BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

ESSAYS IN ECONOMETRICS: BIAS CORRECTIONS AND ROBUST INFERENCE

by

SHUOWEN CHEN

B.A., University of Rochester, 2015
M.A., The University of Texas at Austin, 2016

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2022

Approved by

First Reader

Iván Fernández-Val, Ph.D
Professor of Economics

Second Reader

Hiroaki Kaido, Ph.D
Associate Professor of Economics

Third Reader

Jean-Jacques Forneron, Ph.D
Assistant Professor of Economics

Fourth Reader

Marc Rysman, Ph.D
Professor of Economics

To Yuqi Guo

Acknowledgments

I express my foremost gratitude to Iván Fernández-Val, who has been the principal guide of my apprenticeship. He gave me the green light to explore what I am interested in, and I owe much to him for his patience, tolerance, the rigorous training and the amount of time he lent me. His insights of panel data analysis set this thesis in motion, and his work ethic will always motivate me to push harder.

I am deeply grateful to Hiroaki Kaido for introducing me to the world of misspecification and partial identification. He changed my perspective of econometrics, and it is hard to specify his innumerable insights which will continue to permeate my econometrics life. I thank him for the chances of coauthorship, teaching and research assistance. I will emulate his approach to econometrics for years to come.

The most memorable part of my graduate school experience is countless discussions with Jean-Jacques Forneron. His passion for econometrics and computation is contagious, and his tremendous knowledge of simulation–based estimation inspired me to dig deeper and propose my own solutions. I will miss our long conversations about econometrics, coding tricks, mathematics, coffee and life choices.

I thank Pierre Perron for encouraging me to switch to econometrics in his time series course. His philosophy that econometrics can be presented using simple intuitive arguments has reshaped my writings, teaching and presentations. I will always carry a curious mind and keep asking questions.

Zhongjun Qu convinced me to work on econometrics in my second year, and I vividly remember him saying "...within four years you will thank yourself for making the right decision today." I do not regret my decision at all, and I cannot help but wonder if he has the crystal ball.

I express my gratitude to Susan Athey, Victor Chernozhukov, Andrew Chesher, Guido Imbens and Francesca Molinari. Their recognition on my works at different

stages give me more confidence to do my own things and take risks. Even though I have decided not to pursue an academic career, I will bring this confidence to my daily work and strive to make an impact.

I would like to thank other teachers that have generously shared with me their knowledge on a range of topics. Marc Rysman showed me how to synthesize econometric theory and empirics. Jihye Jeon taught me how to write referee reports: after having pointed out a problem, always think about giving feasible suggestions. Stephen Terry's course on firm dynamics sparked my interest in firm heterogeneity and enabled me to produce a relevant project. Adam Guren's slides on how to give talks will always be my go–to reference for presentations.

I thank colleagues who have had positive peer effects on my research. Among them are Undral Byambadalai, Alessandro Casini, Vince Weize Chen, Qifan Han, Eric Hardy, Chuqing Jin, Paul Koh, Yan Liu, Siyi Luo, Yang Ming, Julio Ortiz, Stefano Pica, Dongling Su, Dongwei Xu, Guang Zhang and Beixi Zhou. Yan proofread the whole dissertation and caught many typos. Dongwei nudged me to econometrics when I was in self–doubt. Conversations with Chuqing have always been intellectually stimulating, and us not being able to work on something together in graduate school is my biggest regret. I enjoy discussing macroeconomics and Schopenhauer with Yang, and I wish him all the best in the academia.

Yaomin Pan and Lynn Qiaoyu Qi are my two best buddies. Our friendships went way back to Rochester, and over these years we supported each other through broken relationships, unemployment and the job market. Hope to see them soon.

I am indebted to my parents for their unconditional support and tolerance. Hopefully I have made them proud.

It takes time to learn how to balance between exploration and exploitation. I did not lose the game. I just ran out of time.

ESSAYS IN ECONOMETRICS: BIAS CORRECTIONS AND ROBUST INFERENCE

SHUOWEN CHEN

Boston University, Graduate School of Arts and Sciences, 2022

Major Professor: Iván Fernández-Val, Ph.D
Professor of Economics

ABSTRACT

Unobserved heterogeneity is common in economic data and has nontrivial impacts on modelling, estimation and statistical inference. This dissertation consists of three chapters that explore and illustrate the implications of unobserved heterogeneity for different types of data. The first two chapters focus on panel data, and the third chapter focuses on cross–sectional data.

A popular way to control for unobserved heterogeneity in panel data models is to include fixed effects, but fixed effect estimators of dynamic and nonlinear panel models are subject to the incidental parameter problem. This problem has two implications for applied research: (1) point estimators are largely biased, and (2) confidence intervals have incorrect coverages. Chapter 1 proposes a new method for bias reduction based on indirect inference. The method simulates data using the model with estimated individual effects, and finds estimators of the model parameters by equating the fixed effect estimates obtained from observed and simulated data. The asymptotic framework provides consistency, bias correction, and asymptotic normality results. An application to female labor force participation and numerical simulations illustrate the finite–sample performance of the method.

Chapter 2 is coauthored with Victor Chernozhukov at MIT, Iván Fernández-Val at BU, and Hiroyuki Kasahara and Paul Schrimpf at The University of British Columbia. It is based on the observation that existing jackknife methods that deal with the incidental parameter problem require stationary variables. However, many applications feature covariates of interests that have trends or structural breaks. This chapter proposes a new jackknife bias correction method that relaxes stationarity. The method is named crossover jackknife because it partitions the panel in two halves, each including half of the time series observations for each cross sectional unit, but where the time periods are crossed over between the two halves of the cross section units. We derive the theoretical properties of this method and illustrate its finite–sample performance via calibrated numerical simulations.

Chapter 3 is coauthored with Hiroaki Kaido at BU. In many important discrete choice models, whether the model makes a unique prediction or not depends on its policy-relevant features and can be examined by testing restrictions on underlying structural parameters. Imposing strong assumptions completes the model and allows to predict unique outcomes, but it masks heterogeneity of the data and affects statistical inference. We provide a new test of model incompleteness using a score-based statistic. Our test statistic remains computationally tractable even with a moderate number of nuisance parameters because they only need to be estimated in the restricted complete model. Two empirical applications illustrate the computational feasibility of the method. A Monte Carlo experiment shows the score test outperforms existing tests in terms of local power.

Contents

List of Tables

List of Figures

List of Abbreviations and Notations

$\|\cdot\|_{op}$	.....	the operator norm for linear mappings
$\|\cdot\|_{\mathcal{F}}$	.....	the supremum norm over $\mathcal{F}$
$a \lesssim b$	.....	$a \leq Mb$ for some constant M
CDF	.....	cumulative distribution function
CLT	.....	central limit theorem
DGP	.....	data generating process
LFP	.....	least favorable pair
LLN	.....	law of large numbers
$N(\epsilon, \mathcal{F}, \|\cdot\|)$	.....	covering number of size ϵ for $\mathcal{F}$ under norm $\|\cdot\|$
$N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$	.....	bracketing number of size ϵ for $\mathcal{F}$ under norm $\|\cdot\|$
PDF	.....	probability density function
$\mathbb{R}^2$	.....	the real plane
SMM	.....	simulated method of moments
$X_n \overset{P^n}{\rightsquigarrow} X$	.....	X_n weakly converges to X under $\{P^n\}$
$\xrightarrow{d}$	.....	convergence in distribution
$\xrightarrow{p}$	.....	convergence in probability
$x \propto y$	.....	x is proportional to y
ULLN	.....	uniform law of large numbers

Chapter 1

Indirect Inference for Nonlinear Panel Models with Fixed Effects

1.1 Introduction

Panel data refers to data on multiple entities (e.g., individuals, firms, etc.) observed at two or more time periods. Unobserved heterogeneity across entities often accounts for a large fraction of the variation in panel data. When this heterogeneity is correlated with the explanatory variables in the regression specifications, the resulting omitted variable bias renders point estimates inconsistent.

Adding individual fixed effects, α_{i0}'s, is the main approach to control for time–invariant unobserved heterogeneity in panel data models. Compared to other approaches like random effects and correlated random effects, the fixed effect approach does not impose distributional assumptions on α_{i0}'s or restrict their relationships with other explanatory variables. Instead, each α_{i0} is treated as a parameter to be estimated. However, because the number of α_{i0}'s increases with the sample size and each α_{i0} is estimated using only entity i's time series observations, adding fixed effects introduces the incidental parameter problem in estimating the vector of parameters of interest θ_0. It has two consequences for applied research: (1) point estimates are subject to large biases, and (2) confidence intervals have incorrect coverages.

This paper proposes a new method to debias fixed effect estimators in a class of nonlinear panel models. The method is named *indirect fixed effect estimation* and

features two main steps: the first one is to simulate data by using estimated individual effects $\widehat{\alpha}_i$'s from the observed data. The second step is to find the vector of parameters that matches the fixed effect estimators using observed and simulated data.

The method has two advantages: first, it does not require an explicit characterization of the bias term, which can be hard to derive in complex models. Instead, the method finds the solution by automatically correcting the bias because the vector of parameter values that is the closest to θ_0 renders similar bias in fixed effect estimations. Second, standard errors can be derived using the delta method, so there is no need to use the bootstrap, which is computationally intensive.

The two properties are inherited from a precedent simulation–based estimation approach called indirect inference, which was first developed by Gouriéroux et al. (1993) and Smith (1993). In a nutshell, indirect inference uses an auxiliary model to summarize the statistical properties of the observed data and simulated data, and finds values of model parameters that match the parameters of the auxiliary model, estimated using the observed and simulated data, in terms of a minimum–distance criterion function. Because the same regression is run on observed and simulated data, matched estimators have the same bias structure and thus the bias gets cancelled.

The theory of indirect inference, however, is not directly applicable to nonlinear panel models, which are widely used in various fields of economics like industrial organization and labor. Because the individual effects cannot be differenced out, data simulations seem infeasible without imposing a parametric specification on their distributions, and the bias term is a complicated function of θ_0 and α_{i0}'s.

To simulate data, this paper proposes using the estimated individual effects $\widehat{\alpha}_i$'s. These are informative proxies for the unknown individual effects α_{i0}'s because they become more accurate estimates when each individual's number of time series observations T grows large. Intuitively speaking, although data simulated using $\widehat{\alpha}_i$'s do

not perfectly mimic the observed data, such a difference vanishes when T increases.

The indirect fixed effect estimator then debiases by matching the fixed effect estimates using observed and simulated data. This brings two advantages for the implementation and theoretical analysis of the new estimator. First, the minimum–distance criterion function for matching is just–identified because the dimensions of the fixed effect estimates are identical. Therefore, there is no need to consider an estimation of an optimal weighting matrix. It further implies that the matching can be made as exact as machine precision permits. The second advantage is with respect to the relationship between the vector of parameters of interest θ_0 and the unique maximizer of the limiting log–likelihood function for fixed effect estimation. To back out point estimates of θ_0 from fixed effect estimators using simulated data, this relationship should be invertible. Because the unique maximizer is θ_0, the relation turns out to be an identity function. Therefore, invertibility is satisfied trivially.

This paper derives consistency, bias correction and asymptotic normality results for the indirect fixed effect estimator. As usual in the indirect inference literature, consistency requires that the fixed effect estimates using observed and simulated data converge to the unique maximizer of the limiting log likelihood. Although the pointwise convergence of $\widehat{\theta}$ to θ_0 is a standard result in the large–T panel literature, three important differences arise in the analysis of fixed effect estimates using simulated data and pose theoretical challenges.

First, the simulated data are generated using $\widehat{\alpha}_i$'s instead of α_{i0}'s. To justify this practice, the corresponding log likelihood function should uniformly well approximate the one rendered by data simulated using the true individual effects. Otherwise, simulated fixed effect estimator could not be pointwise convergent. The proof of this statement, however, is complicated by the fact that the log likelihood function using simulated data is typically nonsmooth for important types of nonlinear panel

models, with binary choice models as leading examples. Intuitively speaking, when the dependent variable is discrete, a small change in the parameter values can lead to discrete changes in the simulated data. As a result, the sample log likelihood function using simulated data is discontinuous.

Simulations often generate discontinuous objective functions (e.g., McFadden, 1989; Pakes and Pollard, 1989), but this paper confronts a second difference: the fixed effect estimator using simulated data is nonsmooth with respect to the parameters of the data generating process (DGP). Therefore, standard proof strategies in the panel literature (e.g., Hahn and Newey, 2004; Hahn and Kuersteiner, 2011) cannot be directly applied to characterize its limiting behavior.

Empirical process theory provides ample tools to handle nonsmoothness functions and moments in econometrics (Andrews, 1994), but the analysis of a nonsmooth fixed effect estimator is further complicated by the third difference: the presence of incidental parameters, whose number increases with the sample size n.

To prove uniform convergence with nonsmoothness, this paper follows Newey (1991) by establishing pointwise convergence and stochastic equicontinuity of the fixed effect estimator in the simulation world. Intuitively speaking, pointwise convergence is equivalent to uniform convergence for any finite number of grid points, but without smoothness, the gap between any two grids can behave rather erratically. The stochastic equicontinuity condition is hence required to restrict such behaviors in probability.

The theoretical analysis of the indirect fixed effect estimator relies on some key structures of the panel data and the log likelihood function. Under the assumption that panel data are independent along the cross section dimension, this paper first justifies data simulation with $\widehat{\alpha}_i$'s by proving that the corresponding log likelihood function uniformly approximates the one from simulated data generated by α_{i0}'s. As

such, a uniform law of large number can be established and pointwise convergence in the simulation world follows from the standard consistency argument (Newey and McFadden, 1994). To verify the stochastic equicontinuity condition of fixed effect estimators using simulated data, this paper uses the concavity property of the profiled log likelihood to verify one of the primitive conditions for stochastic equicontinuity in Andrews (1994). The proof strategy might be of independent interest.

Regarding the asymptotic unbiasedness and normality of the new estimator, due to non–smoothness in the simulation world, the conventional strategy in indirect inference that relies on the implicit function theorem (e.g., Gouriéroux et al., 1993) is not directly applicable. A regularity conditions is thus imposed, which, combined with consistency, allows to explore bias correction and asymptotic normality through the lens of fixed effect estimates. More specifically, the fixed effect estimators in both worlds have the same structures regarding the bias terms and influence functions. The difference is that the ones using the real data are functions of θ_0 and α_{i0}'s while those using the simulated data are functions of θ_0 and $\widehat{\alpha}_i$'s.

This paper currently imposes two high–level conditions to ensure that the bias term and the influence function from data simulated using θ_0 and $\widehat{\alpha}_i$'s are uniformly close to their infeasible counterparts from data simulated using θ_0 and α_{i0}'s with asymptotically negligible approximation errors. The infeasible bias converges to the same probability limit as does the bias obtained from observed data, while the infeasible influence function converges to the same normal distribution as does the influence function from observed data. Therefore, the theory of indirect inference can be invoked to establish bias cancellation and asymptotic normality.

Like other simulation–based estimation methods, the asymptotic variance of the new estimator is inflated by the inverse of the number of simulation draws. The result can also be intepreted as a reflection of the classic bias–variance tradeoff. As shown in

the application and Monte Carlo simulations, however, the finite–sample performance of the indirect fixed effect estimator is comparable to the leading methods in terms of bias correction and outperforms half–panel bias correction methods in terms of standard errors.

Related Literature

The indirect fixed effect estimator presented in this paper combines four strands of literature, of which this section provides a non–exhaustive overview. The incidental parameter problem is first discussed by Neyman and Scott (1948). When T is fixed, fixed effect estimators of nonlinear models are in general inconsistent because estimation errors of $\widehat{\alpha}_i$'s do not vanish even when the cross–section sample size n is very large (Chamberlain, 1984; Lancaster, 2000). Only some special models like static linear and logit specifications feature fixed–T consistent estimators (Andersen, 1970). A key insight of the large–T panel data literature is that the incidental parameter problem becomes an asymptotic bias problem when T grows with the sample size n. When n and T grow at the same rate, fixed effect estimators are consistent and asymptotically normal, but they have a bias comparable to standard errors.

In the search for asymptotically unbiased estimators, there are two leading approaches. For certain types of models, the bias terms have been characterized analytically and corrected using a plug–in approach (Hahn and Kuersteiner, 2002; Hahn and Newey, 2004; Fernández-Val, 2009; Hahn and Kuersteiner, 2011). However, such terms can be hard to derive for complicated models. Under further sampling and regularity conditions, bias terms can be automatically corrected using jackknife. For example, Hahn and Newey (2004) proposed leave–one–out panel jackknife for data that do not have dependencies among observations of the same unit. Dhaene and Jochmans (2015) relaxed the assumption to stationarity along the time series, and proposed a half–panel method. Under an unconditional homogeneity assumption,

Fernández-Val and Weidner (2016) allowed for two–way fixed effects and propose a jackknife method that corrects biases from both dimensions. See Arellano and Hahn (2007) and Fernández-Val and Weidner (2018) for recent surveys. Standard errors are typically obtained using panel bootstrap, which can be computationally intensive.[1]

Another popular simulation–based method that can achieve bias correction is bootstrap (Horowitz, 2001, 2019). Gonçlaves and Kaffo (2015) proposed the bootstrap bias correction methods for dynamic linear panel models without covariates. Kim and Sun (2016) proposed a parametric bootstrap bias correction (BBC) method for the nonlinear panel models considered in this paper. Compared to the indirect fixed effect estimator, the BBC estimator mimics the bias term and removes it from the fixed effect estimate explicitly, and thus the proof strategies are very different.

Second, this paper extends the existing theory and practice of indirect inference. Since the introduction of the method, its asymptotic theory has mainly been focused on times series data (Gouriéroux et al., 1993; Smith, 1993; Gallant and Tauchen, 1996). Some recent papers explore asymptotic properties in panel data with discrete dependent variables, but there are two key differences with this paper. First, their settings hold time series dimension fixed and study different types of models. For example, Bruins et al. (2018) did not consider models with fixed effects, Frazier et al. (2019) imposed normality on individual effects, and Taber and Sauer (2021) assumed a bivarite normal distribution on the types of individuals. Second, they deal with nonsmoothness by smoothing the discontinuous parts and showing that the resultant bias can be corrected.

Gouriéroux et al. (2010) is the first paper that establishes theoretical properties of indirect inference for a class of large–T panel models. They applied indirect inference

[1]For example, to obtain one debiased point estimate, fixed effect estimations are run three times: one for the whole sample, and twice for the two split samples. If the number of bootstraps is set to be 500, then the total number of fixed effects estimations becomes 1500. In addition, in practice it is often recommended to use multiple sample splits to improve the finite–sample performance.

to dynamic panel linear models, whose fixed effect estimators are known to be biased (Nickell, 1981). The linear structure allows them to eliminate the individual fixed effects α_{i0}'s by first–difference. As such, α_{i0}'s do not show up in the bias term, and data can be simulated without information on them. However, first difference does not remove the α_{i0}'s to nonlinear panel models, and this paper fills the gap by extending the theory to handle the presence of α_{i0}'s in data simulation and the bias term.

Indirect inference is popular in various fields of economics, including empirical industrial organization (Collard-Wexler, 2013), labor economics (Altonji et al., 2013) and macroeconomics (Guvenen and Smith, 2014; Berger and Vavra, 2019). However, finding an informative auxiliary model is not a trivial task, and researchers often have to assume invertibility of the limiting relationship between auxiliary parameters and parameters of interest. This paper provides an alternative choice, namely the log likelihood function from the nonlinear panel model, for researchers that employ panel data with fixed effects. The estimation procedures are simple to implement as fixed effect estimation schemes are available in free software like R and Julia.

Nonsmooth objective functions are common in econometrics, and empirical process methods are standard tools for asymptotic analysis (Andrews, 1994; Newey and McFadden, 1994; van der Vaart and Wellner, 1996). The seminal work on simulation–based methods by Pakes and Pollard (1989) is predicated on the independence assumption of cross section data and therefore is not suitable for panel data, which feature dependence for each individual time series. Dedecker and Louhichi (2002) provided an overview of maximal inequalities for empirical central limit theorems for dependent data. Kato et al. (2012) provided new stochastic inequalities for mixing sequences and also established stochastic equicontinuity in the presence of nuisance parameters, but their analysis focused on a different class of nonlinear models, namely panel quantile regression models.

Simulation–based methods like simulated method of moments (McFadden, 1989; Pakes and Pollard, 1989; Lee and Ingram, 1991; Duffie and Singleton, 1993) and indirect inference are widely used to estimate models that do not render tractable moments or likelihood functions. See Gouriéroux and Monfort (1997) for an overview. These methods typically require models to be fully specified, but economic theory does not always provide guidance on functional forms, distributions of shocks or measurement error of observed data. Therefore, the resultant estimators can be subject to misspecification.

This paper considers a class of nonlinear panel models that do not impose distributional assumptions on the individual effects, and hence contributes to a burgeoning literature that considers simulations for models that relax the full parametric specifications in various ways. Dridi and Renault (2000) and Dridi et al. (2007) embedded the semiparametric structural model into a full model for data simulation, and proposed an encompassing principle where parameters of interest are consistently estimated even though nuisance parameters are inconsistently estimated due to misspecification of the full model. Schennach (2014) considered parameters estimation in moment conditions that contain unobservable variables, and proposed a simulation–based method that constructs equivalent moments involving only observable variables. Gospodinov et al. (2017) considered parameter estimation of autoregressive distributed lag models in which covariates are contaminated by serially correlated measurement errors. They proposed a method such that simulated covariates preserve the dependence structure observed in the data even though the dynamics of latent covariates or measurement errors are not specified. Forneron (2020) approximated the distribution of shocks by sieves and proposed a sieve–SMM estimator that jointly estimates structural parameters and the distribution of shocks.

