulations

Table B.1: Model with Correlated Random Coefficients and Asymmetric Instrument, N=1000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9966	.4415	.6241	1.4465	.4882	.4743
$\hat{\psi}_{o1}$	1.9876	1.1375	1.5160	1.8806	.9356	.9522
$\hat{\psi}_{o2}$	2.9873	1.5026	1.6005	2.8735	1.3147	1.4189
$\hat{\Gamma}_0$	4.0006	.4631	.3825	4.0018	.5842	.5104
$\hat{\Gamma}_1$	4.9921	.4946	.5386	4.9997	.6364	.7890
$\hat{\Gamma}_2$	6.0035	.9609	.8968	6.0045	1.0026	.9493
$\widehat{ate_{10}}$	1.9824	1.2580	1.6561	1.4319	.9489	.9999
$\widehat{ate_{20}}$	3.9936	1.4499	1.5281	3.4297	1.1465	1.1986
$bias(\widehat{ate_{10}})$	-.0175			-.5680		
$bias(\widehat{ate_{20}})$	-.0063			-.5702		
$se(\hat{\psi}_{o0})$		.0451	.0998		.0587	.1086
$se(\hat{\psi}_{o1})$		.1323	.3922		.0834	.1666
$se(\hat{\psi}_{o2})$		.1012	.1372		.0962	.1195
$se(\hat{\Gamma}_0)$		.0846	.0826		.1001	.1660
$se(\hat{\Gamma}_1)$		.0682	.1232		.1197	.2831
$se(\hat{\Gamma}_2)$		.2847	.1956		.3037	.2192
$se(\widehat{ate_{10}})$		.1252	.3850		.0755	.1054
$se(\widehat{ate_{20}})$		.0962	.1070		.3108	.3111

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.2: Model with Correlated Random Coefficients and Asymmetric Instrument, N=2000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9984	.4606	.4343	1.4539	.3380	.3354
$\hat{\psi}_{o1}$	1.9788	1.2932	1.0743	1.8739	.8000	.6762
$\hat{\psi}_{o2}$	2.9758	1.1536	1.1198	2.8585	1.0442	.9905
$\hat{\Gamma}_0$	3.9980	.2806	.2704	3.9970	.3582	.3554
$\hat{\Gamma}_1$	5.0014	.5149	.3838	5.0072	.6448	.5441
$\hat{\Gamma}_2$	6.0034	.6197	.6247	6.0037	.6989	.6603
$\widehat{ate_{10}}$	1.9837	1.4239	1.1759	1.4302	.7204	.6984
$\widehat{ate_{20}}$	3.9827	1.1211	1.0674	3.4113	.8897	.8360
$bias(\widehat{ate_{10}})$	-.0162			-.5697		
$bias(\widehat{ate_{20}})$	-.0172			-.5886		
$se(\hat{\psi}_{o0})$		.0394	.0527		.0331	.0536
$se(\hat{\psi}_{o1})$		.1935	.2028		.0958	.0786
$se(\hat{\psi}_{o2})$		.0672	.0715		.0567	.0642
$se(\hat{\Gamma}_0)$		.0386	.0494		.0587	.0832
$se(\hat{\Gamma}_1)$		.1412	.0725		.1612	.1390
$se(\hat{\Gamma}_2)$		.0873	.1101		.1078	.1220
$se(\widehat{ate_{10}})$		.1925	.1983		.0446	.0486
$se(\widehat{ate_{20}})$		.0591	.0531		.1260	.1658

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.3: Model with Correlated Random Coefficients and Asymmetric Instrument, N=5000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9991	.2943	.2708	1.4524	.2536	.2120
$\hat{\psi}_{o1}$	1.9929	.7410	.6711	1.8796	.4351	.4225
$\hat{\psi}_{o2}$	2.9799	.7581	.7062	2.8625	.6654	.6292
$\hat{\Gamma}_0$	3.9989	.1703	.1712	4.0002	.2785	.2235
$\hat{\Gamma}_1$	4.9972	.2316	.2404	4.9989	.2817	.3393
$\hat{\Gamma}_2$	6.0019	.4364	.3886	6.0021	.4593	.4144
$\widehat{ate_{10}}$	1.9920	.8080	.7314	1.4259	.4550	.4386
$\widehat{ate_{20}}$	3.9837	.7243	.6704	3.4120	.5582	.5309
$bias(\widehat{ate_{10}})$	-.0079			-.5740		
$bias(\widehat{ate_{20}})$	-.0162			-.5879		
$se(\hat{\psi}_{o0})$		.0166	.0226		.0232	.0246
$se(\hat{\psi}_{o1})$		.0553	.0835		.0230	.0347
$se(\hat{\psi}_{o2})$		.0277	.0303		.0235	.0289
$se(\hat{\Gamma}_0)$		.0128	.0234		.0419	.0367
$se(\hat{\Gamma}_1)$		.0299	.0340		.0329	.0608
$se(\hat{\Gamma}_2)$		.0401	.0505		.0451	.0562
$se(\widehat{ate_{10}})$		.0487	.0817		.0205	.0189
$se(\widehat{ate_{20}})$		.0214	.0213		.0578	.0768

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.4: Model with Correlated Random Coefficients and Asymmetric Instrument, N=10000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9996	.2047	.1920	1.4527	.1563	.1507
$\hat{\psi}_{o1}$	2.0032	.4443	.4737	1.8782	.2873	.3000
$\hat{\psi}_{o2}$	2.9818	.5233	.5051	2.8677	.4609	.4467
$\hat{\Gamma}_0$	4.0007	.1480	.1212	3.9998	.1712	.1580
$\hat{\Gamma}_1$	4.9993	.1403	.1713	5.0036	.2339	.2392
$\hat{\Gamma}_2$	5.9993	.2952	.2764	5.9971	.3104	.2927
$\widehat{ate_{10}}$	2.0022	.4839	.5173	1.4293	.3146	.3093
$\widehat{ate_{20}}$	3.9807	.4874	.4759	3.4123	.3870	.3744
$bias(\widehat{ate_{10}})$	.0022			-.5706		
$bias(\widehat{ate_{20}})$	-.0192			-.5876		
$se(\hat{\psi}_{o0})$		.0133	.0124		.0105	.0129
$se(\hat{\psi}_{o1})$		.0182	.0429		.0115	.0177
$se(\hat{\psi}_{o2})$		.0160	.0160		.0123	.0155
$se(\hat{\Gamma}_0)$		.0157	.0127		.0152	.0191
$se(\hat{\Gamma}_1)$		.0092	.0185		.0253	.0321
$se(\hat{\Gamma}_2)$		.0215	.0275		.0221	.0304
$se(\widehat{ate_{10}})$		.0173	.0420		.0099	.0092
$se(\widehat{ate_{20}})$		.0097	.0106		.0302	.0414

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.5: Model with Correlated Random Coefficients and Symmetric Instrument, N=1000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	1.0048	.3626	.4217	1.1852	.3861	.4384
$\hat{\psi}_{o1}$	1.9784	1.0168	1.0007	2.0025	.7629	.8378
$\hat{\psi}_{o2}$	2.9972	1.7117	1.5230	2.8303	1.3924	1.3015
$\hat{\Gamma}_0$	3.9989	.3458	.4340	3.9994	.4091	.5091
$\hat{\Gamma}_1$	4.9947	.4341	.5159	5.0007	.5900	.6871
$\hat{\Gamma}_2$	5.9930	.7844	.8203	5.9953	.8397	.8670
$\widehat{ate_{10}}$	1.9692	1.0704	1.1056	1.8184	.7862	.8484
$\widehat{ate_{20}}$	3.9864	1.5429	1.4001	3.6409	1.1777	1.1125
$bias(\widehat{ate_{10}})$	-.0307			-.1815		
$bias(\widehat{ate_{20}})$	-.0135			-.3590		
$se(\hat{\psi}_{o0})$		.0319	.0710		.0410	.0940
$se(\hat{\psi}_{o1})$		.1121	.2218		.0784	.1220
$se(\hat{\psi}_{o2})$		.1380	.1235		.1060	.1021
$se(\hat{\Gamma}_0)$		.0558	.0987		.0665	.1430
$se(\hat{\Gamma}_1)$		.0825	.1166		.1733	.2103
$se(\hat{\Gamma}_2)$		.1532	.1691		.1783	.1901
$se(\widehat{ate_{10}})$		.1019	.2119		.0614	.0747
$se(\widehat{ate_{20}})$		.1119	.0981		.2098	.2682

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.6: Model with Correlated Random Coefficients and Symmetric Instrument, N=2000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9999	.2632	.3012	1.1856	.3106	.3119
$\hat{\psi}_{o1}$	1.9787	.7716	.7118	1.9991	.5969	.5887
$\hat{\psi}_{o2}$	2.9903	1.0527	1.0615	2.8146	.9205	.9192
$\hat{\Gamma}_0$	3.9979	.3050	.3080	3.9972	.3552	.3574
$\hat{\Gamma}_1$	5.0001	.3452	.3666	5.0026	.4424	.4791
$\hat{\Gamma}_2$	6.0027	.5958	.5686	6.0055	.6699	.6063
$\widehat{ate_{10}}$	1.9810	.8465	.7882	1.8189	.6158	.5946
$\widehat{ate_{20}}$	3.9952	.9880	.9801	3.6373	.7780	.7825
$bias(\widehat{ate_{10}})$	-.0189			-.1810		
$bias(\widehat{ate_{20}})$	-.0047			-.3626		
$se(\hat{\psi}_{o0})$		.0235	.0434		.0312	.0564
$se(\hat{\psi}_{o1})$		.0812	.1151		.0369	.0636
$se(\hat{\psi}_{o2})$		.0510	.0638		.0436	.0556
$se(\hat{\Gamma}_0)$		.0550	.0595		.0656	.0818
$se(\hat{\Gamma}_1)$		.0482	.0689		.0586	.1133
$se(\hat{\Gamma}_2)$		.0731	.0954		.0815	.1060
$se(\widehat{ate_{10}})$		.0784	.1097		.0377	.0361
$se(\widehat{ate_{20}})$		.0440	.0479		.1054	.1525