Structure of the Paper

The rest of the paper proceeds as follows: Section 1.2 introduces the model and describes the fixed effect estimator and incidental parameter problem. Section 1.3 provides an overview of the indirect fixed effect estimator and its implementation. Section 1.4 presents the theoretical properties of the estimator. Section 1.5 applies the method to a dataset on labor force participation to illustrate the finite–sample properties of the estimator. Section 1.6 uses numerical simulations to compare the new estimator with other bias correction methods. Section 1.7 concludes and discusses open questions. Appendices A and E consist of proofs and computation details.

1.2 Nonlinear Panel Model and Fixed Effect Estimator

This section starts with a description of nonlinear panel models with fixed effects. Let the data observations be denoted by $\{z_{it} = (y_{it}, x_{it})\colon\ i = 1, \dots, n; t = 1, \dots, T\}$, where y_{it} is the dependent variable and x_{it} is a $p \times 1$ vector of explanatory variables. The observations are independent across entity i and weakly dependent across time t. The DGP of outcome y_{it} takes the following form:

$$y_{it} \mid x_i^T, \alpha_i, \theta \sim f(\cdot \mid x_{it}; \alpha_i, \theta), \tag{1.1}$$

where $x_i^T := (x_{i1}, \dots, x_{iT})$, θ is a $p \times 1$ vector of model parameters and α_i is a scalar individual effect. The explanatory variable x_{it} is strictly exogenous. The model is semiparametric in that neither the distribution of α_i nor its relationship with x_{it} is specified. The conditional density f denotes the parametric part of the model and its form depends on the parametric family of distributions $\{u_{it}\}$. Depending on the specification of f, this type of models have been used to study many different questions of economic interest.

Example 1 (Discrete Choice Model)**.** Let y_{it} denote a binary variable and F_u a cumu-

lative distribution function (CDF), e.g., the standard normal or logistic distribution. Suppose the binary variable is generated by the following single index process with additive individual effects:

$$y_{it} = \mathbf{1}\{x_{it}'\theta + \alpha_i \geq u_{it}\}, \quad u_{it} \mid x_i^T, \alpha_i \sim F_u,$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Then the conditional distribution of y_{it} is expressed as

$$f(y_{it} \mid x_{it}; \alpha_i, \theta) = F_u(x_{it}'\theta + \alpha_i)^{y_{it}}(1 - F_u(x_{it}'\theta + \alpha_i))^{1-y_{it}}.$$

Helpman et al. (2008) modeled a country's export decision as and estimate a gravity equation with country fixed effects. Fernández-Val (2009) used a specification to estimate the determinants of females' labor force participation decisions in the presence of time–invariant heterogeneity such as willingness to work and ability. Collard-Wexler (2013) used a binary logit specification with market–fixed effects to study whether a ready–mix concrete plant decides to be active in a market.

Example 2 (Poisson Regression Model)**.** The Poisson distribution is useful in modeling count data. Let y_{it} denote the number of arrivals of new events within a certain time interval for entity i in year t. For $\lambda_{it} = \exp(x_{it}'\theta + \alpha_i)$, the conditional density is modeled as

$$f(y_{it} \mid x_{it}; \alpha_i, \theta) = \frac{\lambda_{it}^{y_{it}} \exp(-\lambda_{it})}{y_{it}!}\mathbf{1}\{y_{it} \in \{0, 1, \dots\}\}.$$

Using the number of citation–weighted patents as a proxy for innovation, Aghion et al. (2005) employed this specification to study the relationship between innovation and competition with industry fixed effects.

Model (1.1) admits a log likelihood function. The true values of the parameters, denoted by θ_0 and α_{i0}'s, are one solution to the population conditional maximum likelihood problem

$$(\theta_0, \alpha_{10}, \dots, \alpha_{n0}) \in \underset{(\theta,\alpha_1,\dots,\alpha_n)\in\Theta\times\Gamma_\alpha^n}{\arg\max} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\Big[\ln f(y_{it} \mid x_{it}; \alpha_i, \theta)\Big], \tag{1.2}$$

where Θ and Γ_α are the parameter spaces for θ and α_i respectively, the expectation is with respect to the distribution of the observed data, conditional on the unobserved

effects and initial conditions. Section 1.4 discusses assumptions under which the log likelihood function is concave in all parameters and the solution uniquely exists. The indirect fixed effect estimator relies on the uniqueness condition for consistency.

1.2.1 The Fixed Effect Estimator

The fixed effect estimator of θ is obtained by doing maximum likelihood estimation on the sample analog of the population problem (1.2), treating each α_i as a parameter to be estimated.

$$(\widehat{\theta}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_n) \in \underset{(\theta,\alpha_1,\ldots,\alpha_n)\in\Theta\times\Gamma_\alpha^n}{\arg\max} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f(y_{it} \mid x_{it}; \alpha_i, \theta).$$

To facilitate theoretical analysis, this equation is rewritten such that the individual effects are profiled out. More specifically, given θ, the optimal $\widehat{\alpha}_i(\theta)$ for each i is defined as

$$\widehat{\alpha}_i(\theta) \in \underset{\alpha\in\Gamma_\alpha}{\arg\max} \frac{1}{T} \sum_{t=1}^{T} \ln f(y_{it} \mid x_{it}; \alpha, \theta).$$

The estimators $\widehat{\theta}$ and $\widehat{\alpha}_i$ are then

$$\widehat{\theta} \in \underset{\theta\in\Theta}{\arg\max} \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \ln f(y_{it} \mid x_{it}; \widehat{\alpha}_i(\theta), \theta), \quad \widehat{\alpha}_i = \widehat{\alpha}_i(\widehat{\theta}). \tag{1.3}$$

Section 1.4 discusses assumptions under which these estimators exist and are unique with probability approaching one as n and T become large.

1.2.2 The Incidental Parameter Problem

In panel models, the individual effects are incidental parameters, i.e., nuisance parameters whose dimension grows with the number of cross sectional observations n. As equation (1.3) shows, the fixed effect estimator $\widehat{\theta}$ cannot generally be separated from the estimator of individual effects $\widehat{\alpha}_i$'s. Because each $\widehat{\alpha}_i$ is only estimated using the T observations for entity i, its estimation error does not vanish if T is fixed, even

as n grows. These estimation errors in turn contaminate $\widehat{\theta}$. This is the incidental parameter problem for fixed effects estimation. Mathematically,

$$\widehat{\theta} \xrightarrow{p} \theta_T := \arg\max_{\theta\in\Theta} \operatorname{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{T}\sum_{t=1}^{T} \ln f(y_{it} \mid x_{it}, \theta, \widehat{\alpha}_i(\theta))\Big),$$

whereas

$$\theta_0 := \arg\max_{\theta\in\Theta} \operatorname{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\ln f(y_{it} \mid x_{it}, \theta, \alpha_i(\theta))\Big),$$

where

$$\alpha_i(\theta) = \arg\max_{\alpha\in\Gamma_\alpha} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(\ln f(y_{it} \mid x_{it}, \theta, \alpha)).$$

With fixed T, $\widehat{\alpha}_i(\theta) \neq \alpha_i(\theta)$ in general. Therefore, $\theta_T \neq \theta_0$.

To illustrate the problem, suppose y_{it} has the normal distribution with mean α_{i0} and variance θ_0, and the goal is to estimate θ_0. The fixed effect estimator is $\widehat{\theta} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(X_{it} - \widehat{\alpha}_i)^2$, where $\widehat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$. When T is fixed and n approaches infinity, Neyman and Scott (1948) show that

$$\widehat{\theta} \xrightarrow{p} \theta_0 - \frac{\theta_0}{T}.$$

On the other hand, when T also grows to infinity, the bias term $-\frac{\theta_0}{T}$ approaches zero. The large–T panel literature generalizes this insight and shows that the incidental parameter problem becomes an asymptotic bias problem when n and T grow at the same rate.

1.3 The Indirect Fixed Effect Estimator

The key feature of the indirect fixed effect estimator is to match $\widehat{\theta}$ with a fixed effect estimator from simulated data generated by $\widehat{\alpha}_i$'s and a given θ. To avoid confusion, it is necessary to introduce notations to distinguish parameters in the simulation world from those in Section 1.2. More specifically, this paper uses β and γ_i to denote the

vector of parameters of interest and individual effects in the log likelihood function using simulated data.

To clarify the notations and introduce the implementation of indirect fixed effect estimator, this section first revisits the Neyman–Scott example. Using the panel model as a concrete example, this section then illustrates the challenges associated with the presence of nonsmoothness and discusses the general estimation procedures.

1.3.1 Neyman–Scott Example Revisited

If $y_{it} \mid \alpha_{i0} \sim \mathcal{N}(\alpha_{i0}, \theta_0)$ is i.i.d over n and t, then the DGP of the observed data is:

$$y_{it}(\alpha_{i0}, \theta) = \alpha_{i0} + \sqrt{\theta} u_{it}, \quad u_{it} \sim \mathcal{N}(0, 1).$$

This equation cannot be simulated without information on the distribution of α_{i0}'s. The indirect fixed effect estimator uses $\widehat{\alpha}_i$'s instead, and the simulated data have the following DGP:

$$y^h_{it}(\widehat{\alpha}_i, \theta) = \widehat{\alpha}_i + \sqrt{\theta} u^h_{it}, \quad u^h_{it} \sim \mathcal{N}(0, 1),$$

where the superscript h denotes a simulation path. The fixed effect estimator using $\{y^h_{it}(\widehat{\alpha}_i, \theta)\}$ is

$$\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}}) := \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (y^h_{it}(\widehat{\alpha}_i, \theta) - \widehat{\gamma}_i)^2 = \frac{\theta}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (u^h_{it} - \frac{1}{T} \sum_{t=1}^{T} u^h_{it})^2,$$

where $\widehat{\boldsymbol{\alpha}} := (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)$ and $\widehat{\gamma}_i := \frac{1}{T} \sum_{t=1}^{T} y^h_{it}(\widehat{\alpha}_i, \theta)$. The interpretation of $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$ is that the estimator changes if a different value of θ is used to simulate the data. Note that the $\widehat{\alpha}_i$'s are fixed throughout the simulation process. The indirect fixed effect estimator $\widetilde{\theta}$ is the solution to

$$\widehat{\theta} = \widehat{\beta}^h(\widetilde{\theta}, \widehat{\boldsymbol{\alpha}}).$$

Figure (C·1) illustrates the issues with $\widehat{\theta}$ and the performance of $\widetilde{\theta}$ in this example. The density of the fixed effect estimator $\widehat{\theta}$ conveys two messages: (1) fixed effect estimator is subject to a large bias and (2) a confidence interval around $\widehat{\theta}$ would not have the correct coverage. The density of $\widetilde{\theta}$ illustrates that (1) the new estimator corrects the bias significantly and (2) a confidence interval around $\widetilde{\theta}$ would have a much larger coverage than the one around $\widehat{\theta}$.

Remark 1 (Caveat)**.** Due to the simple structure of the data, $\widehat{\theta}$ and $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$ have closed–form expressions, and $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}}) = \widehat{\beta}^h(\theta)$, i.e. the fixed effect estimator with the simulated data does not depend on the fixed effects. However, it is not generally the case in model (1.1).

1.3.2 Challenges due to Nonsmoothness with Incidental Parameters

Consider the binary choice panel model as a concrete example. Given θ, $\widehat{\alpha}_i$'s and x_{it}, the simulated dependent variable is

$$y_{it}^h(\theta, \widehat{\alpha}_i) = \mathbf{1}(x_{it}'\theta + \widehat{\alpha}_i > u_{it}^h), \quad u_{it}^h \sim \mathcal{N}(0,1),$$

where u_{it}^h are simulation draws from the standard normal distribution. The corresponding log likelihood function is

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}^h(\theta, \widehat{\alpha}_i)\log\Big(\Phi(x_{it}'\beta + \gamma_i)\Big) + (1 - y_{it}^h(\theta, \widehat{\alpha}_i))\log\Big(1 - \Phi(x_{it}'\beta + \gamma_i)\Big). \quad (1.4)$$

It illustrates the three different aspects of simulated fixed effect estimator $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$. Because $y_{it}^h(\theta, \widehat{\alpha}_i)$ is discontinuous in θ and $\widehat{\alpha}_i$'s, equation (1.4) is discontinuous, which carries over to its maximizer $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$. In addition, estimating $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$ involves incidental parameters γ_i's. The population version of equation (1.4) does not have

randomness due to data sampling, use of simulations and $\widehat{\alpha}_i$'s, and takes the form

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\Big[y_{it}(\theta,\alpha_{i0})\log\Big(\Phi(x_{it}'\beta+\gamma_i)\Big)+(1-y_{it}(\theta,\alpha_{i0}))\log\Big(1-\Phi(x_{it}'\beta+\gamma_i)\Big)\Big], \tag{1.5}$$

where the expectation is over u_{it}^h and x_{it}, and $\widehat{\alpha}_i$'s are replaced by α_{i0}'s.

Remark 2 (A comparison with panel quantile regression (QR) models)**.** One important type of nonlinear panel models that is not included in model (1.1) but also features nonsmoothness is panel QR models.[2] Kato et al. (2012) considered the following QR model with individual effects:

$$Q_\tau(y_{it}\mid x_{it},\gamma_{i0}(\tau))=\gamma_{i0}(\tau)+x_{it}'\beta_0(\tau),$$

where $\tau\in(0,1)$ is a quantile index, and $Q_\tau(y_{it}\mid x_{it},\gamma_{i0}(\tau))$ is the conditional τ–quantile of y_{it} given $(x_{it},\gamma_{i0}(\tau))$. The fixed effects quantile regression (FE–QR) estimator for this model is

$$(\widehat{\gamma}_{\text{FE-QR}},\widehat{\beta}_{\text{FE-QR}})\in\arg\min\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\rho_\tau(y_{it}-\gamma_i-x_{it}'\beta),$$

where $\gamma:=(\gamma_1,\ldots,\gamma_n)'$ and $\rho_\tau(u):=\{\tau-\mathbf{1}\{u\le 0\}\}u$ is the check function. The FE–QR estimator $\widehat{\beta}_{\text{FE-QR}}$ also contains estimated individual fixed effects and is non–smooth with respect to β or γ_i's because the check function is not differentiable. However, the estimator proposed in this paper is different in two aspects. First, the criterion function (1.4) is still smooth in β and γ_i's. Instead, the source of non–smoothness comes from data simulations. Second, in panel QR models, the DGP does not involve estimated individual fixed effects. Therefore, the theoretical results in Kato et al. (2012) do not apply here.

1.3.3 General Estimation Procedures

From the known distribution F_u, the simulated unobservables $\{u_{it}^h\}$ are independently drawn for $h=1,\ldots,H$, where H denotes the number of simulated panel data sets. For a given value of θ, let $y_{it}^h(\theta,\widehat{\alpha}_i)$ denote the simulated dependent variable for

[2] See Galvao and Kato (2018) for a recent survey.

simulation path h, then the sample log likelihood function using the h–th simulated data is

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\ln f(y_{it}^{h}(\theta,\widehat{\alpha}_i)\mid x_{it};\beta,\gamma_i), \tag{1.6}$$

where β and γ_i respectively denote the finite–dimensional parameter and incidental parameter in the simulation world. The fixed effect estimator to this problem is

$$\widehat{\beta}^{h}(\theta,\widehat{\boldsymbol{\alpha}}) = \arg\max_{\beta\in\mathbb{R}^p}\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\ln f(y_{it}^{h}(\theta,\widehat{\alpha}_i)\mid x_{it};\beta,\widehat{\gamma}_i(\beta,\theta,\widehat{\alpha}_i)),$$

where

$$\widehat{\gamma}_i(\beta,\theta,\widehat{\alpha}_i) = \arg\max_{\gamma\in\mathbb{R}}\frac{1}{T}\sum_{t=1}^{T}\ln f(y_{it}^{h}(\theta,\widehat{\alpha}_i)\mid x_{it};\beta,\gamma).$$

Repeating this estimation for all the simulated panel data, the following average can be computed:

$$\widehat{\beta}_H(\theta,\widehat{\boldsymbol{\alpha}}) := \frac{1}{H}\sum_{h=1}^{H}\widehat{\beta}^{h}(\theta,\widehat{\boldsymbol{\alpha}}),$$

and the indirect fixed effect estimator $\widetilde{\theta}^{H}$ is the solution to

$$\widehat{\theta} = \widehat{\beta}_H(\widetilde{\theta}^{H},\widehat{\boldsymbol{\alpha}}), \tag{1.7}$$

where the superscript H stresses that the finite–sample performance depends on the number of simulations conducted. The box below summarizes the steps required to compute the estimator.

Algorithm: Computing the indirect fixed effect estimator

(i) Obtain $(\widehat{\theta}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)$ using the observed data.

(ii) Set a random seed and H. For each i and t, draw unobservables $\{u_{it}^h\}_{h=1}^H$ from F_u.

(iii) Given θ and $\widehat{\alpha}_i$'s, use model (1.1) and $\{u_{it}^h\}$ to simulate dependent variable $\{y_{it}^h(\theta, \widehat{\alpha}_i)\}$; construct data $\{y_{it}^h(\theta, \widehat{\alpha}_i), x_{it}\}$, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

(iv) Obtain $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$ using the simulated data in Step (iii).

(v) Repeat steps (ii) and (iii) for all $h = 1, \ldots, H$ and solve for equation (1.7).

Remark 3 (Common random number)**.** Step (ii) follows the standard practice for simulations (Glasserman and Yao, 1992) by drawing unobserved shocks at the beginning of the algorithm. It implies that $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$ and $\widehat{\beta}^{h'}(\theta, \widehat{\boldsymbol{\alpha}})$ are independent for $h \neq h'$ conditional on x_{it}.

Remark 4 (The role of H)**.** The number H affects the finite–sample performance of the estimator, and increasing H reduces the asymptotic variance. Just like SMM and indirect inference, there is a trade off between precision of the estimator and intensity of computation. The estimation method, however, is different from the simulated maximum likelihood (Manski and Lerman, 1981), which is inconsistent for fixed H due to a nonlinear transformation of simulated choice probabilities.

1.4 Asymptotic Properties

This section starts with a discussion of the main assumptions that lead to theoretical properties of $\widehat{\theta}$ and $\widehat{\alpha}_i$'s. These assumptions are standard in large–T panel data models (Hahn and Kuersteiner, 2011), and they also impose certain structures that help establish the asymptotic properties of the indirect fixed effect estimator. Additional assumptions are imposed to ensure the simulations do not affect the panel data structure.

Assumption 1 (Large T asymptotics)**.** $n, T \to \infty$ *such that* $nT^{-1} \to \kappa \in (0, \infty)$.

Assumption 1 requires that the time series dimension grows at the same rate as the cross section dimension. The assumption defines the large–T asymptotics framework

and allows to transform the incidental parameter problem from a consistency to a bias problem, the latter of which can be quantified. From a practitioner's point of view, if the ratio T/n is not negligible, then it is reasonable to consider the large T asymptotics.

Assumption 2 (Sampling of observed data)**.** *(i)* $\{z_{it}\}_{t=1}^{\infty}$ *are independent across* i*; (ii) For each* i*,* $\{z_{it}\}_{t=1}^{\infty}$ *is a stationary* α*–mixing sequence with mixing coefficient* $\alpha_i(m)$ *such that* $\sup_i|\alpha_i(m)| \leq Ka^m$ *for some* a *such that* $0 < a < 1$ *and some* $K > 0$*.*

Assumption $2(i)$ imposes independence along the cross–section dimension. Assumption $2(ii)$ imposes a weak temporal dependence on each individual time series. The quantity $\alpha_i(m)$ measures for each i how much dependence exists between data separated by at least m time periods, and a uniform bound is imposed so as to bound covariances and moments when using law of large numbers (LLN) and central limit theorem (CLT). Interested readers can refer to Section 3.4 in White (2000) for definitions and properties. Note that Assumption 2 rules out non–stationary explanatory variables such as time effects and linear trends.

Assumption 3 (Identification)**.** *Denote* $G_{(i)}(\theta, \alpha_i) \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\ln f(y_{it} \mid x_{it}; \theta, \alpha_i)\Big]$*. For each* $\eta > 0$*,*

$$\inf_i \Big[G_{(i)}(\theta_0, \alpha_{i0}) - \sup_{(\theta,\alpha):\|(\theta,\alpha)-(\theta_0,\alpha_{i0})\|\geq\eta} G_{(i)}(\theta, \alpha)\Big] > 0.$$

Assumption 3 is a sufficient condition that ensures the log likelihood function admits a unique maximizer based on the time series variation. This assumption allows to prove the consistency of fixed effect estimators under large–T asymptotics. The indirect fixed effect estimator also requires this assumption for consistency.

Assumption 4 (Envelope condition)**.** *(i) The parameter* $\varphi_i := (\theta, \alpha_i) \in int\ \Theta \times \Gamma_\alpha$*, where* Θ *and* Γ_α *are compact, convex subsets of* $\mathbb{R}^p$ *and* $\mathbb{R}$ *respectively.*
(ii) There exists an envelope function $M(z_{it})$ *such that*

$$\|D^\nu \ln f(y_{it} \mid x_{it}; \varphi_1) - D^\nu \ln f(y_{it} \mid x_{it}; \varphi_2)\| \leq M(z_{it})\|\varphi_1 - \varphi_2\|,$$

where $\nu := (\nu_1, \ldots, \nu_{p+1})$ *is a vector of non–negative integers* ν_j *and* $|\nu| = \sum_{j=1}^{p+1} \nu_j$. *In addition, for* $\ln f(y_{it} \mid x_{it}; \varphi) := \ln f(y_{it} \mid x_{it}; \theta, \alpha_i)$,

$$D^{\nu} \ln f(y_{it} \mid x_{it}; \varphi) := \partial^{|\nu|} \ln f(y_{it} \mid x_{it}; \varphi) / (\partial \varphi_1^{\nu_1} \ldots \partial \varphi_{p+1}^{\nu_{p+1}})$$

and $|\nu| \leq 5$, *the function* $M(z_{it})$ *satisfies*

$$\sup_{\varphi_i \in int\Theta \times \Gamma_\alpha} \| D^{\nu} \ln f(y_{it} \mid x_{it}; \varphi) \| \leq M(z_{it})$$

and $\sup_i \mathbb{E}(|M(z_{it})|^{(10+10q)/(1-10\tilde{\nu})+\delta}) < \infty$ *for some integer* $q \geq p/2 + 2$, $\delta > 0$ *and* $0 < \tilde{\nu} < 1/10$.

Assumption $4(i)$ imposes compactness of parameter space, which is standard for establishing asymptotic properties of extremum estimators (Newey and McFadden, 1994). Compactness is convenient for proving uniform convergence with nonsmooth criterion functions (Newey, 1991). Assumption $4(ii)$ imposes a Lipschitz condition on the log likelihood function and a moment condition on the envelope function. This allows to establish uniform law of large number (ULLN) of sample log likelihood function and hence the pointwise consistency of $\widehat{\theta}$.

Under these assumptions and some regularity conditions on the Hessian matrix, Hahn and Kuersteiner (2011) established the following two results:

$$\max_{1 \leq i \leq n} |\widehat{\alpha}_i - \alpha_{i0}| = o_p(1), \tag{1.8}$$

$$\widehat{\theta} = \theta_0 + \frac{A(\theta_0, \boldsymbol{\alpha_0})}{\sqrt{nT}} + \frac{B(\theta_0, \boldsymbol{\alpha_0})}{T} + o_p\Big(\frac{1}{T}\Big), \tag{1.9}$$

where $\boldsymbol{\alpha_0} := (\alpha_{10}, \ldots, \alpha_{n0})$. Equation (1.8) states that the maximal deviation of $\widehat{\alpha}_i$ from α_{i0} converges to zero. This uniform consistency result is crucial for the theory of indirect fixed effect estimator because it justifies the usage of $\widehat{\alpha}_i$'s for data simulations. Equation (1.9) characterizes the asymptotic relationship between $\widehat{\theta}$ and θ_0. The term $A(\theta_0, \boldsymbol{\alpha_0})$ is the influence function that satisfies the central limit theorem (CLT) with

mean zero. The term $B(\theta_0, \boldsymbol{\alpha_0})$ converges to its expected value. Therefore, $\widehat{\theta}$ is consistent, asymptotically normal, but biased. Hahn and Kuersteiner (2011) derived the analytical forms of both terms, which are complicated functions of θ_0 and $\boldsymbol{\alpha_0}$.

Because the new estimator involves simulations, the following regularity condition is required so that the simulated data still maintain the mixing properties. Another regularity condition is that the parameter space in the simulation world is compact. Because (β, γ) is just a change of notation from (θ, α), this assumption is natural.

Assumption 5 (Simulation)**.** *(i) Assumption 2 holds for the simulated process for all $\theta \in \Theta$. (ii) The parameter spaces for β and γ_i, Θ_β and Γ_γ are compact.*

In sum, Assumption 5 allows for an asymptotic representation of simulated fixed effect estimator $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$ that resembles the one for $\widehat{\theta}$, i.e., equation (1.9).

Remark 5 (Inferring $\widetilde{\theta}^H$ from fixed effect estimators)**.** Backing out $\widetilde{\theta}^H$ from $\widehat{\beta}_H(\widetilde{\theta}^H, \widehat{\boldsymbol{\alpha}})$ requires an invertible relationship $\theta \mapsto \beta(\theta, \boldsymbol{\alpha_0})$, where $\beta(\theta, \boldsymbol{\alpha_0})$ is the maximizer of the limiting function for equation (1.6)

$$\frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E}\Big[\ln f(y_{it}(\theta, \alpha_{i0}) \mid x_{it}; \beta, \gamma_i) \Big].$$

The expectation is taken over simulation draws and sampling of observed data, and $\widehat{\alpha}_i$ is replaced with α_{i0}. This function is essentially the population function for the fixed effects estimation problem, except that β and γ_i are used to denote parameters for estimation in the simulation world. Assumption 3 thus ensures the uniqueness of $\beta(\theta)$, which is θ. As such, invertibility is satisfied trivially.[3] Although $\beta(\theta)$ is an identity function, for the rest of the paper this notation is kept to avoid the confusion between maximum of the limit and a parameter for data generation.

[3]For readers who are familiar with indirect inference, the relationship means that the binding function is an identity. This is because the auxiliary model is identical to the structural model, and thus the parameters in the two models coincide. Many papers that employ indirect inference often have to assume invertibility of the binding function (Collard-Wexler, 2013; Gospodinov et al., 2017), but this assumption is guaranteed in this paper.

1.4.1 Consistency

In order for the indirect inference–type estimator to be consistent, three conditions should be satisfied (Gouriéroux et al., 1993): an invertible relationship between θ and $\beta(\theta, \boldsymbol{\alpha_0})$, pointwise convergence of $\widehat{\theta}$ to $\beta(\theta_0, \boldsymbol{\alpha_0})$, and uniform convergence of $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$ to $\beta(\theta, \boldsymbol{\alpha_0})$ over the compact parameter space Θ. The first condition is satisfied because $\beta(\theta, \boldsymbol{\alpha_0})$ is an identity, and equation (1.9) gives the second condition. The following proposition states the uniform convergence condition.

Proposition 1.4.1 (Uniform convergence of fixed effect estimator using simulated data)**.** *Under Assumptions 1–5,*

$$\sup_{\theta \in \Theta} \| \widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}}) - \beta(\theta, \boldsymbol{\alpha_0}) \| \xrightarrow{p} 0.$$

The current proof specializes to panel models, but it is generalizable to other models that feature concavity and smoothness in (β, γ_i). Details are available in Appendix A.2, and here the main ideas are discussed.

Proving the uniform convergence condition with nonsmoothness requires two steps: pointwise convergence of $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$ to $\beta(\theta, \boldsymbol{\alpha_0})$, and a stochastic equicontinuity condition as follows:

$$\mathbb{E}\Big(\sup_{\|\theta_1 - \theta_2\| \leq \delta} \| \widehat{\beta}^h(\theta_1, \widehat{\boldsymbol{\alpha}}) - \widehat{\beta}^h(\theta_2, \widehat{\boldsymbol{\alpha}}) \| \Big) \leq C\delta, \tag{1.10}$$

where C is a constant and δ is a positive scalar.

Following the standard argument in Newey and McFadden (1994), pointwise convergence requires a ULLN result of log likelihood function using simulated data (1.4) to the limiting log likelihood (1.5). The log likelihood (1.4) has two sources of randomness: the first source comes from sampling variation of observed data, and the other is from simulations of unobservables. The non–standard part, however, is that data are simulated using $\widehat{\alpha}_i$'s. Therefore, it is necessary to first show that (1.4) uniformly

well approximates the log likelihood using data generated by α_{i0}'s:

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\int_{U}\Big[y_{it}^{h}(\theta,\alpha_{i0})\log\Big(\Phi(x_{it}'\beta+\gamma_i)\Big)+(1-y_{it}^{h}(\theta,\alpha_{i0}))\log\Big(1-\Phi(x_{it}'\beta+\gamma_i)\Big)\Big]dF_u, \tag{1.11}$$

where the integration is with respect to the distribution of simulation draws u_{it}^{h} to eliminate randomness from simulations. The details are available in Lemma A.1.1, and intuition is provided here. Because panel data are independent along the cross section, it suffices to show that each individual's log likelihood:

$$\frac{1}{T}\sum_{t=1}^{T}y_{it}^{h}(\theta,\widehat{\alpha}_i)\log\Big(\Phi(x_{it}'\beta+\gamma_i)\Big)+(1-y_{it}^{h}(\theta,\widehat{\alpha}_i))\log\Big(1-\Phi(x_{it}'\beta+\gamma_i)\Big),$$

satisfies this property. Given θ, this individual log likelihood is an additive and multiplicative combination of indicator functions of scalar $\widehat{\alpha}_i$ and smooth functions of (β,γ_i), which belongs to classes of functions that satisfy stochastic equicontinuity (van der Vaart and Wellner, 1996). Therefore, its empirical process:

$$\begin{aligned}\nu_T(\alpha_i)=\frac{1}{T}\sum_{t=1}^{T}\Big[&y_{it}^{h}(\theta,\alpha_i)\log\Big(\Phi(x_{it}'\beta+\gamma_i)\Big)+(1-y_{it}^{h}(\theta,\alpha_i))\log\Big(1-\Phi(x_{it}'\beta+\gamma_i)\Big)\\&-\int_{U}\Big(y_{it}^{h}(\theta,\alpha_i)\log\Big(\Phi(x_{it}'\beta+\gamma_i)\Big)+(1-y_{it}^{h}(\theta,\alpha_i))\log\Big(1-\Phi(x_{it}'\beta+\gamma_i)\Big)\Big)dF_u\Big],\end{aligned}$$

is stochastic equicontinuous. Combined with uniform consistency result of $\widehat{\alpha}_i$'s and LLN of $\nu_T(\alpha_{i0})$, an application of the triangular inequality leads to the uniform approximation result. Now that (1.11) only has randomness from observed data, its uniform convergence to the limiting log likelihood (1.5) follows the argument as in Hahn and Kuersteiner (2011). As such, the pointwise convergence of $\widehat{\beta}^{h}(\theta,\widehat{\boldsymbol{\alpha}})$ follows through.[4] To verify the stochastic equicontinuity condition (1.10), note that the

[4]Details are available in Lemma A.1.2.

profiled log likelihood:

$$\widehat{Q}(\beta;\theta) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\Big[y_{it}^{h}(\theta,\widehat{\alpha}_i)\log\Big(\Phi(x_{it}'\beta+\widehat{\gamma}_i(\beta))\Big) + (1-y_{it}^{h}(\theta,\widehat{\alpha}_i))\log\Big(1-\Phi(x_{it}'\beta+\widehat{\gamma}_i(\beta))\Big)\Big]$$

is concave in β. By definition, $\widehat{\beta}^h(\theta_1,\widehat{\boldsymbol{\alpha}})$ satisfies $\partial\widehat{Q}(\widehat{\beta}^h(\theta_1,\widehat{\boldsymbol{\alpha}});\theta_1)/\partial\beta = 0$. A first–order Taylor expansion with respect to $\widehat{\beta}^h(\theta_1,\widehat{\boldsymbol{\alpha}})$ around $\widehat{\beta}^h(\theta_2,\widehat{\boldsymbol{\alpha}})$ and positive–definiteness of the Hessian shows that $\widehat{\beta}^h(\theta_1,\widehat{\boldsymbol{\alpha}}) - \widehat{\beta}^h(\theta_2,\widehat{\boldsymbol{\alpha}})$ is bounded by

$$\Big|\frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2,\widehat{\boldsymbol{\alpha}});\theta_2)}{\partial\beta} - \frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2,\widehat{\boldsymbol{\alpha}});\theta_1)}{\partial\beta}\Big|,$$

which, by the Cauchy–Schwarz inequality, is bounded by the product of two terms: a smooth function of (β,γ_i) and

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}^{h}(\theta_1,\widehat{\alpha}_i) - y_{it}^{h}(\theta_2,\widehat{\alpha}_i)). \tag{1.12}$$

Therefore, it suffices to bound the two terms in expectation. The technical challenge mainly comes from proving this for equation (1.12) that features nonsmooth components. Although indicator functions are well–known to have controlled complexities (Andrews, 1994), and a similar result on the difference of indicator functions with univariate variable is given in Chen et al. (2003), here θ's can be multi–dimensional. It turns out that the expectation of (1.12) satisfies the L^2–smoothness regularity condition in Andrews (1994).[5] Therefore, the stochastic equicontinuity condition for $\widehat{\beta}^h(\theta,\widehat{\boldsymbol{\alpha}})$ is verified.

Armed with Proposition 1.4.1, the consistency of $\widetilde{\theta}^H$ follows the arguments as in Gouriéroux et al. (1993). The proof is straightforward because there is no need to consider a weighting matrix.

[5]See proof of Proposition 1.4.1 for details.

Theorem 1.4.1 (Consistency of indirect fixed effect estimator)**.** *Under Assumptions 1–5,*

$$\widetilde{\theta}^H \xrightarrow{p} \theta_0.$$

1.4.2 Bias Correction and Asymptotic Normality

Recall that the indirect fixed effect estimator using H simulations $\widetilde{\theta}^H$ is the solution to $\widehat{\theta} = \widehat{\beta}_H(\widetilde{\theta}^H, \widehat{\boldsymbol{\alpha}})$. Non–differentiability of $\theta \mapsto \widehat{\beta}_H(\theta, \widehat{\boldsymbol{\alpha}})$ means that the techniques in the indirect inference literature (e.g., Gouriéroux et al., 1993) are not applicable. The following stochastic equicontinuity assumption is imposed.

Assumption 6. *For all positive deterministic sequences $\delta_{nT} \downarrow 0$,*

$$\sup_{\|\theta_1-\theta_2\|\leq\delta_{nT}} \sqrt{nT}\|\widehat{\beta}_H(\theta_1, \widehat{\boldsymbol{\alpha}}) - \widehat{\beta}_H(\theta_2, \widehat{\boldsymbol{\alpha}}) - \mathbb{E}(\widehat{\beta}_H(\theta_1, \widehat{\boldsymbol{\alpha}}) - \widehat{\beta}_H(\theta_2, \widehat{\boldsymbol{\alpha}}))\| \xrightarrow{p} 0.$$

Assumption 6 requires that the difference between $\widehat{\beta}_H(\theta_1, \widehat{\boldsymbol{\alpha}})$ and $\widehat{\beta}_H(\theta_2, \widehat{\boldsymbol{\alpha}})$ can be approximated by its expectation at a $\sqrt{nT}$ rate. Combined with consistency of $\widetilde{\theta}^H$ and the mean value theorem, it allows to analyze the asymptotic normality of $\widetilde{\theta}^H$ through the lens of fixed effect estimators as follows:

$$\sqrt{nT}(\widetilde{\theta}^H - \theta_0) = \sqrt{nT}(\widehat{\theta} - \widehat{\beta}_H(\theta_0, \widehat{\boldsymbol{\alpha}})) + o_p(1). \tag{1.13}$$

Recall that equation (1.9) characterizes the representation of $\widehat{\theta} - \theta_0$. Because the same regression is run on simulated data h and the likelihood is smooth in (β, γ_i), the same structure of representation arises, namely that

$$\widehat{\beta}^h(\theta_0, \widehat{\boldsymbol{\alpha}}) - \theta_0 = \frac{A^h(\theta_0, \widehat{\boldsymbol{\alpha}})}{\sqrt{nT}} + \frac{B^h(\theta_0, \widehat{\boldsymbol{\alpha}})}{T} + o_p\Big(\frac{1}{T}\Big). \tag{1.14}$$

The terms $A^h(\theta_0, \widehat{\boldsymbol{\alpha}})$ and $B^h(\theta_0, \widehat{\boldsymbol{\alpha}})$ reflect that the data are generated using θ_0, $\widehat{\boldsymbol{\alpha}}$ and simulated unobservables $\{u_{it}^h\}$. A combination of (1.9), (1.13) and (1.14) therefore

leads to

$$\sqrt{nT}(\widetilde{\theta}^H - \theta_0) = \Big(A(\theta_0, \boldsymbol{\alpha_0}) - \frac{1}{H}\sum_{h=1}^{H} A^h(\theta_0, \widehat{\boldsymbol{\alpha}})\Big) + \sqrt{\frac{n}{T}}\Big(B(\theta_0, \boldsymbol{\alpha_0}) - \frac{1}{H}\sum_{h=1}^{H} B^h(\theta_0, \widehat{\boldsymbol{\alpha}})\Big) + o_p(1). \quad (1.15)$$

This equation reflects two observations. First, $\widehat{\theta}$ is unbiased if $B(\theta_0, \boldsymbol{\alpha_0})$ and $B^h(\theta_0, \widehat{\boldsymbol{\alpha}})$ both converge to the same limit. Second, $\widehat{\theta}$ is asymptotically normal if $A(\theta_0, \boldsymbol{\alpha_0})$ and $A^h(\theta_0, \widehat{\boldsymbol{\alpha}})$ converge to the same limiting distribution, but the variance is inflated by a factor of $1/H$. The rest of the section provides the main ideas of the proof.

The intuition can be gained by setting $H = 1$ and considering an infeasible fixed effect estimator $\widehat{\beta}_H(\theta_0, \boldsymbol{\alpha_0})$, which is obtained from data simulated by (θ_0, α_0). Then the representation of $\widehat{\beta}_H(\theta_0, \boldsymbol{\alpha_0}) - \theta_0$ takes the form

$$\sqrt{nT}(\widehat{\beta}_H(\theta_0, \boldsymbol{\alpha_0}) - \theta_0) = A^h(\theta_0, \boldsymbol{\alpha_0}) + \sqrt{\frac{n}{T}} B^h(\theta_0, \boldsymbol{\alpha_0}) + o_p(1).$$

The theory of indirect inference implies that $B(\theta_0, \boldsymbol{\alpha_0})$ and $B^h(\theta_0, \boldsymbol{\alpha_0})$ converge to the same probability limit. Because the actual simulated data are generated by $\widehat{\alpha}_i$'s, it suffices to show that $B^h(\theta_0, \widehat{\alpha})$ uniformly well approximates $B^h(\theta_0, \boldsymbol{\alpha_0})$ such that the approximation error is asymptotically negligible. More specifically, the bias term using simulated data takes the following form,

$$B^h(\theta_0, \widehat{\boldsymbol{\alpha}}) = -\Big[\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big]^{-1} \frac{1}{n}\sum_{i=1}^{n} B_i^h(\theta_0, \widehat{\alpha}_i),$$

where $\mathcal{I}_i(\theta_0, \widehat{\alpha}_i)$ is individual i's information matrix, and it is a smooth function of all its arguments. Therefore, $\mathcal{I}_i(\theta_0, \widehat{\alpha}_i) \xrightarrow{p} \mathcal{I}_i(\theta_0, \alpha_{i0})$ for each i. Each $B_i^h(\theta_0, \widehat{\alpha}_i)$ is nonsmooth in $\widehat{\alpha}_i$, and the following assumption is imposed such that $B^h(\theta_0; \widehat{\boldsymbol{\alpha}})$ replaces $B^h(\theta_0, \boldsymbol{\alpha_0})$ with negligible errors:

Assumption 7. *(Restricting changes due to using $\widehat{\alpha}_i$)*

$$\mathbb{E} \max_{1 \leq i \leq n} \left| B_i^h(\theta_0, \widehat{\alpha}_i) - B_i^h(\theta_0, \alpha_{i0}) \right|^2 = o_p(1)$$

Proposition 1.4.2 (Bias correction of $\widetilde{\theta}^H$)**.** *Under Assumptions 1–7,*

$$|B^h(\theta_0, \widehat{\boldsymbol{\alpha}}) - B^h(\theta_0, \boldsymbol{\alpha_0})| \xrightarrow{p} 0.$$

As such, the indirect fixed effect estimator corrects the bias. Hahn and Kuersteiner (2011) derived the analytical expression of the term $A(\theta_0, \boldsymbol{\alpha_0})$. The term $A^h(\theta_0, \widehat{\boldsymbol{\alpha}})$ has the same structure, namely

$$A^h(\theta_0, \widehat{\boldsymbol{\alpha}}) = \Big[\frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_i^h(\theta_0, \widehat{\alpha}_i)\Big]^{-1} \frac{1}{\sqrt{nT}} \sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}^h(\theta_0, \widehat{\alpha}_i),$$

where $A_{it}^h(\theta_0, \widehat{\alpha}_i)$ is a combination of high–order derivatives of the log likelihood. The following high–level assumption is imposed:

Assumption 8. $\frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} U_{it}^h(\theta_0, \widehat{\alpha}_i) = \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} U_{it}^h(\theta_0, \alpha_{i0}) + o_p(1)$.