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.7: Model with Correlated Random Coefficients and Symmetric Instrument, N=5000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9999	.1771	.1919	1.1819	.1778	.1967
$\hat{\psi}_{o1}$	1.9967	.4387	.4444	2.0052	.3864	.3676
$\hat{\psi}_{o2}$	2.9915	.6408	.6680	2.8216	.5644	.5725
$\hat{\Gamma}_0$	4.0007	.1550	.1940	4.0008	.1780	.2239
$\hat{\Gamma}_1$	4.9971	.2730	.2300	5.0006	.3169	.2989
$\hat{\Gamma}_2$	6.0043	.3523	.3571	6.0028	.3806	.3797
$\widehat{ate_{10}}$	1.9931	.5169	.4912	1.8230	.3802	.3730
$\widehat{ate_{20}}$	3.9951	.6033	.6114	3.6416	.4891	.4848
$bias(\widehat{ate_{10}})$	-.0068			-.1769		
$bias(\widehat{ate_{20}})$	-.0048			-.3583		
$se(\hat{\psi}_{o0})$		.0082	.0215		.0102	.0263
$se(\hat{\psi}_{o1})$		.0319	.0493		.0181	.0285
$se(\hat{\psi}_{o2})$		.0204	.0268		.0203	.0244
$se(\hat{\Gamma}_0)$		.0138	.0284		.0168	.0367
$se(\hat{\Gamma}_1)$		.0262	.0320		.0304	.0508
$se(\hat{\Gamma}_2)$		.0423	.0430		.0485	.0480
$se(\widehat{ate_{10}})$		.0314	.0468		.0135	.0141
$se(\widehat{ate_{20}})$		.0172	.0191		.0581	.0683

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.8: Model with Correlated Random Coefficients and Symmetric Instrument, N=10000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	1.0000	.1395	.1352	1.1819	.1364	.1412
$\hat{\psi}_{o1}$	1.9957	.3210	.3159	2.0030	.2636	.2625
$\hat{\psi}_{o2}$	2.9998	.4701	.4757	2.8244	.4104	.4103
$\hat{\Gamma}_0$	4.0007	.1362	.1379	4.0010	.1491	.1599
$\hat{\Gamma}_1$	4.9976	.1701	.1634	4.9973	.2001	.2117
$\hat{\Gamma}_2$	5.9997	.2320	.2514	6.0004	.2445	.2679
$\widehat{ate_{10}}$	1.9927	.3454	.3516	1.8174	.2581	.2628
$\widehat{ate_{20}}$	3.9988	.4324	.4389	3.6418	.3480	.3512
$bias(\widehat{ate_{10}})$	-.0072			-.1825		
$bias(\widehat{ate_{20}})$	-.0011			-.3581		
$se(\hat{\psi}_{o0})$		.0143	.0126		.0102	.0150
$se(\hat{\psi}_{o1})$		.0210	.0260		.0088	.0143
$se(\hat{\psi}_{o2})$		.0107	.0141		.0095	.0131
$se(\hat{\Gamma}_0)$		.0143	.0162		.0124	.0204
$se(\hat{\Gamma}_1)$		.0142	.0175		.0144	.0264
$se(\hat{\Gamma}_2)$		.0144	.0234		.0169	.0259
$se(\widehat{ate_{10}})$		.0183	.0247		.0059	.0069
$se(\widehat{ate_{20}})$		.0099	.0095		.0230	.0374

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.9: Model with Correlated Random Coefficients, Misspecification, Asymmetric Instrument, N=1000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9461	1.1410	1.2723	1.4027	.4736	.5037
$\hat{\psi}_{o1}$	1.5961	2.0282	2.7527	1.6021	1.0796	1.1983
$\hat{\psi}_{o2}$	4.9293	2.8969	2.2047	3.0860	2.1529	2.1102
$\hat{\Gamma}_0$	4.0019	.3766	.2769	4.0088	1.1592	1.3138
$\hat{\Gamma}_1$	4.9902	.3993	.3731	4.9867	2.4190	2.9119
$\hat{\Gamma}_2$	5.9875	.8826	.5114	5.9920	1.1375	1.5279
$\widehat{ate_{10}}$	1.6196	1.5444	2.0207	1.1407	1.0203	.8939
$\widehat{ate_{20}}$	5.9849	4.5498	4.5059	3.6704	1.3177	1.0434
$bias(\widehat{ate_{10}})$	-.3803			-.8592		
$bias(\widehat{ate_{20}})$	1.9849			-.3295		
$se(\hat{\psi}_{o0})$		.1203	.3022		.1654	.2450
$se(\hat{\psi}_{o1})$		.1422	.3879		.3317	.4835
$se(\hat{\psi}_{o2})$		.3846	.3618		.8319	1.6053
$se(\hat{\Gamma}_0)$		.0749	.0556		1.7195	3.3484
$se(\hat{\Gamma}_1)$		.0893	.0769		4.1807	6.9957
$se(\hat{\Gamma}_2)$		.2508	.0931		1.0852	3.1281
$se(\widehat{ate_{10}})$		.0918	.2787		.4833	.3081
$se(\widehat{ate_{20}})$		.3983	.8435		1.4886	3.1142

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.10: Model with Correlated Random Coefficients, Misspecification, Asymmetric Instrument, N=2000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9287	.8776	.9087	1.4031	.3142	.3414
$\hat{\psi}_{o1}$	1.5697	2.0558	1.9266	1.5992	.7486	.8057
$\hat{\psi}_{o2}$	5.0589	1.6261	1.5390	3.0892	1.2035	1.3993
$\hat{\Gamma}_0$	4.0001	.1551	.1987	4.0046	.6245	.8184
$\hat{\Gamma}_1$	4.9958	.3668	.2645	4.9881	2.5550	1.8173
$\hat{\Gamma}_2$	5.9928	.3897	.3578	5.9977	.8872	.9898
$\widehat{ate_{10}}$	1.6532	1.4764	1.4045	1.1761	.6393	.5918
$\widehat{ate_{20}}$	6.1530	3.0933	3.1422	3.6827	.6620	.6659
$bias(\widehat{ate_{10}})$	-.3467			-.8238		
$bias(\widehat{ate_{20}})$	2.1530			-.3172		
$se(\hat{\psi}_{o0})$		.0745	.1603		.0533	.0912
$se(\hat{\psi}_{o1})$		.1144	.1968		.1379	.1830
$se(\hat{\psi}_{o2})$		.1617	.1897		.1586	.3747
$se(\hat{\Gamma}_0)$		.0167	.0331		.2910	.8311
$se(\hat{\Gamma}_1)$		.0389	.0448		1.2377	1.7261
$se(\hat{\Gamma}_2)$		.0365	.0522		.3916	.7780
$se(\widehat{ate_{10}})$		.0720	.1417		.1356	.1116
$se(\widehat{ate_{20}})$		.2153	.4445		.2736	.6014

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.11: Model with Correlated Random Coefficients, Misspecification, Asymmetric Instrument, N=5000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.8966	.4653	.5729	1.4067	.2315	.2000
$\hat{\psi}_{o1}$	1.5689	1.1492	1.2345	1.5912	.5330	.4752
$\hat{\psi}_{o2}$	5.1376	.8912	.9638	3.0910	.9123	.8222
$\hat{\Gamma}_0$	3.9991	.1160	.1266	3.9969	.7580	.4583
$\hat{\Gamma}_1$	4.9961	.1627	.1676	5.0041	1.5389	1.0147
$\hat{\Gamma}_2$	5.9979	.2006	.2263	5.9940	.7654	.5656
$\widehat{ate_{10}}$	1.6626	.8186	.9052	1.1816	.3769	.3542
$\widehat{ate_{20}}$	6.2447	1.6778	1.9784	3.6827	.4266	.3931
$bias(\widehat{ate_{10}})$	-.3373			-.8183		
$bias(\widehat{ate_{20}})$	2.2447			-.3172		
$se(\hat{\psi}_{o0})$		.0613	.0715		.0600	.0291
$se(\hat{\psi}_{o1})$		.0417	.0786		.0981	.0583
$se(\hat{\psi}_{o2})$		.0696	.0827		.1465	.0885
$se(\hat{\Gamma}_0)$		.0094	.0160		.4003	.1887
$se(\hat{\Gamma}_1)$		.0121	.0211		.7224	.3914
$se(\hat{\Gamma}_2)$		.0226	.0237		.3021	.1806
$se(\widehat{ate_{10}})$		.0289	.0573		.0485	.0349
$se(\widehat{ate_{20}})$		.1387	.1990		.2000	.1042