Under this assumption, $A^h(\theta_0, \widehat{\boldsymbol{\alpha}})$ can uniformly well approximate $A^h(\theta_0, \boldsymbol{\alpha_0})$ with negligible errors, and the asymptotic normality result in indirect inference literature follows through (Gouriéroux et al., 1993, Proposition 5). Combined with Proposition 1.4.2, the indirect fixed effect estimator is asymptotically unbiased and normal.

Theorem 1.4.2. *Under Assumptions 1–8,*

$$\sqrt{nT}\Big(\widetilde{\theta}^H - \theta_0\Big) \xrightarrow{d} \mathcal{N}\Big(0, \Big(1 + \frac{1}{H}\Big)\Omega\Big),$$

where $\Omega := \mathbb{E}(A(\theta_0, \boldsymbol{\alpha_0})A(\theta_0, \boldsymbol{\alpha_0})')$.

The estimation of the variance–covariance matrix uses the estimated Hessian matrix of the sample log likelihood function from the real data. As previously discussed in Remark 4, the number of simulations H shows up as a factor that inflates the

asymptotic variance. There are two interpretations. The first is in line with other simulation–based methods: using simulations introduces an additional source of uncertainty and it is manifested through an increase in variance. The other interpretation is related to the trade–off between bias and variance. Because the indirect inference estimator debiases fixed effects estimator, the variance is larger, and it is quantified by the number of panel data simulated.

Dhaene and Jochmans (2015) proposed a half–panel method that removes the leading bias. Intuitively, the method splits in half the panel along the time series dimension, obtains the fixed effect estimators for the half samples and applies a linear combination with respect to the full–sample fixed effect estimator. Theoretically, the method does not change the asymptotic variance because the influence function is linear. However, in finite samples the variance is inflated due to an inefficient use of data. The indirect fixed effect method is explicit about the bias–variance tradeoff in that the asymptotic variance is multiplied by H. However, as will be shown in the simulations and applications, this method leads to a smaller standard error compared to the half–panel bias correction method.

1.5 Application to Female Labor Force Participation

Research on the relationship between female labor force participation and fertility is complicated by the presence of unobserved factors that affect both decisions. Following Hyslop (1999), this paper addresses the omitted variable issue by including individual fixed effects into the binary response panel model for the female labor force participation.

The data come from the Panel Study of Income Dynamics (PSID) and constitute a nine–year longitudinal sample spanning from 1979 to 1988. The sample includes 664 women aged 18—60 in 1985 who were continuously married with husbands in the

labor force in each of the sample periods and changed their labor force participation statuses. Consider the following static specification:

$$y_{it} = \mathbf{1}\{x_{it}'\theta + \alpha_i > u_{it}\}, \quad u_{it} \sim \mathcal{N}(0,1),$$

where y_{it} denotes the labor force participation indicator for woman i at time t, and x_{it} denotes a vector of time–varying covariates. These covariates include numbers of children of at most 2 years of age, between 3 and 5 years of age, between 6 and 17 years of age; log of the husband's income,[6] age and age squared. The individual effects α_i's are included to control for time–invariant unobserved heterogeneity such as willingness to work or ability.

Table (D.1) reports estimates of index coefficients using different methods. The standard errors are reported in parentheses. The standard errors for the fixed effects are computed from the Hessian of the profiled log likelihood. IFE–1, IFE–10 and IFE–20 denote indirect fixed effect estimators with H being 1, 10 and 20 respectively, and their respective standard errors are computed by multiplying the FE standard errors by $(1 + \frac{1}{H})$. For comparisons, the table includes results using half–panel jackknife method (HBC) (Dhaene and Jochmans, 2015), analytical bias correction (ABC) (Fernández-Val, 2009) and the leave–one–out jackknife method (BC–HN) (Hahn and Newey, 2004). The ABC has the same standard errors as the uncorrected fixed effect estimators, while the standard error computation for BC–HN and HBC follow the descriptions in Hahn and Newey (2004) and Dhaene and Jochmans (2015) respectively. The results show that the uncorrected estimates of index coefficients are about 15% larger (in absolute value) than their bias-corrected counterparts, indirect fixed effect estimators are closely comparable to ABC and BC–HN, and HBC produces estimates that are larger in magnitude. Because HBC achieves bias correction through sample

[6]This variable serves as a proxy for permanent nonlabor income (Hyslop, 1999).

splitting, the standard errors are larger.

1.6 Monte Carlo Simulations

This section considers Monte Carlo simulations calibrated to the same PSID data. The details of calibration procedures are available in Appendix E.1.1. The indirect inference fixed effect estimator is compared with the fixed effect estimation, the ABC and two jackknife bias correction methods. All simulations are done 1000 times and H is set to 10. The coverage reports the proportion of the times that θ_0 falls within the 95% confidence interval. All of the other statistics are relative to the true parameters and multiplied by 100.

Table (D.2) reports the simulation results of fixed effects and indirect fixed effect estimators. Fixed effect estimators are subject to a bias that is of the same order of magnitude as the standard deviation. This leads to severe under–coverage of the confidence intervals. The indirect fixed effect estimators, on the other hand, reduce bias by a margin without much inflation in the standard deviation. Therefore, the empirical coverage is close to the nominal value of 95%.

Table (D.3) tabulates the simulation results of ABC and two jackknife bias correction methods. Compared with IFE, ABC features smaller biases as it removes the bias term based on a plugged–in estimate, but standard deviations are comparable. Turning to the other two methods that automatically correct bias, first note that BC–HN admits smaller biases and standard deviations than HBC. The simulation results are in line with those reported in Hughes and Hahn (2020), who theoretically showed that HBC has a larger higher–order variance and remaining bias than BC–HN.[7] On the other hand, IFE is comparable with BC–HN in terms of both bias and standard deviation. A theoretical exploration is left for future work.

[7]Therefore, in practice it is recommended to use panel bootstrap to obtain standard errors for HBC.

The current theory is restricted to strictly exogenous explanatory variables, but Monte Carlo simulations in Appendix E.1.2 shows that the method can accommodate lagged dependent variables as well. Naturally, the next step is to extend the current theory to allow for dynamics in the DGP.

Like other bias correction methods, the theoretical properties of the indirect fixed effect estimator are predicated on the large–T assumption. Therefore, to compare how the estimator performs against other methods under varying lengths of time periods, this paper follows the literature (e.g., Hahn and Newey, 2004; Fernández-Val, 2009; Hughes and Hahn, 2020) and considers the following simulation design:

$$
\begin{aligned}
&y_{it} = \mathbf{1}\{\theta_0 x_{it} + \alpha_i - \varepsilon_{it} \geq 0\}, \quad \theta_0 = 1, \quad \alpha_i \sim \mathcal{N}(0,1), \quad \varepsilon_{it} \sim \mathcal{N}(0,1); \\
&x_{it} = t/10 + x_{i,t-1}/2 + u_{it}, \quad x_{i0} = u_{i0}, \quad u_{it} \sim U(-0.5, 0.5).
\end{aligned}
$$

The numerical experiments consider panels with $n = \{100, 200\}$ and $T = \{4, 8, 12\}$.

Table (D.4) reports the simulation results. The fixed effect estimators are subject to large biases, even when the time periods is 12. Because their biases are comparable to the standard deviations, fixed effect estimators exhibit undercoverage, which implies under–rejections. HBC has a poor performance when $T = 4$, and this is because each of the split sample only uses 2 time periods for estimation. HBC substantially reduces the bias when T is 8 or 12, but it is subject to a large dispersion, which reflects a larger confidence interval. As a result, the empirical coverage of HBC is larger than the nominal value of 95%.

The indirect fixed effect estimator has a better bias reduction performance compared to HBC, especially when $T = 4$. This means that the new estimator is less sensitive to the time periods, and thus can be potentially useful for short–panel applications as well. The number of simulation paths affects the dispersion. When $H = 1$, the estimator has a larger dispersion compared to the fixed effect estimator. When

$H = 10$, the dispersion is reduced. The trade–off for setting a large H in practice is an increase in computation time. For example, when $n = 200$ and $T = 12$, it took roughly sixteen minutes to obtain the results for $H = 10$ with ten cores on the MacBook Pro (M1, 2020). For $H = 1$, it took about two minutes. It is interesting to explore algorithms that efficiently search for solutions to non–smoonth functions, but this is beyond the scope of the paper.

It is worth noting that both the application and the simulation design feature non–stationary regressors. The results provide suggestive evidence that the indirect fixed effect estimator can accommodate these variables, which do not satisfy the stationarity condition in Assumption 2. However, the theoretical exploration is beyond the scope of this paper. In Chapter 2, the author proposes a new method called *crossover jackknife* that deals with non–stationarity explicitly.

1.7 Conclusion

Fixed effect estimations of nonlinear panel models are subject to large biases of point estimates and incorrect coverages of confidence intervals. This paper proposes a new estimator that reduces the bias and obtains standard errors without bootstrap.

There are at least three other questions for further explorations. First, average partial effects are often the quantities of interest in nonlinear models. This paper establishes theoretical properties of finite dimensional parameters, and it could be interesting to explore if they can be extended to handle average partial effects, which is a function of explanatory variables, parameters of interest and incidental parameters.

Second, this paper directly works with non–smooth log likelihood function and establishes the asymptotic properties of the new estimator. However, a practical concern of non–smoothness is that gradient–based optimization schemes cannot be used for estimation, and gradient–free schemes like Nelder–Mead face computational

difficulty in high–dimensional problems. The indirect fixed effect estimator might benefit from approaches like kernel smoothing, but the theoretical justification can be nontrivial as smoothing can introduce an additional bias.

Finally, incorporating unobserved heterogeneity into the dynamic discrete choice (DDC) models is an active area of research. One popular approach treats unobserved heterogeneity as an unobserved state variable and assumes individuals can be categorized into a finite number of types (Kasahara and Shimotsu, 2009; Arcidiacono and Miller, 2011). Introducing fixed effects circumvents the need to take a stand on the number of types, but can potentially complicate identification and estimation: the individual effects show up in both the current payoff and the continuation value, the latter of which has to be solved using a fixed–point algorithm. It would be exciting to investigate whether some of the ideas in this paper can be applied to incorporate fixed effects into DDC models.

Chapter 2

Crossover Jackknife Bias Correction for Non–Stationary Nonlinear Panel

Panel data offers the opportunity to control for multiple sources of unobserved heterogeneity. Researchers often include individual and time fixed effects into their models to account for unobserved variables that are either time or cross–sectional invariant. Throughout this chapter, we refer to this type of approach as the two–way fixed effects (TWFE). This approach is popular because it does not impose distributional assumptions on the unobserved effects; nor does it specify the relationship between unobserved effects and observed covariates. Instead, fixed effects are treated as parameters to be estimated. However, for dynamic linear and nonlinear panel models, the fixed effect (FE) estimators suffer from the incidental parameter problem. This problem leads to biased parameter estimates and incorrect coverages of confidence intervals.

Jackknife methods provide an automatic way to deal with the incidental parameter problem. They are popular because researchers do not have to derive the expression of the bias analytically, which can be non–trivial for complicated models. Instead, jackknife methods achieve bias corrections based on re–estimating the model using different sub–panels. Intuitively speaking, the FE estimator from each sub–panel magnifies the bias compared to the one that uses the full panel, so a linear combination of the FE estimators can remove the bias. The theoretical properties of the existing methods, however, rely on additional assumptions on the data that rule out covariates

with trends or known structural changes. Many economic variables like age or income exhibit trends. Moreover, the treatment variable in a difference–in–differences (DID) setting is non-stationary, with staggered dates of implementations or roll outs being an extra complication.

In this paper, we introduce a new jackknife bias correction method that accommodates non–stationary covariates. We name it *crossover jackknife* because it partitions the panel in two halves, each including half of the time series observations for each cross sectional unit, but where the time periods are crossed over between the two halves of the cross section units. Each sub–panel contains half of the data but retains the entire time periods, so the non–stationarity structure is preserved and a linear combination of the FE estimators using the full and sub–panels corrects the bias.

We derive the asymptotic properties of the method. Each sub–panel features a well–defined objective function, which allows us to derive the respective influence functions and asymptotic biases. Under the assumption that the data are identically distributed along the cross section, we show that an average of the FE estimators from the two sub–panels remove the bias. Similar to the current jackknife methods, our approach does not inflate the variance in theory, but we recommend practitioners to make inference using panel bootstrap for a better performance in finite samples.

We conduct two numerical simulations calibrated to real data. The first exercise features a dynamic panel Probit model that includes trended variables, while the second exercise features a dynamic linear panel with a staggered binary treatment. Simulation evidence indicates that our corrections improve the estimation and inference performance of the FE estimators. In addition, our method performs better than existing jackknife methods regarding the bias reductions of estimators and empirical coverage of confidence intervals. The comparison with the analytical bias correction method is more mixed. More specifically, our approach generally has a larger variance

but achieves better bias reduction.

Related Literature

This paper contributes to three stands of literature, and this section presents a non–exhaustive review. The incidental parameter problem is first discussed by Neyman and Scott (1948). For nonlinear or dynamic linear models with individual fixed effects, when the time series dimension T is fixed, FE estimators are in general inconsistent because each individual fixed effect is estimated with a fixed number of time series observations, and estimation errors do not vanish even when the cross–section sample size n is very large (Chamberlain, 1984; Lancaster, 2000). Only some special models like static linear and logit specifications feature fixed–T consistent estimators (Andersen, 1970), and many models are partially identified under fixed–T. (Honoré and Tamer, 2006; Chernozhukov et al., 2013).

A key insight of the large–T panel data literature is that the incidental parameter problem becomes an asymptotic bias problem when T grows with the sample size n. When n and T grow at the same rate, the FE estimators are consistent and asymptotically normal, but they have biases comparable to their standard deviations. In the search for asymptotically unbiased estimators, there are two leading approaches. For certain types of models, the bias terms have been characterized analytically and corrected using a plug–in approach (Hahn and Kuersteiner, 2002; Hahn and Newey, 2004; Fernández-Val, 2009; Hahn and Kuersteiner, 2011). In models with TWFE, since each time effect is estimated using n observations, another asymptotic bias shows up, and its form is derived analytically in Fernández-Val and Weidner (2016).

Under further sampling and regularity conditions, the bias terms can be automatically corrected using jackknife. First introduced in Quenouille (1949) for the cross section data, Hahn and Newey (2004) proposed a leave–one–out panel jackknife for data that do not have dependencies among observations of the same unit. Inspired

by Quenouille (1956)'s jackknife for time series data, Dhaene and Jochmans (2015) relaxed the assumption to stationarity along the time series, and proposed a split–panel method. Under an unconditional homogeneity assumption, Fernández-Val and Weidner (2016) allowed for TWFE and proposed a jackknife method that corrects the biases from both dimensions. Chudik et al. (2018) derived the exact–T bias of linear models with TWFE and weak exogeneity, and showed that a half–panel jackknife method effectively reduces the bias. See Fernández-Val and Weidner (2018) for a recent survey. Our new method allows for covariates with trends and known breaks, and hence is applicable to a larger class of empirical settings.

More generally, sample splitting is one of the most widely used methods in areas such as conformal inference (Lei et al., 2018; Barber et al., 2021), random forest (Wager et al., 2014) and causal inference that involves machine learning algorithms (Chernozhukov et al., 2018). Barber et al. (2021) showed that the reliability of jackknife for predictive inference depends on a stability condition: the fitted model and its leave–one–out version should have similar predictions at the test point. This rules out variables that could have breaks as well.

There is an active literature on DID with TWFE. See Freyaldenhoven et al. (2021), Roth et al. (2022) and Sun and Shapiro (2022) for three recent surveys. The consensus is that the FE estimator does not provide a valid economic interpretation when there is a staggered adoption of policies or heterogeneity in treatment effects. When the control variables include covariates that are time–varying based on the treatment participation, Callaway and Li (2021) and Caetano et al. (2022) showed that the parallel trend assumption is fragile, and conditioning parallel trends on the time–varying covariates is a better alternative. Our approach applies to dynamic linear panel models, which can incorporate such time-varying covariates naturally.

Structure of the Paper

The rest of the paper proceeds as follows: Section 2.1 sets up the model and illustrates the incidental parameter problem regarding the FE estimators. Section 2.2 introduces the crossover jackknife and illustrates how it accommodates variables with non–stationarity. Section 2.3 presents the asymptotic theory. Section 2.4 illustrates the finite–sample properties of the method via two calibrated Monte Carlo simulations. Section 2.5 concludes and discusses work in progress. Details of simulation procedures are in Appendix F.

2.1 Model Set–Up

We observe a panel data set $\{z_{it} := (y_{it}, x_{it}) : 1 \leq i \leq n, 1 \leq t \leq T\}$ for a scalar outcome variable of interest y_{it} and a vector of covariates x_{it}. The subscripts i and t index individual and time periods in traditional panel data, but they might index other dimensions in more general data structures such as firms and industries or siblings and families. The observations are independent across i and weakly dependent across t. Suppose we work with a panel data model with a d_β–vector of parameters of interest β, scalar individual fixed effect α_i and time effect γ_t for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Denote $\phi_{nT} = (\alpha_1, \ldots, \alpha_n, \gamma_1, \ldots, \gamma_T)'$. We consider an M–estimator

$$(\widehat{\beta}, \widehat{\phi}_{nT}) \in \underset{(\beta, \phi_{nT}) \in \mathbb{R}^{\dim \beta + \dim \phi_{nT}}}{\arg\max} \psi_{nT}(z_{it}; \beta, \phi_{nT}), \tag{2.1}$$

for some criterion function ψ_{nT}. We assume that $\psi_{nT}(\cdot)$ has a structure such that if n and T grow to infinity, the estimator $(\widehat{\beta}, \widehat{\phi}_{nT})$ is consistent for the true values of the parameters (β^0, ϕ^0_{nT}), which are the unique solutions, up to location nomalizations on

the individual and time effects, to the population program:[1]

$$\max_{(\beta,\phi_{nT})\in\mathbb{R}^{\dim\beta+\dim\phi_{nT}}} \mathbb{E}_{\phi}[\psi_{nT}(z_{it};\beta,\phi_{nT})]. \tag{2.2}$$

Here, $\mathbb{E}_{\phi}$ denotes the expectation with respect to the data distribution, conditional on the unobserved effects and initial conditions, which include strictly exogenous variables as a special case. We discuss two main examples.

Example 1 (Linear models with weak exogeneity)**.** Consider the specification:

$$y_{it} = \alpha_i + \gamma_t + x_{it}'\beta + u_{it}, \tag{2.3}$$

where u_{it} is normalized to have zero mean for each i and t and also satisfies the weak exogeneity condition

$$u_{it} \perp \mathcal{I}_{it}, \quad \mathcal{I}_{it} = \{\alpha_i, (\gamma_s, x_{is})_{s=1}^{t}\}, t = 1, \ldots, T.$$

As such, the covariate x_{it} allows for lagged dependent variables. The criterion function ψ_{nT} is the OLS objective function:

$$\psi_{nT}(z_{it};\beta,\phi_{nT}) = \frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - x_{it}'\beta - \alpha_i - \gamma_t)^2, \tag{2.4}$$

and the estimator of β^0 is

$$\widehat{\beta} = \Big(\sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}^{*}x_{it}^{*'}\Big)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}^{*}y_{it}^{*}.$$

Here, $x_{it}^{*} = x_{it} - \overline{x}_{\cdot t} - \overline{x}_{i\cdot} + \overline{x}$, where $\overline{x}_{\cdot t} = n^{-1}\sum_{i=1}^{n} x_{it}$ is the cross–section average at time t, $\overline{x}_{i\cdot} = T^{-1}\sum_{t=1}^{T} x_{it}$ is the temporal average for unit i, and $\overline{x} = (nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} x_{it}$ denotes the overall average. The variable y_{it}^{*} is defined in a similar way.

Example 2 (Nonlinear models)**.** For each i and t:

$$y_{it} \mid x_i^t, \alpha_i, \gamma_t, \beta \sim f_y(\cdot \mid x_{it}, \alpha_i, \gamma_t; \beta), \tag{2.5}$$

[1]See Section 2.3 for details.

where $x_i^t = (x_{i1}, \ldots, x_{it})$ and f_y is a known density with respect to some dominating measure. The covariates x_{it} are predetermined with respect to y_{it} and might include lags of y_{it} to accommodate dynamics. If the covariates x_{it} are strictly exogenous, then x_i^t can be replaced by $x_i^T := (x_{i1}, \ldots, x_{iT})$ in the conditioning set. The model is semiparametric because we do not specify the distribution of these effects or their relationship with x_{it}. The criterion function is the log likelihood function of the following form:

$$\psi_{nT}(z_{it}; \beta, \phi_{nT}) = \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f_y(y_{it} \mid x_{it}, \alpha_i, \gamma_t, \beta). \tag{2.6}$$

This example nests popular models like normal linear, probit, logit, tobit, negative binomial and Poisson. The specific form of f_y depends on the parametric family of distributions. For a binary response model,

$$f_y(y_{it} \mid x_{it}; \alpha_i, \gamma_t, \beta) = [F(x_{it}'\beta + \alpha_i + \gamma_t)]^{y_{it}} \times [1 - F(x_{it}'\beta + \alpha_i + \gamma_t)]^{(1-y_{it})},$$

where F is a CDF, e.g., the standard normal or standard logistic distribution.