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.12: Model with Correlated Random Coefficients, Misspecification, Asymmetric Instrument, N=10000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9182	.3991	.4102	1.4081	.1619	.1376
$\hat{\psi}_{o1}$	1.5635	.9078	.8644	1.5939	.3615	.3257
$\hat{\psi}_{o2}$	5.1640	.6609	.6831	3.0917	.6498	.5702
$\hat{\Gamma}_0$	3.9998	.0934	.0898	3.9976	.4809	.3100
$\hat{\Gamma}_1$	4.9986	.1156	.1189	5.0024	.9661	.6872
$\hat{\Gamma}_2$	5.9994	.1680	.1595	5.9973	.5499	.3867
$\widehat{ate_{10}}$	1.6472	.6604	.6297	1.1786	.2635	.2438
$\widehat{ate_{20}}$	6.2517	1.3442	1.4030	3.6831	.2873	.2768
$bias(\widehat{ate_{10}})$	-.3527			-.8213		
$bias(\widehat{ate_{20}})$	2.2517			-.3168		
$se(\hat{\psi}_{o0})$		.0259	.0400		.0506	.0133
$se(\hat{\psi}_{o1})$		.0323	.0400		.0543	.0274
$se(\hat{\psi}_{o2})$		.0292	.0440		.0813	.0400
$se(\hat{\Gamma}_0)$		.0090	.0089		.3922	.0808
$se(\hat{\Gamma}_1)$		.0130	.0116		.4640	.1707
$se(\hat{\Gamma}_2)$		.0197	.0128		.1716	.0805
$se(\widehat{ate_{10}})$		.0207	.0296		.0277	.0162
$se(\widehat{ate_{20}})$		.0716	.1129		.2243	.0420

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.13: Model with Correlated Random Coefficients, Misspecification, Symmetric Instrument, N=1000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	1.0018	.3845	.3526	1.0872	.6492	.4221
$\hat{\psi}_{o1}$	1.6606	2.1580	2.1724	1.8613	.5639	.6317
$\hat{\psi}_{o2}$	3.2989	1.7881	1.9619	3.0133	1.4502	1.3144
$\hat{\Gamma}_0$	3.9996	.2767	.2475	3.9994	1.6010	.5225
$\hat{\Gamma}_1$	4.9916	.2989	.3493	5.0007	3.6601	2.0802
$\hat{\Gamma}_2$	5.9948	.5331	.4996	5.9950	2.8154	1.5871
$\widehat{ate_{10}}$	1.6584	1.7267	1.7348	1.7727	.5862	.4716
$\widehat{ate_{20}}$	4.2994	1.7651	1.9578	3.9298	.9728	.9005
$bias(\widehat{ate_{10}})$	-.3415			-.2272		
$bias(\widehat{ate_{20}})$	.2994			-.0701		
$se(\hat{\psi}_{o0})$		.0446	.0571		.2017	.1696
$se(\hat{\psi}_{o1})$		.2181	.3373		.2615	.2193
$se(\hat{\psi}_{o2})$		.1588	.3299		.7145	.4557
$se(\hat{\Gamma}_0)$		.0402	.0484		.7055	.6112
$se(\hat{\Gamma}_1)$		.0642	.0724		3.4152	3.7772
$se(\hat{\Gamma}_2)$		.0713	.0944		2.3872	2.6757
$se(\widehat{ate_{10}})$		.1761	.2670		.4531	.1380
$se(\widehat{ate_{20}})$		.1537	.3287		.7081	.9806

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.14: Model with Correlated Random Coefficients, Misspecification, Symmetric Instrument, N=2000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9960	.2807	.2514	1.0906	.3491	.2816
$\hat{\psi}_{o1}$	1.6299	1.5503	1.5333	1.8668	.6082	.4177
$\hat{\psi}_{o2}$	3.3335	1.5427	1.3816	3.0193	1.1811	.8750
$\hat{\Gamma}_0$	4.0002	.1557	.1768	3.9995	.4435	.3454
$\hat{\Gamma}_1$	4.9937	.2193	.2476	4.9937	1.9177	1.3231
$\hat{\Gamma}_2$	5.9928	.3564	.3531	5.9912	1.3576	1.0265
$\widehat{ate_{10}}$	1.6279	1.2361	1.2241	1.7631	.3395	.3183
$\widehat{ate_{20}}$	4.3370	1.5352	1.3837	3.9303	.7097	.6005
$bias(\widehat{ate_{10}})$	-.3720			-.2368		
$bias(\widehat{ate_{20}})$	.3370			-.0696		
$se(\hat{\psi}_{o0})$		.0261	.0304		.0185	.0608
$se(\hat{\psi}_{o1})$		.1602	.1750		.1178	.0888
$se(\hat{\psi}_{o2})$		.1080	.1676		.2395	.1691
$se(\hat{\Gamma}_0)$		.0217	.0289		.0654	.1816
$se(\hat{\Gamma}_1)$		.0273	.0419		.8022	1.0389
$se(\hat{\Gamma}_2)$		.0493	.0530		.5274	.7355
$se(\widehat{ate_{10}})$		.1264	.1392		.0463	.0530
$se(\widehat{ate_{20}})$		.1035	.1672		.2598	.3569

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.15: Model with Correlated Random Coefficients, Misspecification, Symmetric Instrument, N=5000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9968	.1632	.1606	1.0898	.1727	.1682
$\hat{\psi}_{o1}$	1.6423	.9970	.9633	1.8780	.2419	.2494
$\hat{\psi}_{o2}$	3.3268	.9881	.8717	3.0120	.5201	.5197
$\hat{\Gamma}_0$	4.0005	.1253	.1133	3.9986	.2132	.2023
$\hat{\Gamma}_1$	4.9984	.1355	.1565	5.0057	.6606	.7635
$\hat{\Gamma}_2$	5.9990	.2729	.2225	5.9934	.5288	.5934
$\widehat{ate_{10}}$	1.6422	.7958	.7680	1.7818	.1913	.1873
$\widehat{ate_{20}}$	4.3286	.9701	.8700	3.9251	.3747	.3623
$bias(\widehat{ate_{10}})$	-.3577			-.2181		
$bias(\widehat{ate_{20}})$	.3286			-.0748		
$se(\hat{\psi}_{o0})$		.0107	.0130		.0181	.0196
$se(\hat{\psi}_{o1})$		.0527	.0734		.0226	.0296
$se(\hat{\psi}_{o2})$		.0431	.0668		.0396	.0552
$se(\hat{\Gamma}_0)$		.0201	.0138		.0436	.0497
$se(\hat{\Gamma}_1)$		.0125	.0198		.1721	.2697
$se(\hat{\Gamma}_2)$		.0214	.0241		.1081	.1931
$se(\widehat{ate_{10}})$		.0420	.0585		.0169	.0170
$se(\widehat{ate_{20}})$		.0426	.0668		.0600	.1171

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

Table B.16: Model with Correlated Random Coefficients, Misspecification, Symmetric Instrument, N=10000, and I=10000

	CF Approach			IV Approach		
	Estimate	BS. SE	M.C. SD	Estimate	BS. SE	M.C. SD
$\hat{\psi}_{o0}$	.9956	.1190	.1134	1.0911	.1063	.1173
$\hat{\psi}_{o1}$	1.6620	.6463	.6849	1.8662	.1499	.1725
$\hat{\psi}_{o2}$	3.3391	.5993	.6144	3.0261	.3323	.3619
$\hat{\Gamma}_0$	4.0006	.0878	.0801	3.9990	.1280	.1401
$\hat{\Gamma}_1$	4.9986	.1270	.1105	5.0050	.4043	.5244
$\hat{\Gamma}_2$	5.9973	.1518	.1573	5.9919	.3469	.4090
$\widehat{ate_{10}}$	1.6654	.5150	.5514	1.7798	.1150	.1315
$\widehat{ate_{20}}$	4.3411	.6041	.6154	3.9312	.2490	.2519
$bias(\widehat{ate_{10}})$	-.3345			-.2201		
$bias(\widehat{ate_{20}})$	.3411			-.0687		
$se(\hat{\psi}_{o0})$		.0083	.0067		.0072	.0094
$se(\hat{\psi}_{o1})$		.0241	.0369		.0089	.0139
$se(\hat{\psi}_{o2})$		.0238	.0338		.0181	.0261
$se(\hat{\Gamma}_0)$		.0100	.0076		.0141	.0230
$se(\hat{\Gamma}_1)$		.0134	.0109		.0628	.1220
$se(\hat{\Gamma}_2)$		.0144	.0131		.0502	.0884
$se(\widehat{ate_{10}})$		.0189	.0294		.0059	.0079
$se(\widehat{ate_{20}})$		.0239	.0339		.0353	.0564

Note: N=Sample Size. I=Number of Iteration.
CF= Control Function. IV= Instrumental Variable.
BS. SE=Bootstrapped Standard Error.
M.C. SD= Monte Carlo Standard Deviation.