2.1.1 Fixed Effect Estimators and Incidental Parameter Problem

To analyze the statistical properties of $\widehat{\beta}$, it is convenient to concentrate out ϕ_{nT}. For a given β, define

$$\widehat{\phi}_{nT}(\beta) \in \underset{\phi_{nT} \in \mathbb{R}^{\dim \phi_{nT}}}{\arg\max} \ \psi_{nT}(z_{it}; \beta, \phi_{nT}). \tag{2.7}$$

Hence, the FE estimators of β^0 and ϕ_{nT}^0 are

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^{\dim \beta}}{\arg\max}\, \psi_{nT}(z_{it}; \beta, \widehat{\phi}_{nT}(\beta)), \quad \widehat{\phi}_{nT} = \widehat{\phi}_{nT}(\widehat{\beta}). \tag{2.8}$$

We briefly review the incidental parameter problem regarding the FE estimators. Define

$$\overline{\beta} = \underset{\beta \in \mathbb{R}^{\dim \beta}}{\arg\max}\, \mathbb{E}_{\phi}[\psi_{nT}(z_{it}; \beta, \widehat{\phi}_{nT}(\beta))],$$

and note that $\overline{\beta} = \beta^0$ if $\widehat{\phi}_{nT}(\beta)$ were replaced by

$$\phi(\beta) = \underset{\phi \in \mathbb{R}^{\dim \phi_{nT}}}{\arg\max}\, \mathbb{E}_{\phi}[\psi_{nT}(z_{it}; \beta, \phi)].$$

However, under asymptotic approximations where either n or T is fixed, there is only a fixed number of observations to estimate some components in ϕ, namely T for each individual effect and n for each time effect. Therefore, $\widehat{\phi}_{nT}(\beta)$ is not a consistent estimator of $\phi(\beta)$. The nonlinearity of the model propagates the inconsistency to $\widehat{\beta}$.

Under the assumptions in Section 2.3, Fernández-Val and Weidner (2016) characterized the asymptotic distribution of $\widehat{\beta}$ as follows:

$$\widehat{\beta} - \beta^0 \overset{a}{\sim} -H^{-1} \cdot \mathcal{N}\Big(\frac{b}{T} + \frac{d}{n}, \frac{I_{d_\beta}}{nT}\Big). \tag{2.9}$$

Here, H denotes the Hessian matrix with respect to β and I_{d_β} denotes an identity matrix with dimension d_β. The bias term $-H^{-1}b/T$ comes from the individual fixed effects as there are only T observations that are informative about each of them. On the other hand, the bias term $-H^{-1}d/n$ comes from the time effects because there are n observations that are informative about each of them. Therefore, $-H^{-1}d/n$ is not present in Chapter 1, and is small relative to $-H^{-1}b/T$ when $n \gg T$. The bias $-H^{-1}b/T - H^{-1}d/n$ is of the same order as the standard deviation if $T \propto n$, and thus leads to incorrect coverages of inference methods based on the fixed effect estimators even when the sample is large.

2.2 Crossover Jackknife

In this section we give a graphical argument of the main results, leaving the technical details to Section 2.3. We first review the leading jackknife method for models with TWFE and discuss why it does not work with non–stationary variables. Then, we explain how the crossover jackknife method deals with this issue.

2.2.1 Split–Sample Jackknife

Based on the half–panel jackknife method of Dhaene and Jochmans (2015) for one–way model with only individual effects, Fernández-Val and Weidner (2016) proposed the split–sample bias correction (SBC) that can be visualized as in Figure (C·2).

The full panel is split in half along the time series dimension to deal with the bias coming from the individual fixed effects. Intuitively speaking, because each of the sub–panels (A_1 and A_2) contains half of the time series observations for i, the FE estimators $\widetilde{\beta}_{A_1}$ and $\widetilde{\beta}_{A_2}$ have twice the amount of bias than $\widehat{\beta}$, coming from estimation of α_i's. On the other hand, the full panel is split in half along the cross section dimension to deal with the bias due to time effects. Since each sub–panel (B_1 and B_2) only contains half of the individuals, the FE estimators $\widetilde{\beta}_{B_1}$ and $\widetilde{\beta}_{B_2}$ also have double the bias than $\widehat{\beta}$, coming from estimation of γ_t's. Hence, the SBC estimator takes the form:

$$\widetilde{\beta}_{SBC} = 3\widehat{\beta} - \frac{1}{2}(\widetilde{\beta}_{A_1} + \widetilde{\beta}_{A_2} + \widetilde{\beta}_{B_1} + \widetilde{\beta}_{B_2}). \tag{2.10}$$

To gain more intuition why this linear combination works, observe that

$$\widetilde{\beta}_{SBC} - \beta^0 = (\widehat{\beta} - \beta^0) - \Big(\frac{\widetilde{\beta}_{A_1} + \widetilde{\beta}_{A_2}}{2} - \widehat{\beta}\Big) - \Big(\frac{\widetilde{\beta}_{B_1} + \widetilde{\beta}_{B_2}}{2} - \widehat{\beta}\Big),$$

where $\frac{1}{2}(\widetilde{\beta}_{A_1} + \widetilde{\beta}_{A_2}) - \widehat{\beta} \approx -H^{-1}b/T$ and $\frac{1}{2}(\widetilde{\beta}_{B_1} + \widetilde{\beta}_{B_2}) - \widehat{\beta} \approx -H^{-1}d/n$ due to the statement in the previous paragraph. However, this idea is predicated on the following assumption on the data:

Assumption 1 (Unconditional Homogeneity)**.** *For each n and T, the sequences*

$$\{(y_{it}, x_{it}, \alpha_i, \gamma_t) : 1 \leq i \leq n; 1 \leq t \leq T\}$$

is identically distributed across i and strictly stationary across t.

Assumption 1 implies that $\beta^0_{A_1} = \beta^0_{A_2}$ and $\beta^0_{B_1} = \beta^0_{B_2}$, where $\beta^0_{\boldsymbol{S}}$ denotes the

probability limit of the FE estimator for the sub–panel $\boldsymbol{S}$. This assumption rules out time trends and structural changes in the processes for the unobserved effects and observed variables. Without this assumption, $\frac{1}{2}(\widetilde{\beta}_{A_1} + \widetilde{\beta}_{A_2}) - \widehat{\beta}$ might no longer be approximately equal to $-H^{-1}b/T$.

Remark 1 (Illustration of a dynamic linear model)**.** Chudik et al. (2018) derived the exact–T bias of the FE estimators for the linear models with weak exogeneity. To establish the theoretical guarantee of the half–panel jackknife, they also imposed an assumption that restricts the difference of dynamics in the two sub–panels (Chudik et al., 2018, equation 19). A counterexample that abstracts away fixed effects is

$$y_{it} = x_{it} + u_{it}, \quad x_{it} = \kappa_t \frac{u_{i,t-1}}{\sigma^2_{u_{i,t-1}}} + \xi_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T,$$

where ξ_{it} is independently distributed of $u_{i't'}$ for any i, i', t, t', $\kappa_t = \kappa_a$ for $t \leq T/2$ and $\kappa_t = \kappa_b$ for $t > T/2$. If $\|\kappa_a - \kappa_b\|$ is not $O(T^{-1})$, then the feedback from the errors to the future values of x_{it} systematically and sufficiently varies over time. As a result, half–panel jackknife cannot remove the bias.

2.2.2 Crossover Jackknife

Figure (C·3) is a graphical illustration of the crossover splitting. Here, we split the panel in four parts and construct the sub–panel that merges the two cross–over parts that do not have any indices i or t in common. The crossover splitting combines sample partitions along the cross section and time series dimensions, and hence can remove biases due to $\widehat{\alpha}_i$ and $\widehat{\gamma}_t$. However, this way of splitting preserves the time series structure of the panel in each sub–panel and hence allows for trends and breaks.

We form the crossover jackknife corrected estimator as follows:

$$\widehat{\beta}_{\text{CBC}} = 2\widehat{\beta} - \frac{1}{2}(\widetilde{\beta}_{\boldsymbol{S_1}} + \widetilde{\beta}_{\boldsymbol{S_2}}), \tag{2.11}$$

where $\widetilde{\beta}_{\mathbf{S}}$ denotes the fixed effect estimator using the sub–panel $\boldsymbol{S}$. To give an idea

how the bias correction works, note that

$$\widehat{\beta}_{CBC} - \beta^0 = (\widehat{\beta} - \beta^0) - \frac{1}{2}(\widetilde{\beta}_{\boldsymbol{S_1}} - \widehat{\beta}) - \frac{1}{2}(\widetilde{\beta}_{\boldsymbol{S_2}} - \widehat{\beta}),$$

where $\widetilde{\beta}_{\boldsymbol{S_1}} - \widehat{\beta} \approx -H^{-1}b/T - H^{-1}d/n$ and $\widetilde{\beta}_{\boldsymbol{S_2}} - \widehat{\beta} \approx -H^{-1}b/T - H^{-1}d/n$. Intuitively speaking, each sub–panel contains half of the individuals and time series observations, and thus both biases are doubled compared to that of $\widehat{\beta}$.

To operationalize this method, we need to impose the following assumption on the cross section dimension:

Assumption 2 (Identical distribution on the cross section)**.** *The sequence*

$$\{(z_{it}, \alpha_i, \gamma_t) : 1 \leq i \leq n, 1 \leq t \leq T\}$$

is identically distributed across i.

Compared to Assumption 1, we relax the stationarity assumption over t, but still require the cross section to have an identical distribution. If the cross–sectional units belong to known clusters like counties, then we can split the sample in a way that preserve such structures. However, this does not work if there is an unknown clustering.

2.3 Asymptotic Theory

For linear models with weak exogeneity, i.e., Example 1, fixed effects can be differenced out and the estimator $\widehat{\beta}$ can be obtained via least–squared estimation on demeaned data. In general, however, the population criterion function (2.2) does not have closed form solution, and thus we resort to asymptotic tools to establish the theoretical properties of the crossover jackknife method. Based on Fernández-Val and Weidner (2016), we impose the following assumptions and discuss how they differ from the ones in Chapter 1.

2.3.1 Technical Assumptions

Assumption 3 (Large T asymptotics)**.** *$n, T \to \infty$ such that $nT^{-1} \to \kappa$, where $\kappa \in (0, \infty)$.*

Assumption 3 defines the large–T asymptotic framework and is the same as in Chapter 1. This allows us to transform the incidental parameter problem from a consistency problem to an asymptotic bias problem, and the relative rate of n and T balances the order of the bias and variance and produces a non–degenerate asymptotic distribution. From a practitioner's standpoint, this assumption is reasonable if the ratio T/n is not negligible. For example, micro datasets like the Panel Study of Income Dynamics (PSID) now feature a long time series (around 30 years) for each household.

Assumption 4 (Sampling)**.** *Conditional on ϕ, $\{(x_i^T, y_i^T) : 1 \leq i \leq n\}$ is independent across i. For each i, $\{(x_{it}, y_{it}) : 1 \leq t \leq T\}$ is α–mixing with mixing coefficients satisfying $\sup_i a_i(m) = \mathcal{O}(m^{-\mu})$ as $m \to \infty$, where*

$$a_i(m) := \sup_t \sup_{A \in \mathcal{A}_t^i, B \in \mathcal{B}_{t+m}^i} |P(A \cap B) - P(A)P(B)|,$$

and for $z_{it} = (x_{it}, y_{it})$, $\mathcal{A}_t^i$ is the sigma field generated by $(z_{it}, z_{i,t-1}, \dots)$ and $\mathcal{B}_t^i$ is the sigma field generated by $(z_{it}, z_{i,t+1}, \dots)$.

Assumption 4 does not impose identical distribution nor stationarity over the time series dimension, conditional on the unobserved effects. This is unlike the sampling assumption in Chapter 1, which rules out time effects or structural breaks. The mixing condition is imposed to invoke LLN and CLT.

Assumption 5 (Asymptotic expansion of FE estimator)**.** *Let $\boldsymbol{S}$ denote a sub–panel, which includes the full panel as a special case. Let $|\boldsymbol{S}|$ denote the number of observations in the sub–panel $\boldsymbol{S}$. The FE estimator $\widehat{\beta}_{\boldsymbol{S}}$ for $\boldsymbol{S}$ admits an asymptotic expansion*

$$\sqrt{|\boldsymbol{S}|}(\widehat{\beta}_{\boldsymbol{S}} - \beta_0) = \frac{1}{\sqrt{|\boldsymbol{S}|}} \sum_{i,t \in \boldsymbol{S}} \varphi_{it} + \frac{1}{|\boldsymbol{S}|} \sum_{i,t \in \boldsymbol{S}} \Psi_{it} + R_{\boldsymbol{S}}, \tag{2.12}$$

where

$$\frac{1}{\sqrt{|\boldsymbol{S}|}}\sum_{i,t\in\boldsymbol{S}}\varphi_{it}\xrightarrow{d}\mathcal{N}(0,-H_{\boldsymbol{S}}^{-1}), \tag{2.13}$$

$$\frac{1}{|\boldsymbol{S}|}\sum_{i,t\in\boldsymbol{S}}\Psi_{it}\xrightarrow{p}-H_{\boldsymbol{S}}^{-1}(b_{\boldsymbol{S}}+d_{\boldsymbol{S}})=:-\mathcal{B}_{\boldsymbol{S}}-\mathcal{D}_{\boldsymbol{S}}, \tag{2.14}$$

and $|\boldsymbol{S}|R_{\boldsymbol{S}}\xrightarrow{p}0$ *for any* $\boldsymbol{S}$*. When* $\boldsymbol{S}$ *is the full panel, we drop the subscript* $\boldsymbol{S}$ *to ease the notation.*

Assumption 5 is a high–level condition that charaterizes the asymptotic behavior of the FE estimator. The term φ_{it} is the influence function while the term Ψ_{it} governs the asymptotic bias of the estimator. The remainder term $R_{\boldsymbol{S}}$ is asymptotically negligible. Hahn and Kuersteiner (2002) and Fernández-Val and Weidner (2016) provided primitive conditions for this expansion to hold for dynamic linear and nonlinear models respectively. However, this assumption rules out nonlinear panel models with a factor structure. See Chen et al. (2021) for more details.

2.3.2 Heuristics of Jackknife Methods

To understand why the SBC method requires the unconditional homogeneity assumption, assume the model only has individual fixed effects.[2] Then by Assumption 5, the joint distribution of the five FE estimators are

$$\begin{bmatrix}\widehat{\beta}-\beta^0\\ \widehat{\beta}_{A_1}-\beta^0\\ \widehat{\beta}_{A_2}-\beta^0\\ \widehat{\beta}_{B_1}-\beta^0\\ \widehat{\beta}_{B_2}-\beta^0\end{bmatrix}\overset{a}{\sim}\mathcal{N}\left(\begin{bmatrix}\frac{-H^{-1}b}{T}\\ \frac{-H_{A_1}^{-1}b_{A_1}}{T/2}\\ \frac{-H_{A_2}^{-1}b_{A_2}}{T/2}\\ \frac{-H_{B_1}^{-1}b_{B_1}}{T}\\ \frac{-H_{B_2}^{-1}b_{B_2}}{T}\end{bmatrix},-\begin{bmatrix}\frac{H^{-1}}{nT} & \frac{H_{A_1}^{-1}}{nT} & \frac{H_{A_2}^{-1}}{nT} & \frac{H_{B_1}^{-1}}{nT} & \frac{H_{B_2}^{-1}}{nT}\\ \frac{H_{A_1}^{-1}}{nT} & \frac{2H_{A_1}^{-1}}{nT} & 0 & \frac{H_{A_1}^{-1}}{nT} & \frac{H_{A_2}^{-1}}{nT}\\ \frac{H_{A_2}^{-1}}{nT} & 0 & \frac{2H_{A_2}^{-1}}{nT} & \frac{H_{B_1}^{-1}}{nT} & \frac{H_{B_2}^{-1}}{nT}\\ \frac{H_{B_1}^{-1}}{nT} & \frac{H_{A_1}^{-1}}{nT} & \frac{H_{B_1}^{-1}}{nT} & \frac{2H_{B_1}^{-1}}{nT} & 0\\ \frac{H_{B_2}^{-1}}{nT} & \frac{H_{A_2}^{-1}}{nT} & \frac{H_{B_2}^{-1}}{nT} & 0 & \frac{2H_{B_2}^{-1}}{nT}\end{bmatrix}\right),$$

[2]The bias due to time effects is corrected by splitting the sample along the cross section, which is not affected whether the covariates feature trends/breaks or not.

Recall that the SBC estimator takes the form

$$\widetilde{\beta}_{SBC} = 3\widehat{\beta} - \frac{1}{2}(\widehat{\beta}_{A_1} + \widehat{\beta}_{A_2}) - \frac{1}{2}(\widehat{\beta}_{B_1} + \widehat{\beta}_{B_2}),$$

and thus the bias of $\widetilde{\beta}_{SBC}$, after being multiplied by T, is

$$-3H^{-1}b + H_{A_1}^{-1}b_{A_1} + H_{A_2}^{-1}b_{A_2} + \frac{1}{2}H_{B_1}^{-1}b_{B_1} + \frac{1}{2}H_{B_2}^{-1}b_{B_2}. \tag{2.15}$$

Under Assumption 1, each sub–panel $\boldsymbol{S}$ features the same Hessian and bias as in the full–panel: $H_{\boldsymbol{S}} = H$, $b_{\boldsymbol{S}} = b$, and thus (2.15) is zero. Under Assumption 2,

$$H_{B_1}^{-1}b_{B_1} = H_{B_2}^{-1}b_{B_2} = H^{-1}b,$$

and thus (2.15) becomes $-2H^{-1}b + H_{A_1}^{-1}b_{A_1} + H_{A_2}^{-1}b_{A_2}$. Observe that

$$\begin{aligned} H^{-1}b &= (H_{A_1} + H_{A_2})^{-1}(b_{A_1} + b_{A_2}) \\ &= (H_{A_1} + H_{A_2})^{-1}H_{A_1}\mathcal{B}_{A_1} + (H_{A_1} + H_{A_2})^{-1}H_{A_2}\mathcal{B}_{A_2}. \end{aligned}$$

In order for the SBC to work, $H_{A_1} = (H_{A_1} + H_{A_2})/2$, which holds only if $H_{A_1} = H_{A_2}$ or $\mathcal{B}_{A_1} = \mathcal{B}_{A_2}$. However, the sub–panels A_1 and A_2 have different time series structure when there are trends or known breaks, and thus it is unlikely that the two Hessian matrices or the two biases are equal. On the other hand, crossover splitting preserves the time series structure in both sub–panels and thus provides a promising way to correct the bias. Denote $\mathcal{T}_1 = \{1, \ldots, \lfloor (T+1)/2 \rfloor\}$, $\mathcal{T}_2 = \{\lceil T/2 \rceil + 1, \ldots, T\}$, $\mathcal{N}_1 = \{1, \ldots, \lfloor (n+1)/2 \rfloor\}$, $\mathcal{N}_2 = \{\lceil n/2 \rceil + 1, \ldots, n\}$, $\mathcal{T}_0 = \mathcal{T}_1 \cup \mathcal{T}_2$ and $\mathcal{N}_0 = \mathcal{N}_1 \cup \mathcal{N}_2$. Let S_{jk} denote the sub–panel that contains the observations with cross sectional units in $\mathcal{N}_j$ and time series observations in $\mathcal{T}_k$ for $j, k \in \{0, 1, 2\}$. Then the two sub–sample in the crossover splitting are $\boldsymbol{S_1} := S_{11} \cup S_{22}$ and $\boldsymbol{S_2} := S_{12} \cup S_{21}$.

To characterize the asymptotic expansions of $\widehat{\beta}_{\boldsymbol{S_1}}$ and $\widehat{\beta}_{\boldsymbol{S_2}}$, we first consider the

properties of FE estimators $\widehat{\beta}_{jk}$ using the block S_{jk}. By Assumption 5,

$$\widehat{\beta}_{jk} - \beta^0 = \frac{2}{\sqrt{nT}} \sum_{i,t \in S_{jk}} \varphi_{it}^{jk} + \frac{4}{nT} \sum_{i,t \in S_{jk}} \Psi_{it}^{jk} + R_{nT/4},$$

where

$$\frac{2}{\sqrt{nT}} \sum_{i,t \in S_{jk}} \varphi_{it}^{jk} \xrightarrow{d} \mathcal{N}(0, -H_{jk}^{-1}), \quad \frac{4}{nT} \sum_{i,t \in S_{jk}} \Psi_{it}^{jk} \xrightarrow{p} -H_{jk}^{-1} b_{jk} - H_{jk}^{-1} d_{jk},$$

and $nTR_{nT/4} \xrightarrow{p} 0$. Each S_{jk} features its block–specific Hessian and bias terms.

Regarding $\widehat{\beta}_{\boldsymbol{S_1}}$ and $\widehat{\beta}_{\boldsymbol{S_2}}$, we currently impose a high–level condition on the probability limits of their respective Hessian matrices and bias terms. We are working on deriving more primitive conditions.