APPENDIX C

APPENDIX FOR CHAPTER 3

C.1 Chapter 3: Tables-Simulations

Table C.1: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 4$, $p = 2100$, $p' = 27$, $f = 15$, N=1000, Correlated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	1.0574	.8641	1.0467	.8215	1.0334	.8106	1.0308	.8678	.9958	1.2182
$\widehat{ate_{2,0}}$	1.9784	1.5510	1.9936	1.0332	1.9716	1.0220	1.9777	1.0807	1.9666	1.3507
$bias(\widehat{ate_{1,0}})$	.0574		.0467		.0334		.0308		-.0041	
$bias(\widehat{ate_{2,0}})$	-.0215		-.0063		-.0283		-.0222		-.0333	
# of SV	26.375		25.941		25.941		25.942		-	
# of CSV	11.300		10.933		10.933		10.932		-	
$MAPE$	13.382		13.319		13.315		7.8851		11.533	
$RMSE$	17.839		17.772		17.769		12.409		15.740	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.2: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 4$, $p = 2100$, $p' = 27$, $f = 15$, N=1250, Correlated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	1.0669	.7594	1.0240	.7177	1.0142	.7108	1.0105	.7260	1.0043	1.0938
$\widehat{ate_{2,0}}$	2.0192	1.2388	2.0518	.9188	2.0335	.9109	2.0483	.9656	2.0742	1.2079
$bias(\widehat{ate_{1,0}})$	.0669		.0240		.0142		.0105		.0043	
$bias(\widehat{ate_{2,0}})$	.0192		.0518		.0335		.0483		.0742	
# of SV	26.602		26.254		26.254		26.253		-	
# of CSV	11.533		11.245		11.245		11.244		-	
$MAPE$	13.355		13.323		13.320		7.8101		11.552	
$RMSE$	17.801		17.770		17.767		12.301		15.768	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.3: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 4$, $p = 2100$, $p' = 27$, $f = 15$, N=1500, Correlated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	.9334	.7097	.9491	.6820	.9411	.6765	.9444	.6900	.9197	1.0111
$\widehat{ate_{2,0}}$	1.9567	1.0555	1.9327	.8343	1.9180	.8277	1.9126	.8860	1.8759	1.1009
$bias(\widehat{ate_{1,0}})$	-.0665		-.0508		-.0588		-.0555		-.0802	
$bias(\widehat{ate_{2,0}})$	-.0432		-.0672		-.0819		-.0873		-.1240	
# of SV	26.777		26.491		26.491		26.484		-	
# of CSV	11.713		11.487		11.487		11.482		-	
$MAPE$	13.318		13.288		13.286		7.7574		11.561	
$RMSE$	17.775		17.743		17.741		12.234		15.777	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.4: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 4$, $p = 2100$, $p' = 27$, $f = 15$, N=2000, Correlated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	.9457	.5676	.9594	.5536	.9531	.5498	.9575	.5750	.9473	.8948
$\widehat{ate_{2,0}}$	1.9408	.8869	1.9617	.7436	1.9501	.7392	1.9634	.7578	1.9691	.9272
$bias(\widehat{ate_{1,0}})$	-.0542		-.0405		-.0468		-.0424		-.0526	
$bias(\widehat{ate_{2,0}})$	-.0591		-.0382		-.0498		-.0365		-.0308	
# of SV	26.964		26.776		26.776		26.773		-	
# of CSV	11.893		11.773		11.773		11.772		-	
$MAPE$	13.298		13.289		13.288		7.6724		11.574	
$RMSE$	17.750		17.740		17.739		12.094		15.790	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.5: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 1$, $p = 2100$, $p' = 18$, $f = 15$, N=1000, Correlated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	1.0687	.8674	1.0519	.8176	1.0485	.8169	1.0562	.8211	1.0581	.8194
$\widehat{ate_{2,0}}$	1.9977	1.5211	1.9974	1.0195	1.9941	1.0182	1.9975	1.0214	1.9972	1.0232
$bias(\widehat{ate_{1,0}})$	.0687		.0519		.0485		.0562		.0581	
$bias(\widehat{ate_{2,0}})$	-.0022		-.0025		-.0058		-.0024		-.0027	
# of SV	16.214		16.101		16.101		16.099		-	
# of CSV	1.1850		1.0990		1.0990		1.0980		-	
$MAPE$	10.510		10.431		10.431		7.5310		7.6320	
$RMSE$	14.473		14.402		14.401		11.861		11.879	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.6: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 1$, $p = 2100$, $p' = 18$, $f = 15$, N=1250, Correlated **h**, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	1.0589	.7473	1.0158	.7184	1.0134	.7183	1.0249	.7128	1.0255	.7146
$\widehat{ate_{2,0}}$	2.0210	1.2399	2.0506	.9152	2.0479	.9143	2.0512	.9229	2.0509	.9273
$bias(\widehat{ate_{1,0}})$	.0589		.0158		.0134		.0249		.0255	
$bias(\widehat{ate_{2,0}})$	.0210		.0506		.0479		.0512		.0509	
# of SV	16.437		16.265		16.265		16.253		-	
# of CSV	1.3973		1.2622		1.2622		1.2512		-	
$MAPE$	10.471		10.430		10.430		7.5390		7.6456	
$RMSE$	14.437		14.397		14.397		11.880		11.903	

Note: l_{h_1}= # of Nonzero Variables in **h**. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.7: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 1$, $p = 2100$, $p' = 18$, $f = 15$, N=1500, Correlated **h**, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	.9434	.7025	.9618	.6682	.9597	.6678	.9562	.6790	.9518	.6770
$\widehat{ate_{2,0}}$	1.9527	1.0737	1.9273	.8452	1.9250	.8443	1.9293	.8445	1.9304	.8461
$bias(\widehat{ate_{1,0}})$	-.0565		-.0381		-.0402		-.0437		-.0481	
$bias(\widehat{ate_{2,0}})$	-.0472		-.0726		-.0749		-.0706		-.0695	
# of SV	16.645		16.415		16.415		16.411		-	
# of CSV	1.6110		1.4170		1.4170		1.4100		-	
$MAPE$	10.439		10.402		10.401		7.5440		7.6557	
$RMSE$	14.425		14.388		14.388		11.900		11.928	

Note: l_{h_1}= # of Nonzero Variables in **h**. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.8: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 1$, $p = 2100$, $p' = 18$, $f = 15$, N=2000, Correlated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	.9440	.5737	.9475	.5489	.9461	.5483	.9580	.5714	.9619	.5794
$\widehat{ate_{2,0}}$	1.9386	.8904	1.9605	.7466	1.9584	.7459	1.9601	.7510	1.9624	.7528
$bias(\widehat{ate_{1,0}})$	-.0559		-.0524		-.0538		-.0419		-.0380	
$bias(\widehat{ate_{2,0}})$	-.0613		-.0394		-.0415		-.0398		-.0375	
# of SV	17.159		16.878		16.878		16.869		-	
# of CSV	2.1081		1.8798		1.8798		1.8748		-	
$MAPE$	10.402		10.392		10.392		7.5283		7.6527	
$RMSE$	14.373		14.364		14.364		11.873		11.914	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.9: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 4$, $p = 2100$, $p' = 27$, $f = 15$, N=1000, Uncorrelated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	1.0801	.9276	1.0411	.8367	1.0288	.8295	1.0262	.9350	1.0364	1.0547
$\widehat{ate_{2,0}}$	2.0061	1.6753	1.9714	1.0833	1.9552	1.0752	1.9766	1.1150	1.9921	1.2047
$bias(\widehat{ate_{1,0}})$	.0801		.0411		.0288		.0262		.0364	
$bias(\widehat{ate_{2,0}})$	.0061		-.0285		-.0447		-.0233		-.0078	
# of SV	23.417		22.871		22.871		22.852		-	
# of CSV	8.3620		7.8676		7.8676		7.8525		-	
$MAPE$	12.039		11.955		11.952		8.2623		9.9062	
$RMSE$	16.212		16.131		16.129		12.934		13.892	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.10: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 4$, $p = 2100$, $p' = 27$, $f = 15$, N=1250, Uncorrelated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	1.0771	.7486	1.0501	.7200	1.0400	.7146	.9985	.7692	.9787	.9376
$\widehat{ate_{2,0}}$	2.0233	1.3014	2.0385	.9446	2.0241	.9383	2.0522	.9919	2.0675	1.0649
$bias(\widehat{ate_{1,0}})$	.0771		.0501		.0400		-.0014		-.0212	
$bias(\widehat{ate_{2,0}})$	.0233		.0385		.0241		.0522		.0675	
# of SV	24.240		23.755		23.755		23.743		-	
# of CSV	9.2086		8.7572		8.7572		8.7472		-	
$MAPE$	11.994		11.953		11.951		8.1502		9.9246	
$RMSE$	16.165		16.123		16.121		12.790		13.919	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.11: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 4$, $p = 2100$, $p' = 27$, $f = 15$, N=1500, Uncorrelated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	.9513	.7127	.9561	.6839	.9485	.6792	.9310	.7240	.8889	.8623
$\widehat{ate_{2,0}}$	1.9390	1.1272	1.9145	.8748	1.9025	.8694	1.9048	.9230	1.8970	.9896
$bias(\widehat{ate_{1,0}})$	-.0486		-.0438		-.0514		-.0689		-.1110	
$bias(\widehat{ate_{2,0}})$	-.0609		-.0854		-.0974		-.0951		-.1029	
# of SV	24.822		24.386		24.386		24.380		-	
# of CSV	9.7933		9.3941		9.3941		9.3901		-	
$MAPE$	11.958		11.924		11.922		8.0679		9.9344	
$RMSE$	16.142		16.107		16.105		12.683		13.932	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.12: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 4$, $p = 2100$, $p' = 27$, $f = 15$, N=2000, Uncorrelated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	.9425	.5692	.9549	.5548	.9486	.5513	.9611	.5879	.9554	.7198
$\widehat{ate_{2,0}}$	1.9349	.9067	1.9520	.7481	1.9417	.7442	1.9643	.7734	1.9761	.8275
$bias(\widehat{ate_{1,0}})$	-.0574		-.0450		-.0513		-.0388		-.0445	
$bias(\widehat{ate_{2,0}})$	-.0650		-.0479		-.0582		-.0356		-.0238	
# of SV	25.809		25.480		25.480		25.482		-	
# of CSV	10.788		10.485		10.485		10.484		-	
$MAPE$	11.929		11.917		11.916		7.9188		9.9373	
$RMSE$	16.103		16.093		16.092		12.459		13.933	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.13: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 1$, $p = 2100$, $p' = 18$, $f = 15$, N=1000, Uncorrelated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	1.0685	.8753	1.0493	.8199	1.0460	.8193	1.0591	.8221	1.0606	.8212
$\widehat{ate_{2,0}}$	1.9907	1.5179	1.9988	1.0231	1.9957	1.0217	1.9971	1.0251	1.9966	1.0273
$bias(\widehat{ate_{1,0}})$	.0685		.0493		.0460		.0591		.0606	
$bias(\widehat{ate_{2,0}})$	-.0092		-.0011		-.0042		-.0028		-.0033	
# of SV	16.200		16.095		16.095		16.091		-	
# of CSV	1.1831		1.0990		1.0990		1.0990		-	
$MAPE$	10.509		10.431		10.431		7.5307		7.6317	
$RMSE$	14.472		14.401		14.401		11.861		11.879	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.14: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 1$, $p = 2100$, $p' = 18$, $f = 15$, N=1250, Uncorrelated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	1.0590	.7406	1.0176	.7175	1.0152	.7174	1.0279	.7110	1.0289	.7123
$\widehat{ate_{2,0}}$	2.0256	1.2428	2.0513	.9137	2.0487	.9128	2.0518	.9219	2.0513	.9269
$bias(\widehat{ate_{1,0}})$	.0590		.0176		.0152		.0279		.0289	
$bias(\widehat{ate_{2,0}})$	.0256		.0513		.0487		.0518		.0513	
# of SV	16.410		16.261		16.261		16.250		-	
# of CSV	1.3957		1.2625		1.2625		1.2555		-	
$MAPE$	10.473		10.431		10.431		7.5405		7.6467	
$RMSE$	14.439		14.399		14.398		11.883		11.905	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.15: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 1$, $p = 2100$, $p' = 18$, $f = 15$, N=1500, Uncorrelated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	.9475	.6968	.9581	.6682	.9560	.6678	.9536	.6782	.9501	.6764
$\widehat{ate_{2,0}}$	1.9547	1.0812	1.9261	.8480	1.9239	.8471	1.9281	.8474	1.9297	.8499
$bias(\widehat{ate_{1,0}})$	-.0524		-.0418		-.0439		-.0463		-.0498	
$bias(\widehat{ate_{2,0}})$	-.0452		-.0738		-.0760		-.0718		-.0702	
# of SV	16.622		16.403		16.403		16.392		-	
# of CSV	1.6078		1.4132		1.4132		1.4022		-	
$MAPE$	10.440		10.403		10.402		7.5455		7.6566	
$RMSE$	14.427		14.390		14.390		11.902		11.930	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

Table C.16: High Dimensional Sparse Model with Heterogeneous Counterfactual Errors $l_{h_1} = 1$, $p = 2100$, $p' = 18$, $f = 15$, N=2000, Uncorrelated $\mathbf{h}$, and I=1000

	XPO		DS		PO		LASSO		CF	
	Est.	SD	Est.	SD	Est.	SD	Est.	SD	Est.	SD
$\widehat{ate_{1,0}}$	.9468	.5714	.9525	.5487	.9510	.5480	.9610	.5702	.9647	.5780
$\widehat{ate_{2,0}}$	1.9378	.8879	1.9576	.7447	1.9556	.7439	1.9572	.7491	1.9606	.7528
$bias(\widehat{ate_{1,0}})$	-.0531		-.0474		-.0489		-.0389		-.0352	
$bias(\widehat{ate_{2,0}})$	-.0621		-.0423		-.0443		-.0427		-.0393	
# of SV	17.113		16.869		16.869		16.864		-	
# of CSV	2.1043		1.8796		1.8796		1.8746		-	
$MAPE$	10.401		10.390		10.390		7.5267		7.6523	
$RMSE$	14.372		14.363		14.362		11.870		11.913	

Note: l_{h_1}= # of Nonzero Variables in $\mathbf{h}$. p= # of Potential Variables. p' = # of Variables with Nonzero Coefficients. f= # of Forced Variables. N= Sample Size. I= # of Iterations. Est.=Estimate.
XPO= Cross Partial-Out Estimation. DS= Double Selection Estimation. PO= Partial-Out Estimation.
CF= OLS Using Only Forced Variables. SD= Monte Carlo Standard Deviation.
SV= # of Selected Variables. CSV= # of Correctly Selected Variables.
MAPE= Mean Absolute Prediction Error. RMSE= Root Mean Square Error.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadie, A., and M. D. Cattaneo (2018), "Econometric Methods for Program Evaluation," Annual Review of Economics 10, 647-667.

Akin, J. S., D. K. Guilkey, and R. Sickles (1979), "A Random Coefficient Probit Model with an Application to a Study of Migration," Journal of Econometrics 11, 233-246.

Akresh R. and I. R. Akresh (2011), "Using Achievement Tests to Measure Language Assimilation and Language Bias among the Children of Immigrants," Journal of Human Resources 46, 441-462.

Aldashev, A., J. Gernandt, and S. L. Thomsen (2009), "Language Usage, Participation, Employment and Earnings Evidence for Foreigners in West Germany with Multiple Sources of Selection," Labour Economics 16, 330–341.

Amann, R. A., and T. J. Klein (2012), "Returns to Type or Tenure?," Journal of the Royal Statistical Society: Series A (Statistics in Society) 175, 153-166.

Amemiya, T. (1981), "Qualitative Response Models: A Survey," Journal of Economic Literature 19, 1483-1536.

Amemiya, T. (1985), Advanced Econometrics. Cambridge, Mass: Harvard University Press.

Angrist, J. D. (1991), "Instrumental Variables Estimation of Average Treatment Effects in Econometrics and Epidemiology," National Bureau of Economic Research Technical Working Paper Number 115.

Angrist, J. D., and G. W. Imbens (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," Journal of the American Statistical Association 90, 431-442.

Angrist, J. D., G. W. Imbens, and D. Rubin (1996), "Identification and Casual Effects Using Instrumental Variables," Journal of the American Statistical Association 91, 444-455.

Appelt, S. (2015), "Authorized Generic Entry prior to Patent Expiry: Reassessing for Independent Generic Entry," The Review of Economics and Statistics 97, 654-666.

Athey, S. (2018), "The Impact of Machine Learning on Economics," Retrieved from Stanford Graduate School of Business website:
https://www.gsb.stanford.edu/sites/default/files/publication-pdf/atheyimpactmlecon.pdf

Athey, S., and G. W. Imbens (2019), "Machine Learning Methods that Economists Should Know about," Annual Review of Economics 11, 685-725.

Athey, S., and M. Luca (2019), "Economists (and Economics) in Tech Companies," Journal of Economic Perspectives 33, 209–230.

Athey, S., G. W. Imbens, and S. Wager (2018), "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions," Journal of the Royal Statistical Society. Series B. Statistical Methodology 80, 597-623.

Athey S., J. Tibshirani, and S. Wager (2019), "Generalized Random Forests," The Annals of Statistics 47, 1148–1178.

Athey, S., M. Bayati, N. Doudchenko, G. W. Imbens, and K. Khosravi (2021), "Matrix Completion Methods for Causal Panel Data Models (arXiv No: 1710.10251)," Retrieved from arXiv website: https://arxiv.org/pdf/1710.10251.pdf

Bach, P., S. Klaassen, J. Kueck, and M. Spindler (2020), "Uniform Inference in High-Dimensional Generalized Additive Models (arXiv No: 2004.01623)," Retrieved from arXiv website: https://arxiv.org/pdf/2004.01623.pdf

Bartle, R. G. (1964), The Elements of Real Analysis. New York: Wiley.

Becker, G. S. (1967), Human Capital and the Personal Distribution of Income: An Analytical Approach. W. S. Woytinsky Lecture 1. Ann Arbor : Institute of Public Administration, University of Michigan.

Becker, G. S., and B. R. Chiswick (1966), "Education and the Distribution of Earnings," American Economic Review 56, 358-369.

Bekker, P., and T. Wansbeek (2001), Identification in Parametric Models. In B. H. Baltagi (ed.) A Companion to Theoretical Econometrics. Malden, Mass: Blackwell Publishers, ch.7.

Belloni, A., and V. Chernozhukov (2011a), "High Dimensional Sparse Econometric Models: An Introduction (arXiv No: 1106.5242)," Retrieved from arXiv website: https://arxiv.org/pdf/1106.5242.pdf

Belloni, A., and V. Chernozhukov (2011b), "ℓ1-penalized Quantile Regression in High-dimensional Sparse Models," The Annals of Statistics 39, 82-130.

Belloni, A., and V. Chernozhukov (2013), "Least Squares after Model Selection in High-dimensional Sparse Models," Bernoulli 19, 521-547.

Belloni, A., A. Kaul, and M. Rosenbaum (2019), "Pivotal Estimation via Self-Normalization for High-Dimensional Linear Models with Error in Variables (arXiv No: 1708.08353)," Retrieved from arXiv website: https://arxiv.org/pdf/1708.08353.pdf

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," Econometrica 80, 2369-2429.

Belloni, A., V. Chernozhukov, and A. Kaul (2017), "Confidence Bands for Coefficients in High Dimensional Linear Models with Error-in-variables (arXiv No: 1703.00469)," Retrieved from arXiv website: https://arxiv.org/pdf/1703.00469.pdf

Belloni, A., V. Chernozhukov, and C. Hansen (2011a), "Inference for High-Dimensional Sparse Econometric Models (arXiv No: 1201.0220)," Retrieved from arXiv website: https://arxiv.org/pdf/1201.0220.pdf

Belloni, A., V. Chernozhukov, and C. Hansen (2014a), "Inference on Treatment Effects after Selection among High- Dimensional Controls," Review of Economic Studies 81, 608-650.