Assumption 6 (Asymptotic expansions of crossover sub–panel FE estimators)**.** *The estimators $\widehat{\beta}$, $\widehat{\beta}_{\boldsymbol{S_1}}$ and $\widehat{\beta}_{\boldsymbol{S_2}}$ satisfy:*

$$\begin{bmatrix} \widehat{\beta} - \beta^0 \\ \widehat{\beta}_{\boldsymbol{S_1}} - \beta^0 \\ \widehat{\beta}_{\boldsymbol{S_2}} - \beta^0 \end{bmatrix} \overset{a}{\sim} \mathcal{N} \left(\begin{bmatrix} -\frac{H^{-1}b}{T} - \frac{H^{-1}d}{n} \\ -\frac{2\mathcal{B}_{\boldsymbol{S_1}}}{T} - \frac{2\mathcal{D}_{\boldsymbol{S_1}}}{n} \\ -\frac{2\mathcal{B}_{\boldsymbol{S_2}}}{T} - \frac{2\mathcal{D}_{\boldsymbol{S_2}}}{n} \end{bmatrix}, \begin{bmatrix} \frac{-H^{-1}}{nT} & \frac{-H_{\boldsymbol{S_1}}^{-1}}{nT} & \frac{-H_{\boldsymbol{S_2}}^{-1}}{nT} \\ \frac{-H_{\boldsymbol{S_1}}^{-1}}{nT} & \frac{-2H_{\boldsymbol{S_1}}^{-1}}{nT} & 0 \\ \frac{-H_{\boldsymbol{S_2}}^{-1}}{nT} & 0 & \frac{-2H_{\boldsymbol{S_2}}^{-1}}{nT} \end{bmatrix} \right),$$

where

$$\mathcal{B}_{\boldsymbol{S_1}} = H_{\boldsymbol{S_1}}^{-1}(b_{12} + b_{21}), \quad \mathcal{D}_{\boldsymbol{S_1}} = H_{\boldsymbol{S_1}}^{-1}(d_{12} + d_{21}), \tag{2.16}$$

$$\mathcal{B}_{\boldsymbol{S_2}} = H_{\boldsymbol{S_2}}^{-1}(b_{11} + d_{22}), \quad \mathcal{D}_{\boldsymbol{S_2}} = H_{\boldsymbol{S_2}}^{-1}(d_{11} + d_{22}), \tag{2.17}$$

$$H_{\boldsymbol{S_1}} = H_{12} + H_{21}, \quad H_{\boldsymbol{S_2}} = H_{11} + H_{22}, \tag{2.18}$$

and

$$H_{11} = H_{21} = H_{01}, \quad H_{12} = H_{22} = H_{02}, \tag{2.19}$$

$$b_{11} = b_{21} = b_{01}, \quad b_{12} = b_{22} = b_{02}. \tag{2.20}$$

Assumption 6 contains two parts. First, for $\widehat{\beta}_{\boldsymbol{S_1}}$, its asymptotic variance is well

approximated by $-2(H_{S_{11}} + H_{S_{22}})^{-1}/nT$ and its two biases are well approximated by $-2(H_{S_{11}} + H_{S_{22}})^{-1}(b_{S_{11}} + b_{S_{22}})/T$ and $-2(H_{S_{11}} + H_{S_{22}})^{-1}(d_{S_{11}} + d_{S_{22}})/n$, respectively.[3]

Second, the probability limits of the biases and Hessians for $\widehat{\beta}_{\boldsymbol{S}}$ are not affected by how the panel is split along the cross section. This can be illustrated in Figure (C·4). The upper two subfigures show two possible splits along the time series. Without the strict stationarity assumption, The Hessians for sub–panels A_1, B_1, A_1' and B_2' can be all different. The bottom two subfigures present two possible splits along the cross section, conditional on the same split along the time series. Because the cross section units are identically distributed,

$$H_{A_{11}} = H_{A_{12}} = H_{A_1}, \quad H_{B_{11}} = H_{B_{12}} = H_{B_1},$$
$$H_{A'_{11}} = H_{A'_{12}} = H_{A_1}, \quad H_{B'_{11}} = H_{B'_{12}} = H_{B_1}.$$

In a similar vein, the corresponding b terms for the sub–panel exhibits a similar system of equalities. The d terms, on the other hand, exhibit such a property under Assumption 2. Because each sub–panel contains half of the data for each cross section unit, each $\widehat{\beta}_{\boldsymbol{S}}$ doubles the two biases compared to $\widehat{\beta}$.

Theorem 2.3.1 shows that the CBC method corrects the bias and does not inflate the asymptotic variance. The proof is straightforward and thus is included in the main text.

Theorem 2.3.1. *Under Assumptions 2 – 6,*

$$\widetilde{\beta}_{CBC} - \beta_0 \overset{a}{\sim} \mathcal{N}\Big(0, -\frac{H^{-1}}{nT}\Big).$$

[3]We skip the discussion on $\widehat{\beta}_{\boldsymbol{S_2}}$ as the arguments apply symmetrically.

Proof. Under Assumptions 2 – 6,

$$\begin{bmatrix} \widehat{\beta}-\beta^0 \\ \widehat{\beta}_{\boldsymbol{S_1}}-\beta^0 \\ \widehat{\beta}_{\boldsymbol{S_2}}-\beta^0 \end{bmatrix} \overset{a}{\sim} \mathcal{N}\left(\begin{bmatrix} -\frac{H^{-1}b}{T}-\frac{H^{-1}d}{n} \\ -\frac{2H^{-1}b}{T}-\frac{2H^{-1}d}{n} \\ -\frac{2H^{-1}b}{T}-\frac{2H^{-1}d}{n} \end{bmatrix}, \begin{bmatrix} \frac{-H^{-1}}{nT} & \frac{-H^{-1}}{nT} & \frac{-H^{-1}}{nT} \\ \frac{-H^{-1}}{nT} & \frac{-2H^{-1}}{nT} & 0 \\ \frac{-H^{-1}}{nT} & 0 & \frac{-2H^{-1}}{nT} \end{bmatrix} \right).$$

Then by the properties of multivariable normal distribution,

$$\widetilde{\beta}_{CBC} := 2\widehat{\beta} - \frac{1}{2}(\widehat{\beta}_{\boldsymbol{S_1}} + \widehat{\beta}_{\boldsymbol{S_2}})$$

corrects the bias without variance inflation. Indeed,

$$[2 \quad -1/2 \quad -1/2] \begin{bmatrix} -\frac{H^{-1}b}{T}-\frac{H^{-1}d}{n} \\ -\frac{2H^{-1}b}{T}-\frac{2H^{-1}d}{n} \\ -\frac{2H^{-1}b}{T}-\frac{2H^{-1}d}{n} \end{bmatrix} = 0,$$

and

$$[2 \quad -1/2 \quad -1/2] \begin{bmatrix} \frac{-H^{-1}}{nT} & \frac{-H^{-1}}{nT} & \frac{-H^{-1}}{nT} \\ \frac{-H^{-1}}{nT} & \frac{-2H^{-1}}{nT} & 0 \\ \frac{-H^{-1}}{nT} & 0 & \frac{-2H^{-1}}{nT} \end{bmatrix} \begin{bmatrix} 2 \\ -1/2 \\ -1/2 \end{bmatrix} = \frac{-H^{-1}}{nT}.$$

□

2.4 Calibrated Monte Carlo Simulations

This section documents two calibrated simulation exercises. For dynamic probit panel, we use the labor force participation (LFP) data from Fernández-Val (2009). For the dynamic linear panel, we use a balanced panel subset of the democracy data from Acemoglu et al. (2019). In both designs, data are simulated 500 times with panel ids being reshuffled and the initial conditions are set to be the original observations in the data. We use nonparametric panel bootstrap to compute bootstrap standard errors (normalized interquartile range based on 200 repetitions).[4] Appendix F.1 provides

[4]For democracy application, we further implement the panel weighted bootstrap for comparison.

detailed procedures of the simulations.

We report results for uncorrected FE estimators, analytical bias corrected (ABC), split-sample bias correction (SBC) based on Fernández-Val and Weidner (2016) and crossover bias correction (CBC). For dynamic linear panel, we also compute the Arellano–Bond (AB) estimator (Arellano and Bond, 1991).

2.4.1 Dynamic Probit Panel

Using the same PSID data from 1980–1988 as in Chapter 1, we first consider a dynamic probit model with strictly exogenous covariates and unobserved individual and time fixed effects. The regression specification takes the following form:

$$y_{it} = \mathbf{1}\{\beta_y y_{i,t-1} + \beta_x x_{it} + \alpha_i + \gamma_t \geq \varepsilon_{it}\},$$

where y_{it} denotes binary indicator of female labor force participation, x_{it} includes three fertility variables (the numbers of children aged 0–2, 3–5 and 6–17), the log of the husband's income, and a quadratic function of age, α_i and γ_t respectively denote individual and time fixed effects, and errors ε_{it} are independent draws from the standard normal distribution. The age and income variables are trended. The parameters $(\beta_y, \beta_x, \alpha_i, \gamma_t)$ are calibrated to the FE probit estimates in the PSID 1980–1988 with the observed labor force participation indicator being the dependent variable.

Table (D.7) reports biases, standard deviations, root MSEs (RMSE) and empirical coverage probabilities of confidence intervals with nominal level of 95% for the estimators. All the results, except for the coverage probabilities, are in percentage of the true value of the parameters. We find that the ABC overall does a good job reducing the bias without increase in dispersion. CBC generally improves over SBC in terms of bias reduction and outperforms ABC for many coefficients, but it is subject to a larger dispersion compared to the ABC.

2.4.2 Dynamic Linear Panel

We revisit the application to the causal effect of democracy on economic growth of Acemoglu et al. (2019). To keep the analysis simple, we use a balanced sub–panel of 147 countries over the period from 1987 through 2009 extracted from the data set used in the original paper. Following Acemoglu et al. (2019), we consider the dynamic linear panel specification:

$$y_{it} = \alpha_i + \gamma_t + \beta x_{it} + \sum_{l=1}^{4} \rho_l y_{i,t-l} + \varepsilon_{it},$$

where y_{it} denotes log GDP per capita for a country i in year t and x_{it} denotes the binary treatment variable that equals 1 if country i in year t is considered a democracy and 0 otherwise. The model includes four lagged dependent variables, country and year fixed effects. We assume each idiosyncratic error ε_{it} is normalized to have zero mean and satisfies the weak sequential exogeneity assumption:

$$\varepsilon_{it} \perp \mathcal{I}_{it}, \quad \mathcal{I}_{it} := \{(x_{is}, \gamma_s, y_{s-1}, y_{s-2}, y_{s-3}, y_{s-4})_{s=1}^{t}, \alpha_i\}.$$

This assumption implies that (1) democracy and past GDP are orthogonal to contemporaneous and future GDP shocks and that (2) these shocks are serially uncorrelated. The parameters $(\alpha_i, \gamma_t, \beta, \rho_1, \rho_2, \rho_3, \rho_4)$ are calibrated to the dynamic linear estimates in the balanced democracy data by Acemoglu et al. (2019) with the observed log GDP being the dependent variable. We also consider a long–run effect of the treatment as follows:

$$\beta / \Big(1 - \sum_{j=1}^{4} \rho_j\Big),$$

whose true value in the simulations is constructed using the calibrated β and ρ_l's. The treatment variable is not stationary as some countries either change their status or switch back and forth over the years. This non–stationarity is visualized in

Figure (C·5).[5]

Table (D.8) shows the results with nonparametric bootstrap standard errors. Arellano–Bond estimators can have even larger bias than fixed effect estimation due to the many instruments problem. The CBC drastically reduces bias and performs better than the SBC for all coefficients except for the one of the long run effect. However, the CBC increases dispersions compared to the FE estimators. The ABC reduces the biases in general without an inflation in standard deviations, but its bias reduction performance is worse than the CBC.

2.5 Conclusion

Fixed effect estimators for nonlinear and dynamic linear panel models suffer from the incidental parameter problem. Jackknife bias correction methods can automatically correct the bias, but existing methods cannot allow for variables that have trends or structural changes. We propose a new method to accommodate these variables.

There are at least three avenues for future research. First, for nonlinear panel models, the ultimate parameters are often the *ceteris paribus* or partial effects, i.e. effects in the outcome of changing each covariate while holding the rest of covariates and unobserved effects fixed (Chamberlain, 1984). Fernández-Val and Weidner (2016) showed that the asymptotic bias of estimated average partial effects can be removed using jackknife methods. With some modifications, we are optimistic about extending the crossover jackknife to bias correct average partial effects as well.

Second, the derivation implies that bias correcting the score function directly can also reduce the bias of the fixed effect estimators. Due to the incidental parameter problem, the score function is not centered around zero as well. Dhaene and Jochmans (2015) developed jackknife methods to correct the scores in models with individual

[5]The graph is plotted using the R package from Imai et al. (2021).

effects, and it is promising to develop a crossover method for models with TWFE.

Third, we consider the simple case of estimating a common vector of parameters, but it is worth investigating if the method can be applied to models with heterogeneous treatment effects and individual–specific non–stationarity. As pointed out by Freyaldenhoven et al. (2021), one restriction of the leading specifications in the event–study literature is that they do not include lagged dependent or predetermined variables. It is interesting to synergize the two literature and explore the implications for microeconomic heterogeneity.

Chapter 3

Robust Tests of Model Incompleteness in the Presence of Nuisance Parameters

3.1 Introduction

Models of discrete choice are used widely. In empirical studies, a common strategy is to combine a theory of choice (e.g., utility maximization) that predicts a unique outcome value with distributional assumptions on latent variables. This approach allows a researcher to describe the conditional distribution of the outcome given observable covariates. However, recent economic applications often involve models that permit multiple outcome values, which we call an *incomplete prediction*. Such an incomplete prediction occurs when the researcher is willing to work only with weak assumptions or has limited knowledge of the DGP.

More specifically, we consider a form of incompleteness summarized as follows: an observable discrete outcome variable $Y \in \mathcal{Y}$ satisfies

$$Y \in G(u|X;\theta), \tag{3.1}$$

where G collects all outcome values that are compatible with the model given the unobserved and observed variables (u, X) and a structural parameter θ. This structure arises in a variety of contexts. Multiple outcomes are predicted in single-agent discrete choice models when the agent's choice set is unobservable and consistent with a wide range of choice set formation processes (Barseghyan et al., 2021). In discrete

games such as firms' market entry or household's labor supply decisions, multiple equilibria may exist, but one may not know how an equilibrium outcome gets selected (Bresnahan and Reiss, 1991a; Ciliberto and Tamer, 2009). In panel dynamic discrete choice models, one's theory may be silent about how an initial observation is generated (Heckman, 1978; Honoré and Tamer, 2006).

Recent empirical studies have fruitfully applied econometric methods for such incomplete models in different areas, including English auctions (Haile and Tamer, 2003), strategic voting (Kawai and Watanabe, 2013), product offerings (Eizenberg, 2014; Wollmann, 2018), network formation (de Paula et al., 2018; Sheng, 2020), school choices (Fack et al., 2019), and major choices (Henry et al., 2020).

A natural question is whether a model needs to allow incompleteness to be consistent with data. This question itself motivates tests of the model completeness. Furthermore, whether a model is complete or not is often closely related to policy-relevant features of the underlying structural model. Hence testing the model completeness against incompleteness may provide useful information for the practitioners. For example, in a commonly used market entry model, multiple equilibria exist only if the firms interact strategically. Testing the presence of the strategic interaction effects and inferring their signs can provide critical information for policymaking (de Paula and Tang, 2012). In a triangular model with a binary outcome and a binary treatment variable, we show that taking a control function approach yields an incomplete model only if the treatment assignment is endogenous. Detecting the endogeneity of treatments can help the practitioner choose a suitable framework for evaluating the treatment effects.

In many of these examples, one can state the null hypothesis of model completeness as restrictions on the structural parameter's subvector (called β). We develop a computationally tractable test for such restrictions. The test is based on a novel

score statistic. Advantages of this approach are (i) the score statistic only requires estimation of nuisance parameters in the restricted model, which is complete; (ii) the nuisance parameters can be estimated by standard point estimators (e.g., restricted MLE) using package software; and (iii) one can simulate the statistic's limiting distribution easily.

The basic idea behind our test is as follows. The model incompleteness, in general, implies multiple (typically infinitely many) likelihood functions, which makes it challenging to apply standard likelihood-based tests. However, the class of models we consider has properties that make score-based tests attractive. First, under any value of structural parameter $\theta_0 + h$ violating the null hypothesis locally, there is a "least favorable" data generating density q_{θ_0+h} that is the most difficult to distinguish from the density q_{θ_0} under the null hypothesis. We may then view the map $\theta \mapsto q_\theta$ as a "least favorable parametric model", along which detecting the deviation from θ_0 is most difficult. We show that one can explicitly derive such a model by solving a convex program, which allows one to calculate its score. Second, we show that the score-based test maximizes a measure of local discrimination (between θ_0 and $\theta_0 + h$) based on the least favorable parametric model. The resulting test is robust because it detects any local deviation from the null hypothesis regardless of how Y is selected from the predicted set $G(u|X, \theta)$.

As in other subvector inference problems, we need to deal with nuisance components δ of the parameter vector. Exploiting the property that the model is complete under the null hypothesis, we show that one can construct a $\sqrt{n}$-consistent point estimator of δ and plug it into the score. This plug-in procedure is computationally tractable because it avoids evaluating the test statistic over a grid of nuisance parameters. Among $\sqrt{n}$-consistent estimators of δ, we recommend using the *restricted maximum likelihood estimator* (MLE). This estimator maximizes the likelihood sub-

ject to the restrictions under the null hypothesis and can be computed using standard package software. The score-based statistic has a limiting distribution that can be easily simulated if one uses the restricted MLE. For other estimators of δ, we provide an orthogonalized version of the score test following the insights of Neyman's $C(\alpha)$-test (Neyman, 1959, 1979). The orthogonalization makes the distribution of the statistic insensitive to the effects of the estimated nuisance parameters, and its limiting distribution is also easy to simulate.

While this paper focuses on testing the model completeness, estimating some or all components of θ may be the researcher's ultimate goal in some applications. In such a case, our restricted maximum likelihood estimator provides a consistent estimator of δ if the null hypothesis is true. In contrast, one can use robust subvector inference methods in the literature if the alternative hypothesis is true. Our test, therefore, can be viewed as a specification test, which naturally raises a question regarding its impact on any post-model selection inference. While we defer a formal analysis to another work, we propose a hybrid procedure that aims at controlling the potential distortion of the model selection step using a shrinkage method borrowing the insights from the moment selection literature (Andrews and Soares, 2010; Romano et al., 2014).

Related Literature

Our paper belongs to the literature on inference in incomplete models. The seminal work of Tamer (2003) showed an incomplete model induces multiple distributions and implies partially identifying restrictions on parameters. Recent developments in the literature (Galichon and Henry, 2011; Beresteanu et al., 2011; Chesher and Rosen, 2017) provided tools to systematically derive so-called *sharp identifying restrictions*, which convert all model information into a set of equality and inequality restrictions on the conditional moments of the observables. Inference methods based on the sample analogs of such moment restrictions are extensively studied (see Canay

and Shaikh, 2017, and references therein). Our approach builds on the recent developments on likelihood-based inference methods for incomplete models (Chen et al., 2018; Kaido and Zhang, 2019). In particular, we combine the sharp identifying restrictions with the tools from Kaido and Zhang (2019) (KZ19, henceforth) to derive the least favorable parametric model. We then construct a test statistic using the score function associated with the least favorable parametric model. To our knowledge, this approach is new. One can view our procedure as an analog of deriving a score function using a parametric specification in a complete model.

Hypothesis testing in incomplete models has been studied extensively. As discussed earlier, many of them are based on the sample analogs of conditional or unconditional moment restrictions. One of the challenges surrounding (subvector) inference is the high computational cost for implementing the existing methods (Molinari, 2020, sec. 6). There are attempts to improve the computational tractability of the moment-based inference methods within a particular class of models or testing problems. In particular, Andrews et al. (2019) and Cox and Shi (2020) assumed that moment inequality restrictions implied by the model are linear conditional on some observable variables. This paper focuses on another class in which the model is complete under the null hypothesis. This structure allows us to make our test computationally tractable by combining (i) the score function associated with the least favorable parametric model and (ii) a point estimator of the nuisance components.

Practitioners can use this paper's framework to test a variety of hypotheses. For example, one can test the presence of strategic interaction effects and multiple equilibria in static complete information games. Related problems have been studied in other classes of models. For incomplete information games, de Paula and Tang (2012) introduced a semiparametric inference procedure on the signs of strategic interaction effects. For finite-state Markov games, Otsu et al. (2016) provided techniques to

test whether the conditional choice probabilities, state transition, and other features of games are homogeneous across cross-sectional units. Rejection of their null hypothesis could occur when multiple equilibria are present. In the context of network formation with many agents, Pelican and Graham (2021) developed a procedure to test whether agents' preferences over networks are interdependent. Using a Logit specification, they proposed conditional tests and introduce an MCMC algorithm to implement their test. One can also apply our framework to triangular systems involving a binary outcome and a binary endogenous variable. We show that taking a control function approach in such a setting leads to a model with an incomplete prediction. Namely, the model involves a set-valued control function. Our framework can be used to test the endogeneity of treatment assignments with weak assumptions. To our knowledge, this test is new to the literature and provides an alternative to the existing proposal by Wooldridge (2014) who made additional high-level assumptions.

Structure of the Paper

The rest of the paper proceeds as follows: Section 3.2 introduces the set up with motivating examples. Section 3.3 defines the hypothesis testing problem and discusses our new test statistic. Section 3.4 illustrates the usage via two empirical applications. Section 3.5 presents simulation evidence for the local power property. Section 3.6 concludes. Details of proofs, derivations and computations are available in Appendices B and G.