Belloni, A., V. Chernozhukov, and C. Hansen (2014b), "High-dimensional Methods and Inference on Structural and Treatment Effects," The Journal of Economic Perspectives 28, 29-50.

Belloni, A., V. Chernozhukov, and K. Kato (2019), "Valid Post-Selection Inference in High-Dimensional Approximately Sparse Quantile Regression Models," Journal of the American Statistical Association 114, 749-758.

Belloni, A., V. Chernozhukov, and L. Wang (2011), "Square-root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming," Biometrika 98, 791–806.

Belloni, A., V. Chernozhukov, and Y. Wei (2016), "Post-Selection Inference for Generalized Linear Models With Many Controls," Journal of Business & Economic Statistics 34, 606-619.

Belloni, A., V. Chernozhukov, C. Hansen, and D. Kozbur (2016), "Inference in High-dimensional Panel Models with an Application to Gun Control," Journal of Business and Economic Statistics 34, 590–605.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2018b), "Uniformly Valid Post-Regularization Confidence Regions for Many Functional Parameters in Z-Estimation Framework," The Annals of Statistics 46, 3643-3675.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018a), "High-Dimensional Econometrics and Regularized GMM (arXiv No: 1806.01888)," Retrieved from arXiv website: https://arxiv.org/pdf/1806.01888.pdf

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017), "Program Evaluation and Causal Inference with High-dimensional Data," Econometrica 85, 233–298.

Ben-Akiva, M., and S. R. Lerman (1985), Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, Mass: MIT Press.

Berman, E., K. Lang, and E. Siniver (2003), "Language-skill Complementarity: Returns to Immigrant Language Acquisition," Labour Economics 10, 265–290.

Billingsley, P. (1995), Probability and Measure. New York: Wiley.

Bjorklund, A., and R. Moffitt (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," The Review of Economics and Statistics 69, 42-49.

Bleakley, H. and A. Chin (2004), "Language Skills and Earnings: Evidence from Childhood Immigrants," The Review of Economics and Statistics 86, 481-496.

Bleakley, H. and A. Chin (2010), "Age at Arrival, English Proficiency, and Social Assimilation Among US Immigrants," American Economic Journal: Applied Economics 2, 165-192.

Borjas, G. J. (1985), "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants," Journal of Labor Economics 3, 463-489.

Borjas, G. J. (1995), "Assimilation and Changes in Cohort Quality Revisited: What Happened to Immigrant Earnings in the 1980s," Journal of Labor Economics 13, 201-245.

Bound, J., D. A. Jaeger, and R. M. Baker (1995), "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak" Journal of the American Statistical Association 90, 443-450.

Bourguignon, F., M. Fournier, and M. Gurgand (2007), "Selection Bias Correction Based on the Multinomial Logit Model: Monte Carlo Comparisons," Journal of Economic Surveys 21, 174-205.

Budra, S., and P. Swedberg (2012), "The Impact of Language Proficiency on Immigrants Earnings in Spain (IZA Discussion Paper No. 6957)," Retrieved from IZA, Institute of Labor Economics website: http://ftp.iza.org/dp6957.pdf

Cameron, A. C., and P. K. Trivedi (2005), Microeconometrics: Methods and Applications. New York: Cambridge University Press.

Caner, M., and A. B. Kock (2019), "High Dimensional Linear GMM (arXiv No: 1811.08779)," Retrieved from arXiv website: https://arxiv.org/pdf/1811.08779.pdf

Card, D. (2001), The Causal Effect of Education on Earnings. In D. Card and O. Ashenfelter (ed.) Handbook of Labor Economics, Vol. 3A. Amsterdam: North-Holland Publishers, ch. 30.

Carliner, G. (1981), "Wage Differences by Language Group and the Market for Language Skills in Canada," Journal of Human Resources 16, 384-399.

Chernozhukov, V., C. Hansen, and M. Spindler (2015a), "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," Annual Review of Economics 7, 649–688.

Chernozhukov, V., C. Hansen, and M. Spindler (2015b), "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments," Annual Review of Economics: Papers & Proceedings 105, 486–490.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017), "Double/Debiased/Neyman Machine Learning of Treatment Effects," American Economic Review: Papers & Proceedings 107, 261–265.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," The Econometrics Journal 21, C1–C68.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2018), "Exact and Robust Conformal Inference Methods for Predictive Machine Learning With Dependent Data," Proceedings of the 31st Conference On Learning Theory, Proceedings of Machine Learning Research 75, 732-749.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2019), "Inference on Average Treatment Effects in Aggregate Panel Data Settings (Cemmap Working Paper No. CWP32/19)," Retrieved from EconStor website: https://www.econstor.eu/handle/10419/211125

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2020), "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments (arXiv No: 1712.04802)," Retrieved from arXiv website: https://arxiv.org/pdf/1712.04802.pdf

Chernozhukov, V., W. K. Hardle, C. Huang, and W. Wang (2020), "LASSO-Driven Inference in Time and Space (arXiv No: 1806.05081)," Retrieved from arXiv website: https://arxiv.org/pdf/1806.05081.pdf

Chernozhukov, V., W. K. Newey, and R. Singh (2021), "Automatic Debiased Machine Learning of Causal and Structural Effects (arXiv No: 1809.05224)," Retrieved from arXiv website: https://arxiv.org/pdf/1809.05224.pdf

Chesher, A., and A. M. Rosen (2013), "What Do Instrumental Variable Models Deliver with Discrete Dependent Variables?," American Economic Review 103, 557-562.

Chesher, A., and A. M. Rosen (2017), "Generalized Instrumental Variable Models," Econometrica 85, 959-989.

Chiswick, B. R. (1974), Income Inequality: Regional Analyses within a Human Capital Framework. New York: Columbia University Press for NBER.

Chiswick, B. R. (1998), "Hebrew Language Usage: Determinants and Effects on Earnings among Immigrants in Israel," Journal of Population Economics 11, 253-271.

Chiswick, B. R., and G. Repetto (2000), "Immigrant Adjustment in Israel: Literacy and Fluency in Hebrew and Earnings (IZA Discussion Paper No. 177)," Retrieved from IZA, Institute of Labor Economics website: http://ftp.iza.org/dp177.pdf

Chiswick, B. R., and J. Mincer (1972), "Time-Series Changes in Personal Income Inequality in the United States from 1939, with Projections to 1985," Journal of Political Economy, 80 (3, Part 2), S34-S66.

Chiswick, B. R., and P. W. Miller (1985), "Immigrant Generation and Income in Australia," Economic Record 61, 540-553.

Chiswick, B. R., and P. W. Miller (1995), "The Endogeneity between Language and Earnings: International Analyses," Journal of Labor Economics 13, 246-288.

Chiswick, B. R., and P. W. Miller (2002), "Immigrant Earnings: Language Skills, Linguistic Concentrations and the Business Cycle," Journal of Population Economics 15, 31- 57.

Chiswick, B. R., and P. W. Miller (2003), "The Complementarity of Language and Other Human Capital: Immigrant Earnings in Canada ," Economics of Education Review 22, 469-480.

Chiswick, B. R., Y. L. Lee, and P. W. Miller (2005), "Immigrant Earnings: A Longitudinal Analysis," Review of Income and Wealth 51, 485-503.

Cochran, W. G. (1957), "Analysis of Covariance: Its Nature and Uses," Biometrics 13, 261-281.

Daganzo, C. (1979), Multinomial Probit: The Theory and Its Application to Demand Forecasting. London: Academic Press.

Dahl, G. B. (2002), "Mobility and the Returns to Education: Testing a Roy Model with Multiple Markets," Econometrica 70, 2367–2420.

Danquah, M., A. M. Iddrisu, E. O. Boakye, and S. Owusu (2021), "Do Gender Wage Differences within Households Influence Women's Empowerment and Welfare? Evidence from

Ghana (WIDER Working Paper No. 2021/40)," Retrieved from ResearchGate website: https://www.researchgate.net/profile/Michael-Danquah-5/publication/349564027

Davidson, J. (1994), Stochastic Limit Theory. Oxford: Oxford University Press.

Dhrymes, P. J. (2013), Mathematics for Econometrics. New York: Springer.

Di Paolo, A. and J. L. Raymond (2012), "Language Knowledge and Earnings in Catalonia," Journal of Applied Economics 15, 89–118.

Dixmier, J. (1984), General Topology. New York: Springer.

Donaldson, D., and A. Storeygard (2016), "The View from Above: Applications of Satellite Data in Economics," Journal of Economic Perspectives 30, 171–198.

Dong, Y. (2010), "Endogenous Regressor Binary Choice Models without Instruments, with an Application to Migration," Economics Letters 107, 33-35.

Dubin, J. A., and D. L. McFadden (1984), "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption," Econometrica 52, 345-362.

Duo, Q. (1993), The Formation of Econometrics: A Historical Perspective. New York: Oxford University Press.

Dustmann C. and F. Fabbri (2003), "Language Proficiency and Labour Market Performance of Immigrants in the UK," Economic Journal 113, 695-717.

Dustmann, C. (1994), "Speaking Fluency, Writing Fluency and Earnings of Migrants," Journal of Population Economics 7, 133-156.

Dustmann, C. and A. V. van Soest (2001), "Language Fluency and Earnings: Estimation with Misclassified Language Indicators," The Review of Economics and Statistics 83, 663-674.

Dustmann, C. and A. V. van Soest (2002), "Language and the Earnings of Immigrants," Industrial and Labor Relations Review 55, 473-492.

Eckstein Z. and Y. Weiss (2004), "On the Wage Growth of Immigrants: Israel, 1990-2000," Journal of the European Economic Association 2, 665-695.

Escanciano, J. C., D. Jacho-Chávez, and A. Lewbel (2016), "Identification and Estimation of Semiparametric Two-Step Models," Quantitative Economics 7, 561-589.

Espenshade, T. and H. Fu (1997), "An Analysis of English Language Proficiency Among U.S. Immigrants," American Sociological Review 62, 288-305.

Ettner, S. L. (1995), "The Impact of Parent Care on Female Labor Supply Decisions," Demography 32, 63-80.

Ettner, S. L. (1996), "The Opportunity Costs of Elder Care," Journal of Human Resources 31, 189-205.

Farrell, M. H., T. Liang, and S. Misra (2021), "Deep Neural Networks for Estimation and Inference," Econometrica 89, 181-213.

Fisher, F. (1966), The Identification Problem in Econometrics. New York: McGraw-Hill.

Frisch, R. (1934), Statistical Confluence Analysis by Means of Complete Regression Systems. Oslo: Universitetets Økonomiske Institutt.

Gebel, M., and F. Pfeiffer (2007), "Educational Expansion and its Heterogeneous Returns for Wage Workers," Discussion Paper No. 07-010, Centre for European Economic Research.

Gonzalez, A. (2003), "The Education and Wages of Immigrant Children: The Impact of Age at Arrival," Economics of Education Review 22, 203–212.

Gonzalez, L. (2005), "Nonparametric Bounds on the Returns to Language Skills," Journal of Applied Econometrics 20, 771-795.

Greene, W. H. (2012), Econometric Analysis. Essex, England: Pearson Education.

Grenier, G. (1984), "The Effects of Language Characteristics on the Wages of Hispanic–American Males," Journal of Human Resources 19, 35–52.

Gwartney, J. D., and J. E. Long (1978), "The Relative Earnings of Blacks and Other Minorities," Industrial and Labor Relations Review 31, 336-346.

Haavelmo, T. (1943), "The Statistical Implications of a System of Simultaneous Equations," Econometrica 11, 1-12.

Hahn, J., P. Todd, and W. van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," Econometrica 69, 201-209.

Hastie, T., R. Tibshirani, and J. Friedman (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, N.Y.: Springer.

Hayashi, F. (2000), Econometrics. Princeton, N.J.: Princeton University Press.

Heckman, J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System," Econometrica 46, 931-959.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error," Econometrica 47, 153-161.

Heckman, J. (1990), "Varieties of Selection Bias," Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, American Economic Review 80, 313-318.

Heckman, J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," Journal of Human Resources 32, 441-462.

Heckman, J., and E. Vytlacil (1998), "Instrumental Variables Methods for the Correlated Random Coefficient Model," Journal of Human Resources 33, 974-987.

Heckman, J., and E. Vytlacil (2007a), Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. In J. Heckman and E. Leamer (eds.) Handbook of Econometrics, Volume 6B. New York: Elsevier Science, ch. 70.

Heckman, J., and E. Vytlacil (2007b), 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. In J. Heckman and E. Leamer (eds.) Handbook of Econometrics, Volume 6B. New York: Elsevier Science, ch. 71.

Heckman, J., and R. Robb (1985a), "Alternative Methods for Evaluating the Impact of Interventions: An Overview," Journal of Econometrics 30, 239-267.

Heckman, J., and R. Robb (1985b), Alternative Methods for Evaluating the Impact of Interventions. In J. Heckman and B. Singer (eds.) Longitudinal Analysis of Labor Market Data (Econometric Society Monographs). New York: Cambridge University Press, ch. 4.

Heckman, J., and X. Li (2004), "Selection Bias, Comparative Advantage and Heterogeneous Returns to Education: Evidence from China in 2000," Pacific Economic Review 9, 155-171.

Heckman, J., D. Schmierer, and S. Urzua (2010), "Testing the Correlated Random Coefficient Model," Journal of Econometrics 158, 177-203.

Heckman, J., Justin L. Tobias, and E. Vytlacil (2003), "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," The Review of Economics and Statistics 85, 748-755.

Hsiao, C. (1983), Identification. In Z. Griliches and M. D. Intriligator (eds.) Handbook of Econometrics, Volume 1. Amsterdam and New York: North-Holland, ch.4.

Hurwicz, L. (1950), Generalization of the Concept of Identification. In T. Koopmans (ed.) Statistical Inference in Dynamic Economic Models. New York: Wiley, ch. 4.

Imbens, G. W., and D. B. Rubin (2015), Causal Inference in Statistics, Social, and Biomedical Sciences. New York: Cambridge University Press.

Imbens, Guido W., and Jeffrey M. Wooldridge (2009), "Recent Developments in the Econometrics of Program Evaluation," Journal of Economic Literature 47, 5-86.

Imbens, Guido W., and T. Lemieux (2008), "Regression Discontinuity Designs: A Guide to Practice," Journal of Econometrics 142, 615-635.

Iskhakov, F., J. Rust, and B. Schjerning (2020), "Machine Learning and Structural Econometrics: Contrasts and Synergies," The Econometrics Journal 23, S81–S124.

Keane, M. P. (1992), "A Note on Identification in the Multinomial Probit Model," Journal of Business and Economic Statistics 10, 193-200.

Klein, L. (1953), A Textbook of Econometrics. Evanston, Ill: Row, Peterson and Co.

Kock, A. B. (2016), "Oracle Inequalities, Variable Selection and Uniform Inference in High-dimensional Correlated Random Effects Panel Data Models," Journal of Econometrics 195, 71–85.

Koopmans, T. C., and O. Reiersøl (1950), "The Identification of Structural Characteristics," The Annals of Mathematical Statistics, 165-181.

Koopmans, T. C., H. Rubin, and R. B. Leipnik (1950), Measuring the Equation Systems of Dynamic Economics. In T. Koopmans (ed.) Statistical Inference in Dynamic Economic Models. New York: Wiley, ch. 2.

Kossoudji, S. A. (1988), "English Language Ability and the Labor Market Opportunities of Hispanic and East-Asian Immigrant Men," Journal of Labor Economics 6, 205–228.

Laub, A. J. (2005), Matrix Analysis for Scientists and Engineers. Philadelphia, Penn: Society for Industrial and Applied Mathematics.

Lee, L. F. (1978), "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables," International Economic Review 19, 415– 433.

Lee, L. F. (1983), "Generalized Econometric Models with Selectivity," Econometrica 51, 507–512.

Leslie, D. and J. Lindley (2001), "The Impact of Language Ability on Employment and Earnings of Britain's Ethnic Communities," Economica 272, 587-606.

Leung, S. F., and S. Yu (1996), "On the Choice between Sample Selection and Two-Part Models," Journal of Econometrics 72, 197-229.

Lewbel, A. (2012), "Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models," Journal of Business and Economic Statistics 30, 67-80.

Lewbel, A. (2018), "The Identification Zoo: Meanings of Identification in Econometrics," Journal of Economic Literature 57, 835-903.

Lewbel, A., Y. Dong, and T. T. Yang (2012), "Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors," Boston College Working Papers in Economics Number 789.

Lewis, E. G. (2011), "Immigrant-Native Substitutability: The Role of Language Ability (NBER Working Paper Series No. 17609)," Retrieved from NBER, The National Bureau of Economic Research website: https://www.nber.org/papers/w17609

Lewis, H. G. (1983), "Union Relative Wage Effects: A Survey of Macro Estimates," Journal of Labor Economics 1, 1-27.

Lin, W. (2013), "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." The Annals of Applied Statistics 7, 295–318.

Linden, A., S. D. Uysal, A. Ryan, and J. L. Adams (2016), "Estimating Causal Effects for Multivalued Treatments: A Comparison of Approaches," Statistics in Medicine 35, 534–552.

Little, R. J. A. (1985), "A Note About Models for Selectivity Bias," Econometrica 53, 1469-1474.

Maddala, G. S. (1986), Limited-dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.

Maddala, G. S., and A. Flores-Lagunes (2001), Qualitative Response Models. In B. H. Baltagi (ed.) A Companion to Theoretical Econometrics. Malden, Mass: Blackwell Publishers, ch.17.

Manski, C. F. (1988), Analog Estimation Methods in Econometrics. New York, N.Y.: Chapman and Hall.

Manski, C. F. (1990), "Nonparametric Bounds on Treatment Effects," The American Economic Review 80, 319-323.

Manski, C. F. (1995), Identification Problems in the Social Sciences. Cambridge, Mass: Harvard University Press.

Manski, C. F. (2003), Partial Identification of Probability Distributions. New York: Springer.

Manski, C., and J. Pepper (2000), "Monotone Instrumental Variables: with an Application to the Returns to Schooling," Econometrica 68, 997–1010.

Marschak, J., and W. H. Andrews (1944), "Random Simultaneous Equations and the Theory of Production," Econometrica 12, 143-205.

Matzkin R. L. (2007), Nonparametric Identification. In J. Heckman and E. Leamer (eds.) Handbook of Econometrics, Volume 6B. Amsterdam: Elsevier, ch. 73.

McFadden, D. L. (1973), Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (ed.) Frontiers in Econometrics. New York: Academic Press, ch. 4.

McFadden, D. L. (1984), Econometric Analysis of Qualitative Response Models. In Z. Griliches and M. D. Intrilligator (eds.) Handbook of Econometrics, Volume 2. Amsterdam and New York: North-Holland, ch.24.