3.2 Set Up

Let Y be a discrete outcome taking values in a finite set $\mathcal{Y}$. Let $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ be a vector of observable covariates and let $u \in U \subseteq \mathbb{R}^{d_U}$ be a vector of unobservable variables. Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ be a finite–dimensional parameter.

The prediction of a structural model is summarized by a weakly measurable set-valued map $G : U \times \mathcal{X} \times \Theta \rightsquigarrow \mathcal{Y}$. We assume that Y takes one of the values in $G(u|X;\theta)$ with probability one. The map G describes how the observable and unobservable characteristics of individuals and economic environments translate into a set of possible outcome values. It reflects restrictions imposed by theory such as the functional form of utility/profit functions, forms of strategic interaction, and any equilibrium or optimality concepts. It is important to note that G can be set-valued. This feature allows to encode the researcher's lack of understanding of some part of the structural model. More specifically, it does not require the knowledge of the *selection mechanism* according to which the observed outcome y is *selected* from $G(u|x;\theta)$. This formulation nests the classic setting in which the model is characterized by a *reduced form equation*:

$$Y = g(u|X;\theta), \tag{3.2}$$

for a function $g : U \times \mathcal{X} \times \Theta \to \mathcal{Y}$. This corresponds to the setting in which G is almost surely singleton-valued, i.e. $G(u|x;\theta) = \{g(u|x;\theta)\}, a.s.$ If this is the case, we say the model makes a *complete prediction*.

Throughout, we assume that u's law belongs to a parametric family $F = \{F_\theta, \theta \in \Theta\}$, where, for each θ, F_θ is a probability distribution on U. To keep notation concise, we use the same θ for parameters that show up in G and that index F_θ. Also, we focus on settings in which u is independent of X. However, the framework can be easily extended to settings where u is correlated with X, and the researcher specifies its conditional distribution $F_\theta(u|x)$. Furthermore, we note that our framework does accommodate settings in which some of the observable covariates are endogenous, but one can construct a set-valued control function (See Example 3 below).

3.2.1 Motivating Examples

Below, we illustrate the objects introduced above with examples studied in the literature. Our first two examples are discrete games of complete information (Bresnahan and Reiss, 1991a; Ciliberto and Tamer, 2009).

Example 1 (Discrete Games of Strategic Substitution)**.** There are two players (e.g. firms). Each player may either choose $y^{(j)} = 0$ or $y^{(j)} = 1$. The payoff of player j is

$$\pi^{(j)} = y^{(j)}\big(x^{(j)\prime}\delta^{(j)} + \beta^{(j)}y^{(-j)} + u^{(j)}\big), \tag{3.3}$$

where $y^{(-j)} \in \{0, 1\}$ denotes the other player's action, $x^{(j)}$ is player j's observable characteristics of player j, and $u^{(j)}$ is an unobservable payoff shifter. The payoff is summarized below and is assumed to belong to the players' common knowledge.

		Player 2	
		$y^{(2)} = 0$	$y^{(2)} = 1$
Player 1	$y^{(1)} = 0$	$0, 0$	$0, x^{(2)\prime}\delta^{(2)} + u^{(2)}$
	$y^{(1)} = 1$	$x^{(1)\prime}\delta^{(1)} + u^{(1)}, 0$	$x^{(1)\prime}\delta^{(1)} + \beta^{(1)} + u^{(1)},\ x^{(2)\prime}\delta^{(2)} + \beta^{(2)} + u^{(2)}$

The key parameter is the *strategic interaction effect* $\beta^{(j)}$ which captures the impact of the opponent's taking $y^{(-j)} = 1$ on player j's payoff. Suppose that $\beta^{(j)} \leq 0$ for both players. For example, if the outcome represents each firm's market entry, $\beta^{(j)}$ measures the effect of the other firm's entry on firm j's profit. Let $\theta = (\beta, \delta)$. Suppose that the players play a pure strategy Nash equilibrium (PSNE). Then, the set of PSNEs predicted by this model is summarized by the following map:

$$G(u|x;\theta) = \begin{cases} \{(0,0)\} & u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}, \\ \{(0,1)\} & u \in U_2, \\ \{(1,0)\} & u \in U_1, \\ \{(1,1)\} & u^{(1)} > -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}, \\ \{(1,0),(0,1)\} & -x^{(j)\prime}\delta^{(j)} < u^{(j)} < -x^{(j)\prime}\delta^{(j)} - \beta^{(j)}, \quad j = 1, 2. \end{cases} \tag{3.4}$$

where $U_1 = \{u^{(1)} > -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\} \cup \{-x^{(1)\prime}\delta^{(1)} < u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}\}$ and $U_2 = \{u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)}\} \cup \{-x^{(1)\prime}\delta^{(1)} < u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} > x^{(2)\prime}\delta^{(2)} - \beta^{(2)}\}$.

Figure C·6 shows the level sets of $u \mapsto G(u|x,\theta)$ for a given (x,θ). When $\beta^{(j)} < 0$ for both players, the model admits multiple equilibria $\{(0,1),(1,0)\}$ when each $u^{(j)}$ is between the two thresholds $x^{(j)\prime}\delta^{(j)}$ and $x^{(j)\prime}\delta^{(j)} - \beta^{(j)}$ (the colored region in Figure C·6). When $\beta^{(j)} = 0$ for either of the players, the model predicts a unique equilibrium for any value of $u = (u^{(1)}, u^{(2)})'$ (see left panel of Figure C·6, in which $\beta^{(j)} = 0, j = 1,2$).

Example 2 (Discrete Games of Strategic Complementarity)**.** Consider the payoff functions in (3.3) again but assume that $\beta^{(j)} \geq 0$. One can use this setting to analyze households' labor supply or retirement decisions, in which household members' labor force participation can be strategically complementary (Bresnahan and Reiss, 1991a).

For each (x,u), the predicted set of PSNE's is given by

$$G(u|x;\theta) = \begin{cases} \{(0,0)\} & u \in U_1, \\ \{(0,1)\} & u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} \geq -x^{(2)\prime}\delta^{(2)} \\ \{(1,0)\} & u^{(1)} \geq -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)} - \beta^{(2)} \\ \{(1,1)\} & u \in U_2, \\ \{(0,0),(1,1)\} & -x^{(j)\prime}\delta^{(j)} - \beta^{(j)} \leq u^{(j)} < -x^{(j)\prime}\delta^{(j)},\ j = 1,2, \end{cases} \tag{3.5}$$

where $U_1 = \{u^{(1)} < -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}\} \cup \{-x^{(1)\prime}\delta^{(1)} - \beta^{(1)} \leq u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}\}$, and $U_2 = \{u^{(1)} \geq -x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, u^{(2)} \geq -x^{(2)\prime}\delta^{(2)}\} \cup \{u^{(1)} \geq -x^{(1)\prime}\delta^{(1)}, -x^{(2)\prime}\delta^{(2)} - \beta^{(2)} \leq u^{(2)} < -x^{(2)\prime}\delta^{(2)}\}$.

When $\beta^{(j)} = 0$ for one of the players, the model makes a complete prediction for almost all u. In contrast, if $\beta^{(j)} > 0$ for both members, both $(0,0)$ and $(1,1)$ can arise as equilibrium outcomes for some value of u.

The following example is a parametric version of the triangular system of nonseparable equations (Chesher, 2003; Shaikh and Vytlacil, 2011). We consider a control function approach to this model.

Example 3 (Triangular model with an incomplete control function)**.** Consider a triangular model, in which a binary outcome y_i is determined by a binary treatment d_i, a vector w of exogenous covariates, and an unobserved variable ϵ_i; the treatment indicator d_i is determined by a vector of instrumental variables z_i and an unobserved

variable v_i:

$$y_i = 1\{\alpha d_i + w_i'\eta + \epsilon_i \geq 0\}, \tag{3.6}$$
$$d_i = 1\{z_i'\gamma + v_i \geq 0\}. \tag{3.7}$$

Suppose that (w_i, z_i) is independent of (ϵ_i, v_i). The unobserved characteristics ϵ_i and v_i may be dependent, making d_i potentially endogenous.

If one could recover the unobservable characteristic v_i from the observables (which would be possible with a continuous d_i), conditioning on v_i would make ϵ_i independent of d_i. This *control function approach* would allow us to recover key model parameters (Imbens and Newey, 2009; Wooldridge, 2015). In the current setting, we may not uniquely recover v_i due to the discreteness of d_i, which makes it difficult to apply the control function approach without further assumptions.[1] However, the model restricts v_i to the following set:

$$\begin{aligned} H(d_i, z_i; \gamma) &\equiv \left\{v \in \mathbb{R} : d_i = 1\{z_i'\gamma + v \geq 0\}\right\} \\ &= \begin{cases} [-z_i'\gamma, \infty) & \text{if } d_i = 1, \\ (-\infty, -z_i'\gamma) & \text{if } d_i = 0. \end{cases} \end{aligned} \tag{3.8}$$

Suppose that ϵ_i's conditional distribution given v_i belongs to a location family and the location parameter is βv_i. Then, one may write $\epsilon_i = \beta v_i + u_i$ for some u_i independent of d_i. Substituting this expression into (3.6) and noting that $v_i \in H(d_i, z_i; \gamma)$, the set of outcome values compatible with the model is

$$\begin{aligned} G(u_i|x_i; \theta) = \Big\{ & y_i \in \{0, 1\} : y_i = 1\{\alpha d_i + w_i'\eta + \beta v_i + u_i \geq 0\} \\ & \text{for some } v_i \in H(d_i, z_i; \gamma) \Big\}, \end{aligned} \tag{3.9}$$

where $x_i = (d_i, w_i', z_i')'$ and $\theta = (\beta, \delta')'$ with $\delta = (\alpha, \eta', \gamma')'$. One of the benefits of the control function approach that one can test the endogeneity of d_i (Wooldridge, 2015). As we show below, this is also the case even if the control function cannot be

[1] Wooldridge (2014) uses the generalized residual $r_i = d_i\lambda(z_i'\gamma) - (1 - d_i)\lambda(-z_i'\gamma)$ from the first stage MLE, where λ is the inverse Mills ratio. He makes additional high-level assumptions so that r_i serves as a sufficient statistic for capturing the endogeneity of d_i and proposes an estimator of the average structural function. Instead of taking this approach, we explore what can be learned from the set-valued control function.

uniquely recovered.[2]

The next example is a panel dynamic discrete choice model (Heckman, 1978; Hyslop, 1999) that features unobserved initial conditions.

Example 4 (Panel Dynamic Discrete Choice Models)**.** An individual makes binary decisions across multiple periods according to

$$y_{it} = 1\{x_{it}'\lambda + y_{it-1}\beta + \alpha_i + \epsilon_{it} \geq 0\},\ i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{3.10}$$

where y_{it} is a binary outcome for individual i in period t, x_{it} is a vector of observable covariates, α_i is an unobservable individual specific effect, and ϵ_{it} is an unobserved idiosyncratic error. If β is nonzero, the individual's choice in period t depends on her past choice, rendering the decision *state dependent*.

Suppose the researcher observes (y_{it}, x_{it}) for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Since y_{i0} is not observable, this leaves the value of $y_{i1}, \ldots, y_{iT}$ not fully determined and makes the model incomplete (Heckman, 1978, 1987; Honoré and Tamer, 2006).[3] For example, consider $T = 2$. Suppose for the moment $y_{i0} = 0$. For a given $(x_i, \alpha_i, \epsilon_{i1}, \epsilon_{i2})$, the observed outcome $y_i = (y_{i1}, y_{i2})$ must satisfy

$$y_{i1} = 1\{x_{i1}'\lambda + \alpha_i + \epsilon_{i1} \geq 0\}, \tag{3.11}$$
$$y_{i2} = 1\{x_{i2}'\lambda + y_{i1}\beta + \alpha_i + \epsilon_{i2} \geq 0\}. \tag{3.12}$$

Similarly, if $y_{i0} = 1$, the outcome must satisfy

$$y_{i1} = 1\{x_{i1}'\lambda + \beta + \alpha_i + \epsilon_{i1} \geq 0\}, \tag{3.13}$$
$$y_{i2} = 1\{x_{i2}'\lambda + y_{i1}\beta + \alpha_i + \epsilon_{i2} \geq 0\}. \tag{3.14}$$

Without further assumptions, the model permits both possibilities. Letting $u_i = (u_{i1}, u_{i2})'$ with $u_{it} = \alpha_i + \epsilon_{it}$, the model prediction can therefore be summarized by

[2] We take a control function approach that conditions on v_i, which only requires specification of the conditional distribution of ϵ_i given v_i. Alternatively, one could specify the joint distribution of (ϵ_i, v_i). This alternative but stronger assumption would imply a complete model; there is a unique value of (y_i, d_i) for a given (ϵ_i, v_i) and exogenous covariates due to the triangular structure (Lewbel, 2007).

[3] As an alternative, one could work with the likelihood function conditional on the initial observation. However, this approach can be problematic if one wants to be internally consistent across a different number of periods (Honoré and Tamer, 2006; Wooldridge, 2005).

the following correspondence

$$G(u_i|x_i;\theta) = \Big\{ y_i = (y_{i1}, y_{i2}) \in \{0,1\}^2 : y_i \text{ satisfies either (3.11)-(3.12) or (3.13)-(3.14)} \Big\}. \tag{3.15}$$

If $\beta \geq 0$, this map can be expressed as follows:[4]

$$G(u_i|x_i;\theta) = \begin{cases} \{(0,0)\} & u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} < -x_{i2}'\lambda, \\ \{(0,1)\} & u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} \geq -x_{i2}'\lambda, \\ \{(1,0)\} & u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} < -x_{i2}'\lambda - \beta, \\ \{(1,1)\} & u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda - \beta, \\ \{(0,0),(1,0)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ u_{i2} \leq -x_{i2}'\lambda - \beta, \\ \{(0,0),(1,1)\} & -x_{it}'\lambda - \beta \leq u_{it} < -x_{it}'\lambda,\ t \in \{1,2\}, \\ \{(0,1),(1,1)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda. \end{cases} \tag{3.16}$$

Similar to the previous examples, the model makes a complete prediction when $\beta = 0$ (see Figure C·11 in the Appendix).

3.3 Testing Hypotheses

Let $\beta \in \Theta_\beta \subset \mathbb{R}^{d_\beta}$ denote the subvector of θ whose value determines whether the model is complete or not. Let $\delta \in \Theta_\delta \subset \mathbb{R}^{d_\delta}$ collect the remaining components of θ. Given a sample of data $(Y_i, X_i), i = 1, \ldots, n$, we consider testing a hypothesis on β. Let the null and alternative hypotheses be

$$H_0 : \beta = \beta_0, \quad \text{v.s.} \quad H_1 : \beta \in B_1, \tag{3.17}$$

where $B_1 \subset \Theta_\beta$ is a set not containing β_0. For instance, in Example 1, the presence of strategic substitution effects can be tested by letting $\beta_0 = 0$ and $B_1 = \{\beta : \beta^{(j)} < 0, j = 1, 2\}$. Similarly, we may test the potential endogeneity of treatment assignments

[4] Appendix G.1.3 provides details and a graphical illustration of G.

(Example 3) and the presence of state dependence (Example 4) by setting $\beta_0 = 0$ and choosing suitable alternative hypotheses. In what follows, we let $\Theta_0 = \{\beta_0\} \times \Theta_\delta$ and $\Theta_1 = B_1 \times \Theta_\delta$ denote the sets of null and alternative parameter values respectively.

Let $\Delta_{Y|X}$ denote the set of conditional distributions (or probability kernels) of Y given $X = x$. For each $\theta = (\beta', \delta')'$, an incomplete model admits the following set of conditional distributions:

$$\mathcal{Q}_\theta = \Big\{ Q \in \Delta_{Y|X} : Q(A|x) = \int_U p(A|x,u) dF_\theta(u),\ \forall A \subseteq \mathcal{Y}, \\ \text{for some } p \in \Delta_{Y|X,u} \text{ such that } p(G(u|x;\theta)|x,u) = 1,\ a.s. \Big\}. \quad (3.18)$$

Here, the conditional distribution $p(\cdot|x,u)$ represents the unknown *selection mechanism* according to which an outcome gets selected from the set of predicted outcome values. Since the model is silent about its specification, we allow any law supported on $G(u|x;\theta)$. Consequently, the model can admit (infinitely) many likelihood functions for a given θ. Let μ be the counting measure on $\mathcal{Y}$. For each θ, define

$$\mathfrak{q}_\theta = \{q_{y|x} : q_{y|x} = dQ(\cdot|x)/d\mu, Q \in \mathcal{Q}_\theta\}. \quad (3.19)$$

This set collects all (conditional) densities that are compatible with a given θ. In the case of discrete games (Examples 1 and 2), this set contains all densities of equilibrium outcomes that are compatible with the game's description. Similarly, in the context of panel discrete choice (Example 4), this set collects all densities of individual choices consistent with arbitrary specifications of the initial condition. The multiplicity of the densities is due to the model incompleteness that admits any selection mechanism $p(\cdot|u,x)$. In this sense, we may think that elements in $\mathfrak{q}_\theta$ are indexed by the unknown selection mechanisms. Observe that $\mathfrak{q}_\theta$ reduces to a singleton set $\{q_\theta\}$ if the model is complete, i.e. $G(u|x;\theta) = \{g(u|x;\theta)\}$ for some function g, in which case $q_\theta = dQ_\theta/d\mu$

with $Q_\theta(A|x) = \int 1\{g(u|x;\theta) \in A\} dF_\theta(u)$.

While the multiplicity of likelihood functions may appear challenging, $\mathfrak{q}_\theta$ can be simplified, and this property enables us to conduct robust tests in a tractable manner. By Artstein's inequality (see e.g. Galichon and Henry, 2011; Molinari, 2020), $\mathfrak{q}_\theta$ can be written as the following set of densities satisfying a finite number of linear inequalities:

$$\mathfrak{q}_\theta = \Big\{ q_{y|x} : \sum_{y \in A} q_{y|x}(y|x) \geq \nu_\theta(A|x),\ A \subseteq \mathcal{Y} \Big\}, \tag{3.20}$$

where

$$\nu_\theta(\cdot|x) = F_\theta(G(u|x;\theta) \subseteq \cdot|x) \tag{3.21}$$

is the conditional *containment functional* (or *belief function*) associated with the random set $G(u|x;\theta)$. This functional gives the sharp lower bound for the conditional probability $Q(A|x)$ across all Q's that belong to $\mathcal{Q}_\theta$.[5] Theoretical properties of the containment functional and methods of numerical approximation are well studied in the literature (Ciliberto and Tamer, 2009; Galichon and Henry, 2011).[6] For us, the fact that $\mathfrak{q}_\theta$ is characterized by a system of linear inequalities is important. Together with an extended Neyman-Pearson lemma, this allows us to construct a computationally tractable score-based test. In the next subsection, we briefly review the existing results we rely on.

3.3.1 Preliminaries

Let $p_0(y|x)$ denote the true conditional distribution of the outcome given the covariates. Let us start with a problem of distinguishing a parameter value θ_0 from another value θ_1. In a parametrically specified complete model $\{p_\theta, \theta \in \Theta\}$, this amounts to

[5] The upper bound for $Q(A|x)$ is given by the *capacity functional* $\nu^*(A|X) = F_\theta(G(u|x;\theta) \cap A \neq \emptyset|x)$ (Molinari, 2020). It is sufficient to use either of the lower or upper bounds in (3.20) because the bounds are related to each other through the conjugate relationship $\nu(A|x) = 1 - \nu^*(A^c|x)$.

[6] We provide a brief review on these in Appendix B.

testing $p_0 = p_{\theta_0}$ against $p_0 = p_{\theta_1}$. It is well known that the most powerful test for such a problem is a likelihood-ratio test, which is the result of the Neyman-Pearson lemma. In incomplete models, corresponding null and alternative hypotheses would be $p_0 \in \mathfrak{q}_{\theta_0}$ and $p_0 \in \mathfrak{q}_{\theta_1}$ rendering both hypotheses composite. KZ19 show that it is possible to extend the Neyman-Pearson lemma to such settings, building on a general result established by Huber and Strassen (1973). Their key observation is that there is a *least favorable pair (LFP)* $(q_{\theta_0}, q_{\theta_1}) \in \mathfrak{q}_{\theta_0} \times \mathfrak{q}_{\theta_1}$ of densities. This pair is such that q_{θ_0} is consistent with θ_0 and is least favorable for controlling the size of a test among all densities belonging to $\mathfrak{q}_{\theta_0}$, whereas q_{θ_1} is consistent with θ_1 and is least favorable for maximizing a measure of power among the densities belonging to $\mathfrak{q}_{\theta_1}$.[7] Furthermore, they show that a likelihood-ratio test based on this pair constitutes a minmax test, which maximizes a robust measure of power among a class of level-α tests (see KZ19 Section 3).