McKenzie, C. R., and M. McAleer (1994), "On the Effects of Misspecification Errors in Models with Generated Regressors," Oxford Bulletin of Economics and Statistics 56, 441-455.

McManus W., W. Gould, and F. Welch (1983), "Earnings of Hispanic Men: The Role of English Language Proficiency," Journal of Labor Economics 1, 101-130.

Meghir, C., and M. Palme (2001), "The Effect of a Social Experiment in Education" Working Paper WP01/11, The Institute for Fiscal Studies.

Mendelsohn, R. (1985), "Identifying Structural Equations with Single Market Data," The Review of Economics and Statistics 67, 525-529.

Mincer, J. A. (1974), Schooling, Experience, and Earnings. New York: NBER, available at: http://www.nber.org/books/minc74-1

Miranda, A., and Y. Zhu (2013), "English Deficiency and the Native–Immigrant Wage Gap," Economics Letters 118, 38–41.

Mocan, H. N., and E. Tekin (2003), "Nonprofit Sector and Part-Time Work: An Analysis of Employer-Employee Matched Data on Child Care Workers," The Review of Economics and Statistics 85, 38-50.

Moffitt, R. (1999), New Developments in Econometric Methods for Labor Market Analysis. In O. Ashenfelter and D. Card (eds.) Handbook of Labor Economics, Volume 3A. Amsterdam and New York: North-Holland, ch.24.

Montmarquette, C., N. Viennot-Briot, and M. Dagenais (2007), "Dropout, School Performance, and Working while in School," The Review of Economics and Statistics 89, 752-760.

Mullainathan, S., and J. Spiess (2017), "Machine Learning: An Applied Econometric Approach," Journal of Economic Perspectives 31, 87-106.

Negi, A., and J. M. Wooldridge (2021), "Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects," Econometric Reviews 40, 504-534.

Newey, W. K. (1984), "A Method of Moments Interpretation of Sequential Estimators," Economics Letters 14, 201–206.

Newey, W. K. (1993), Efficient Estimation of Models with Conditional Moment Restrictions. In G. S. Maddala, C. R. Rao, and H. D. Vinod (eds.) Handbook of Statistics, Volume 11. Amsterdam: North-Holland, ch. 16.

Newey, W. K., and D. McFadden (1994), Large Sample Estimation and Hypothesis Testing. In R. F. Engle and D. McFadden (eds.) Handbook of Econometrics, Volume 4. Amsterdam: North-Holland, ch. 36.

Norton, E. C., and D. O. Staiger (1994), "How Hospital Ownership Affects Access to Care for the Uninsured," The RAND Journal of Economics 25, 171-185.

Olsen, R. J. (1980), "A Least Squares Correction for Selectivity Bias ," Econometrica 48, 1815-1820.

Pagan, A. (1986), "Two Stage and Related Estimators and Their Applications," The Review of Economic Studies 53, 517-538.

Puhani, P. A., and A. M. Weber (2007), "Does the Early Bird Catch the Worm? Instrumental Variable Estimates of Early Educational Effects of Age of School Entry in Germany," Empirical Economics 32, 359-386.

Rao, C. R. (1948), "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation," Mathematical Proceedings of the Cambridge Philosophical Society 44, 50-57.

Reiersøl, O. (1941), "Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis," Econometrica 9, 1-24.

Reilly, K. T. (1996), "Does Union Membership Matter? The Effect of Establishment Union Density on the Union Wage Differential," The Review of Economics and Statistics 78, 547-557.

Reimers, C. W. (1983), "Labor Market Discrimination against Hispanic and Black Men," The Review of Economics and Statistics 65, 570–579.

Robins, J. M. (1989a), The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies. In L. Sechrest, H. Freeman, and A. Bailey (eds.) Health Service Research Methodology: A Focus on AIDS, National Center for Health Services Research, U.S. Public Health Service, pp. 113-159.

Robins, J. M. (1989b), "The Control of Confounding by Intermediate Variables," Statistics in Medicine 8, 679-701.

Robinson, Chris. (1989), "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models," Journal of Political Economy 97, 639-667.

Rosen, S. (1977), Human Capital: A Survey of Empirical Research. In R. Ehrenberg (ed.) Research in Labor Economics, Volume 1. Greenwich, Conn: JAI Press, ch. 1.

Rosenbaum, P. R., and D. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika 70, 41-55.

Rothenberg, T. J. (1971), "Identification in Parametric Models," Econometrica 39, 577-591.

Rubin, D. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," The Annals of Statistics 6, 34-58.

Schaafsma J. and A. Sweetman (2001), "Immigrant Earnings: Age at Immigration Matters," Canadian Journal of Economics 34, 1066-1099.

Schaffner, J. A. (2002), "Heteroskedastic Sample Selection and Developing-Country Wage Equations," The Review of Economics and Statistics 84, 269-280.

Searle, S. R. (1982), Matrix Algebra Useful for Statistics. New York: Wiley.

Semenova, V., M. Goldman, V. Chernozhukov, and M. Taddy (2021), "Estimation and Inference on Heterogeneous Treatment Effects in High-Dimensional Dynamic Panels (arXiv No: 1712.09988)," Retrieved from arXiv website: https://arxiv.org/pdf/1712.09988.pdf

Shen, C. (2013), "Determinants of Health Care Decisions: Insurance, Utilization, and Expenditures," The Review of Economics and Statistics 95, 142-153.

Shields, M. A. and S. W. Price (2002), "The English Language Fluency and Occupational Success of Ethnic Minority Immigrant Men Living in English Metropolitan Areas," Journal of Population Economics 15, 137-160.

Simon C. P. and L. Blume (1994), Mathematics for Economists. New York: Norton.

Sloan, F. A., G. A. Picone, D. H. Taylor Jr., and S. Chou (2001), "Hospital Ownership and Cost and Quality of Care: Is There a Dime's Worth of Difference?," Journal of Health Economics 20, 1-21.

Stock, J. H., and F. Trebbi (2003), "Retrospectives Who Invented Instrumental Variable Regression?," Journal of Economic Perspectives 17(3), 177-194.

Swamy, P. A. V. B. (1970), "Efficient Inference in a Random Coefficient Regression Model," Econometrica 38, 311-323.

Swamy, P. A. V. B., and G. S. Tavlas (2001), Random Coefficient Models. In B. H. Baltagi (ed.) A Companion to Theoretical Econometrics. Malden, Mass: Blackwell Publishers, ch.19.

Swamy, P. A. V. B., and J. S. Mehta (1977), "Estimation of Linear Models with Time and Cross-Sectionally Varying Coefficients," Journal of the American Statistical Association 72, 890-898.

Tainer, E. (1988), "English Language Proficiency and the Determination of Earnings among Foreign-Born Men," Journal of Human Resources 23, 108-122.

Tamer, E. (2010), "Partial Identification in Econometrics," Annual Review of Economics 2, 167- 195.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological) 58, 267-288.

van der Klaauw, W. (2002), "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," International Economic Review 43, 1249-1287.

Vella, F., and M. Verbeek (1999), "Estimating and Interpreting Models with Endogenous Treatment Effects," Journal of Business and Economic Statistics 17, 473-478.

Wald, A. (1950), Statistical Decision Functions. New York: Wiley.

White, H. (1994), Estimation, Inference and Specification Analysis. Cambridge, New York: Cambridge University Press.

White, H. (2001), Asymptotic Theory for Econometricians. San Diego, Calif: Academic Press.

Wilde, J. (2000), "Identification of Multiple Equation Probit Models with Endogenous Dummy Regressors," Economics Letters 69, 309-312.

Willis, R. J., and S. Rosen (1979), "Education and Self-Selection," Journal of Political Economy 87, Part 2: Education and Income Distribution, S7-S36.

Wooldridge, J. M. (1994), Estimation and Inference for Dependent Processes. In R. F. Engle and D. McFadden (eds.) Handbook of Econometrics, Volume 4. Amsterdam: North-Holland, ch. 45.

Wooldridge, J. M. (1997), "On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model," Economics Letters 56, 129-133.

Wooldridge, J. M. (2003), "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model," Economics Letters 79, 185-191.

Wooldridge, J. M. (2008), Instrumental Variables Estimation of the Average Treatment Effect in the Correlated Random Coefficient Model. In T. Fomby, R. C. Hill, D. L. Millimet, J. A. Smith, and E. J. Vytlacil (eds.) Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics, Volume 21. Bingley: Emerald Group Publishing Limited, ch. 4.

Wooldridge, J. M. (2010), Econometric Analysis of Cross Section and Panel Data. Cambridge, Mass: MIT Press.

Wooldridge, J. M. (2015), "Control Function Methods in Applied Econometrics," The Journal of Human Resources 50, 420-445.

Working, E. J. (1927), "What Do Statistical Demand Curves Show?," The Quarterly Journal of Economics 41, 212-235.

Working, H. (1925), "The Statistical Determination of Demand Curves," The Quarterly Journal of Economics 39, 503-543.

Wright, P. G. (1928), The Tariff on Animal and Vegetable Oils, New York: Macmillan.

Yao, Y. and J. C. van Ours (2015), "Language Skills and Labor Market Performance of Immigrants in the Netherlands," Labour Economics 34, 76-85.

Zellner, A. (1969), On the Aggregation Problem: A New Approach to a Troublesome Problem. In K. A. Fox, J.K. Sengupta, and G. V. L. Narasimham (eds.) Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner. Berlin, Heidelberg: Springer, ch. 16.