There is a simple way to compute the LFP through a convex program. For each $x \in \mathcal{X}$, the LFP $(q_{\theta_0}, q_{\theta_1})$ is characterized as

$$(q_{\theta_0}, q_{\theta_1}) = \underset{(q_0, q_1)}{\arg\min} \sum_{y \in \mathcal{Y}} \ln\left(\frac{q_0(y|x) + q_1(y|x)}{q_0(y|x)}\right)(q_0(y|x) + q_1(y|x)) \tag{3.22}$$

$$s.t. \sum_{y \in A} q_0(y|x) \geq \nu_{\theta_0}(A|x), \ A \subseteq \mathcal{Y} \tag{3.23}$$

$$\sum_{y \in A} q_1(y|x) \geq \nu_{\theta_1}(A|x), \ A \subseteq \mathcal{Y}. \tag{3.24}$$

The constraints in (3.23) and (3.24) are the sharp identifying restrictions.[8] In view of (3.20), they are equivalent to saying that q_0 belongs to $\mathfrak{q}_{\theta_0}$ and q_1 belongs to $\mathfrak{q}_{\theta_1}$

[7] They consider the lower envelope of power over $\mathcal{Q}_\theta$.

[8] A common way to use them for identification analysis is to define the *sharp identified set* as $\Theta_I = \{\theta : P(A|x) \geq \nu_\theta(A|x), a.s.\}$. That is, given the conditional probability $P(\cdot|x)$ identified from data, one collects all values of θ satisfying the sharp identifying restrictions. For hypothesis testing, we instead fix θ and ask what would be a distribution among all distributions satisfying the sharp identifying restrictions that is least favorable for controlling the size or maximizing the power.

respectively. For us, these restrictions are useful for computing the LFP because they are linear in (q_0, q_1). The convex problem can be solved numerically in general. In some of the leading examples, it is also possible to compute it analytically.

To illustrate, let us consider Example 1. Suppose that the latent payoff shifters $(u^{(1)}, u^{(2)})$ follow a bivariate standard normal distribution. We may then compute $\nu_\theta(A|x)$ for each event. Let us take $A = \{(1,0)\}$ as an example. Using (3.4) and (3.21), we obtain

$$\begin{aligned}\nu_\theta(\{(1,0)\}|x) &= F_\theta(G(u|x;\theta) \subseteq \{(1,0)\}|x) \\ &= F_\theta(u \in U_1) = (1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})], \quad (3.25)\end{aligned}$$

where $\Phi(\cdot)$ denotes the standard normal CDF. This corresponds to the probability assigned to the green region in Figure C·6 (right panel) and is the sharp lower bound for the probability of $A = \{(1,0)\}$.

Now consider two parameter values $\theta_0 = (0_2', \delta')'$ and $\theta_1 = (\beta', \delta')'$, where $\beta = (\beta^{(1)}, \beta^{(2)})'$ with $\beta^{(j)} < 0$ for $j = 1, 2$. As we discuss in more detail below, the model is complete when $\beta = 0$. One can show that the restrictions in (3.23) reduce to the following equality restrictions:

$$q_0((0,0)|x) = (1 - \Phi(x^{(1)\prime}\delta^{(1)}))(1 - \Phi(x^{(2)\prime}\delta^{(2)})), \quad (3.26)$$

$$q_0((0,1)|x) = (1 - \Phi(x^{(1)\prime}\delta^{(1)}))\Phi(x^{(2)\prime}\delta^{(2)}), \quad (3.27)$$

$$q_0((1,0)|x) = \Phi(x^{(1)\prime}\delta^{(1)})(1 - \Phi(x^{(2)\prime}\delta^{(2)})), \quad (3.28)$$

$$q_0((1,1)|x) = \Phi(x^{(1)\prime}\delta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}). \quad (3.29)$$

These restrictions uniquely determine the least-favorable null density q_{θ_0}. Hence, we

may write

$$q_{\theta_0}(y|x) = [(1-\Phi_1)(1-\Phi_2)]^{1\{y=(0,0)\}}[(1-\Phi_1)\Phi_2]^{1\{y=(0,1)\}} \times [\Phi_1(1-\Phi_2)]^{1\{y=(1,0)\}}[\Phi_1\Phi_2]^{1\{y=(1,1)\}}, \quad (3.30)$$

where, to ease notation, we use Φ_1 and Φ_2 to denote $\Phi(x^{(1)\prime}\delta^{(1)})$ and $\Phi(x^{(2)\prime}\delta^{(2)})$.

When $\beta^{(j)} < 0, j = 1, 2$, there are multiple densities satisfying (3.24). The least favorable alternative density q_{θ_1} can be found by minimizing (3.22) with respect to q_1 subject to (3.24). The solution can be expressed analytically. For example, when player 1's strategic interaction effect on player 2 is relatively high, it is given by the following form:[9]

$$\begin{aligned} q_{\theta_1}(y|x) = {} & [(1-\Phi_1)(1-\Phi_2)]^{1\{y=(0,0)\}}[(1-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)}))\Phi_2]^{1\{y=(0,1)\}} \\ & \times [(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})(\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}))]^{1\{y=(1,0)\}} \\ & \times [\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]^{1\{y=(1,1)\}}. \quad (3.31) \end{aligned}$$

Comparing (3.30) and (3.31), one can see that q_{θ_1} tends to q_{θ_0} as β approaches its null value (i.e. 0). Hence, one may view $\theta \mapsto q_\theta$ a "parametric" model, and we will indeed take this view below. Here is a way to interpret this parametric model. For each θ_1, the density q_{θ_1} corresponds to the DGP that is least favorable in terms of detecting β's deviation from its null value among all densities compatible with θ_1. Behind q_{θ_1}, there is a selection mechanism that induced the least favorable DGP. For our purposes, however, we do not need to know the precise form of this selection mechanism. When we solve the convex program, we are "profiling out" the selection mechanism and directly obtaining the induced density q_{θ_1}. This is why q_{θ_1} no longer involves any selection mechanism.

[9] See Appendix G.1.1 for details.

By varying θ_1, we may trace out a family of such densities and form a parametric model. We therefore call the map $\theta \to q_\theta$ the *least favorable (LF) parametric model*. Equation (3.31) suggests that we may pretend as if data were generated by a parametric discrete choice model with the given density. This is indeed the case if one is interested in maximizing a measure of discrimination between θ_0 and θ_1 based on q_θ. Thanks to this property, most of our analysis below will resemble that of standard discrete choice models, which helps us keep our framework tractable.

3.3.2 Model Completeness Under the Null

The following assumption imposes a key structure on the model.

Assumption 1. *(i) For any null parameter value $\theta_0 = (\beta_0', \delta')'$ with $\delta \in \Theta_\delta$, the set of conditional densities of outcome is a singleton $\mathfrak{q}_{\theta_0} = \{q_{\theta_0}\}$; (ii) For any pair of parameters $\theta_0 = (\beta_0', \delta')'$ and $\theta_1 = (\beta', \delta')'$ with $\beta \in B_1$ and $\delta \in \Theta_\delta$, we have $\mathfrak{q}_{\theta_0} \cap \mathfrak{q}_{\theta_1} = \emptyset$.*

By Assumption 1 (i), we require the model is complete under the null hypothesis in the sense that $\mathfrak{q}_{\theta_0}$ contains a unique density when $\beta = \beta_0$. As discussed earlier, this holds whenever the model makes a complete prediction under the null hypothesis and is satisfied in the examples discussed in Section 3.2.1.

The model can be complete or incomplete under the alternative hypothesis. Assumption 1 (ii) requires that the sets $\mathfrak{q}_{\theta_0}$ and $\mathfrak{q}_{\theta_1}$ are disjoint. If this is the case, it is possible to detect θ_1's local deviation from θ_0 regardless of the unknown selection mechanism. In KZ19, such an alternative hypothesis is called *robustly testable*, and we focus on settings in which this assumption is satisfied.[10] Let us now revisit the examples for illustration.

Example 1 (Binary Response Game of Complete Information)**.** Consider testing the presence of strategic substitution effects by testing $H_0 : \beta^{(1)} = \beta^{(2)} = 0$ against

[10]KZ19 analyze a general case in which this assumption may fail to hold by extending the notion of local alternatives. We conjecture that we may extend our framework similarly. Since our leading examples satisfy Assumption 1 (ii) (see Appendix G.1), we leave this extension elsewhere.

$H_0 : \beta^{(1)} < 0, \beta^{(2)} < 0$. Under the null hypothesis, there is no strategic interaction between the players, which leads to the following complete prediction:

$$G(u|x;\theta_0) = \begin{cases} \{(0,0)\} & u^{(1)} < -x^{(1)\prime}\delta^{(1)}, u^{(2)} < -x^{(2)\prime}\delta^{(2)}, \\ \{(1,1)\} & u^{(1)} > -x^{(1)\prime}\delta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)}, \\ \{(1,0)\} & u^{(1)} > -x^{(1)\prime}\delta^{(1)}, u^{(2)} \leq -x^{(2)\prime}\delta^{(2)}, \\ \{(0,1)\} & u^{(1)} \leq -x^{(1)\prime}\delta^{(1)}, u^{(2)} > -x^{(2)\prime}\delta^{(2)}. \end{cases} \tag{3.32}$$

Hence, for any value of the observed and unobserved variables, $G(u|x;\theta_0)$ contains a unique equilibrium outcome. This corresponds to the left panel of Figure C·6. A similar analysis applies to Example 2.

Example 3 (Triangular Model with an Incomplete Control Function)**.** Consider testing the endogeneity of the treatment by testing the hypothesis that the coefficient β on the control function v is 0. When the null hypothesis is true, the model's prediction reduces to

$$y_i = 1\{\alpha d_i + w_i'\eta + u_i \geq 0\}. \tag{3.33}$$

Hence, for a given (x_i, u_i), the value of y_i is uniquely determined regardless of the value of the control function. Indeed, there is no need to control for v_i because u_i is independent of (d_i, w_i). Hence, this is a model of binary choice with exogenous covariates whose analysis is standard.

Example 4 (Panel Dynamic Discrete Choice Models)**.** Consider testing the presence of state dependence. This can be done by testing whether the coefficient β on the lagged dependent variable y_{it-1} is 0 or not. When $\beta = 0$ in (3.10), the model reduces to the static panel binary choice model

$$y_{it} = 1\{x_{it}'\lambda + \alpha_i + \epsilon_{it} \geq 0\},\ i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{3.34}$$

which makes (3.11)-(3.12) and (3.13)-(3.14) equivalent. Hence, under the null hypothesis, $G(u_i|x_i;\theta_0)$ contains the unique outcome satisfying (3.34).

3.3.3 Score–Based Tests

Score-based tests such as Rao's score (or Lagrange multiplier) test and Neyman's $C(\alpha)$ test are widely used. These tests require estimation of the restricted model only, which is attractive in our setting. The restricted model is complete and hence typically admits point estimation of nuisance parameters under reasonably weak conditions. We take advantage of this property to carry out a score-based test. Below, we briefly review the core ideas behind the classic score tests and discuss extensions to handle potential model incompleteness under the alternative. For expositional purposes, we assume that q_θ is differentiable with respect to θ for now and will weaken this assumption later.

Consider testing the null parameter value $\theta_0 = (\beta_0', \delta')'$ against a local alternative hypothesis $\theta_h = (\beta_0' + h', \delta')'$, where $h \in \mathbb{R}^{d_\beta}$. The most powerful test for this problem is the likelihood-ratio test, which compares q_{θ_h} to q_{θ_0} and rejects H_0 when the ratio of the two is high (KZ19). The test is also robust in the sense that, under Assumption 1 (ii), the log-likelihood ratio can detect any deviation from the null hypothesis with non-trivial power no matter what the selection mechanism is. The log-likelihood ratio can be locally approximated by $\sum_{i=1}^n h' s_\beta(Y_i|X_i;\beta_0,\delta)$, where $s_\beta(y|x;\beta,\delta) = \frac{\partial}{\partial\beta}\ln q_\theta(y|x)|_{\theta=(\beta,\delta)}$ is the score function. Let $\Sigma_{\beta_0} = \mathrm{Var}(\sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\delta))$. For i.i.d. data, $\Sigma_{\beta_0} = nI_{\beta_0}$ where $I_{\beta_0} = E[s_\beta(Y_i|X_i;\beta_0,\delta)s_\beta(Y_i|X_i;\beta_0,\delta)']$. For a fixed h, the normalized quantity

$$\frac{(\sum_{i=1}^n h' s_\beta(Y_i|X_i;\beta_0,\delta))^2}{h'\Sigma_\beta h}, \tag{3.35}$$

serves as a measure of discrimination between β_0 and $\beta_0 + h$. This quantity is a robust measure of local discrimination between β_0 and $\beta_0 + h$ because it is based on an approximation to the log-likelihood ratio $\ln(q_{\theta_h}/q_{\theta_0})$, which serves as a robust (and optimal) test statistic for detecting a local deviation from the null hypothesis.

If one seeks for a direction h that maximizes (3.35), it is given by

$$h^* = I_{\beta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\beta(Y_i|X_i;\beta_0,\delta),$$

which motivates Rao's score statistic:[11]

$$\begin{aligned} T_n &= \sup_{h\in\mathbb{R}^{d_\beta}} \frac{\sum_{i=1}^n h' s_\beta(Y_i|X_i;\beta_0,\delta))^2}{nh' I_{\beta_0} h} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\beta(Y_i|X_i;\beta_0,\delta)' I_{\beta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\beta(Y_i|X_i;\beta_0,\delta). \end{aligned} \tag{3.36}$$

This statistic depends on the unknown nuisance parameter δ. Suppose that the nuisance parameter δ can be estimated by a point estimator $\hat{\delta}_n$. Evaluating the sample mean of the score at $\delta = \hat{\delta}_n$ and imposing the null hypothesis yields

$$g_n(\beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\beta(Y_i|X_i;\beta_0,\hat{\delta}_n). \tag{3.37}$$

A feasible version of (3.36) is

$$\hat{T}_n = g_n(\beta_0)' \hat{V}_n^{-1} g_n(\beta_0), \tag{3.38}$$

where $\hat{V}_n = n^{-1} \sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\hat{\delta}_n) s_\beta(Y_i|X_i;\beta_0,\hat{\delta}_n)'$ is a consistent estimator of the asymptotic variance $V_0 \equiv I_{\beta_0}$. The sampling distribution of the score generally depends on $\hat{\delta}_n$. However, if one uses the restricted MLE (discussed in the next section) to estimate δ, which we recommend, $\hat{T}_n$ converges in distribution to a χ^2-distribution with d_β degrees of freedom under the null hypothesis.

It is also possible to use point estimators other than the restricted MLE, in which case we recommend using an orthogonalized version of the score in the spirit of Neyman's $C(\alpha)$ test (see Remark 2 below).

[11] See Bera and Bilias (2001) for a more detailed argument for complete models. The same argument can be applied to incomplete models by replacing the standard likelihood function with the LF density q_θ.

The analysis so far presumed that q_θ was differentiable and $h \in \mathbb{R}^{d_\beta}$ was unrestricted. These assumptions may be restrictive in our context. For example, in discrete games of complete information, the least favorable parametric model $h \mapsto q_{\theta_h}$ and its score take different functional forms depending on whether the alternative hypothesis admits strategic substitution (i.e. $h < 0$ as in Example 1) or strategic complementarity (i.e. $h > 0$ as in Example 2). It is then natural to analyze these two cases separately. Hence, we weaken the differentiability requirement to accommodate these features and also allow the alternative hypothesis to be restricted (e.g. one-sided). Let $\mathbb{C}(0, \epsilon)$ denote an open cube centered at the origin with edges of length 2ϵ. A set $\Gamma \subseteq \mathbb{R}^d$ is said to be locally equal to set $\Upsilon \subseteq \mathbb{R}^d$ if $\Gamma \cap \mathbb{C}(0, \epsilon) = \Upsilon \cap \mathbb{C}(0, \epsilon)$ for some $\epsilon > 0$ (Andrews, 1999).

Assumption 2 (L^2-directional differentiability)**.** *(i) $B_1 - \beta_0$ is locally equal to a convex cone $\mathcal{V}_1$; (ii) For any $\zeta \in \mathcal{V}_1 \times \mathbb{R}^{d_\delta}$, there exists a square integrable function $s_\theta = (s'_\beta, s'_\delta)' : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^d$ such that*

$$\left\| q_{\theta_0+\tau\zeta}^{1/2} - q_{\theta_0}^{1/2}(1 + \frac{1}{2}\tau\zeta' s_\theta(\cdot|\cdot; \beta_0, \delta)) \right\|_{L^2_\mu} = o(\tau), \tag{3.39}$$

as $\tau \downarrow 0$.

Assumption 2 (i) requires the set of deviations from β_0 can be locally approximated by a convex cone. In Example 1, consider testing $H_0 : \beta = (0, 0)'$ against $H_1 : \beta^{(1)} < 0, \beta^{(2)} < 0$. Then, $B_1 - \beta_0$ is locally equal to

$$\mathcal{V}_1 = \{h = (h^{(1)}, h^{(2)}) : h^{(1)} < 0, h^{(2)} < 0\}. \tag{3.40}$$

Assumption 2 (ii) uses the notion of differentiability in quadratic mean (see e.g van der Vaart, 2000), but it only requires that a unique score, in the sense of the L^2-derivative of the square-root density, exists for the set $\mathcal{V}_1$ of local deviations from the null hypothesis. This weaker assumption is appropriate for incomplete models, and s_θ

can be derived from the least favorable parametric model similar to the standard parametric models.[12]

To accommodate the "one-sided" nature of the alternative hypothesis, we define a test statistic for $H_0 : \beta = \beta_0$ v.s. $H_1 : \beta \in B_1$ by

$$\hat{S}_n = g_n(\beta_0)'\hat{V}_n^{-1}g_n(\beta_0) - \inf_{h \in \mathcal{V}_1}(g_n(\beta_0) - h)'\hat{V}_n^{-1}(g_n(\beta_0) - h). \tag{3.41}$$

This test statistic is a slight modification of (3.38) and follows the construction in Silvapulle and Silvapulle (1995). It requires the same functions of data as T_n, but it is designed to direct power against the local alternatives in $\mathcal{V}_1$. Note that, if the alternative hypothesis is locally unrestricted, i.e. $\mathcal{V}_1 = \mathbb{R}^{d_\beta} \setminus 0$, the test statistic reduces to T_n.

The asymptotic distribution of $\hat{S}_n$ is no longer a χ^2 distribution. However, its critical value is easy to compute using simulations. Let

$$c_\alpha = \inf\{x \in \mathbb{R} : Pr(S \leq x) \geq 1 - \alpha\}, \tag{3.42}$$

where

$$S \equiv Z'V_0^{-1}Z - \inf_{h \in \mathcal{V}_1}(Z - h)'V_0^{-1}(Z - h), \ Z \sim N(0, V_0), \tag{3.43}$$

which can be simulated by drawing Z repeatedly from a zero mean multivariate normal distribution with estimated variance $\hat{V}_n$.

Remark 2. If $\hat{\delta}_n$ is not the restricted MLE, $\hat{g}_n(\beta_0)$'s limiting distribution may depend on that of $\hat{\delta}_n$ in general. Neyman's $C(\alpha)$ statistic addresses this issue by making the statistic insensitive to the estimation error associated with $\hat{\delta}_n$. This is achieved by projecting s_β to s_δ and replacing g_n with the "orthogonalized" (or "residualized")

[12] Appendix G.1 derives s_θ for some of the examples.

score:

$$g_n(\beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\beta(Y_i|X_i; \beta_0, \hat{\delta}_n) - I_{\beta,\delta} I_\delta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\delta(Y_i|X_i; \beta_0, \hat{\delta}_n), \tag{3.44}$$

where $I_{\beta,\delta}$ and I_δ are submatrices of

$$I_\theta = \begin{bmatrix} I_\beta & I_{\beta,\delta} \\ I_{\delta,\beta} & I_\delta \end{bmatrix} = E_{q_\theta}\left[s_\theta(Y_i|X_i) s_\theta(Y_i|X_i)'\right], \tag{3.45}$$

which can be estimated by their sample analogs. The orthogonalized score $g_n(\beta_0)$ constructed this way is robust to the estimation error of δ, and its asymptotic distribution coincides with a version of the statistic which replaces $\hat{\delta}_n$ with the true value δ_0. The asymptotic variance of $g_n(\beta_0)$ is $V_0 = I_\beta - I_{\beta,\delta} I_\delta^{-1} I_{\delta,\beta}$. The test statistic in (3.41) can be constructed in the same way using g_n in (3.44) and a consistent estimator $\hat{V}_n$ of V_0. The way to calculate the critical value remains the same.[13]

3.3.4 Estimation of Nuisance Parameters

Our tests require an estimator $\hat{\delta}_n$ of the nuisance parameter. A natural estimator of δ is the restricted maximum likelihood estimator (MLE) $\hat{\delta}_n$, which is a maximizer of

$$\mathbb{M}_n(\delta) \equiv \frac{1}{n} \sum_{i=1}^{n} \ln q_{\beta_0,\delta}(Y_i|X_i), \tag{3.46}$$

where $q_{\beta_0,\delta}$ is the conditional density of Y_i when $\beta = \beta_0$. Under our assumptions, this density also coincides with the least favorable parametric model evaluated at $\beta = \beta_0$. We therefore call the map $\delta \mapsto \ln q_{\beta_0,\delta}$ the *restricted log-likelihood function*. The complete model (under H_0) is often a standard discrete choice problem. As such, existing package software can be used to estimate δ.

Example 1 (Binary Response Game (continued))**.** Under $H_0 : \beta^{(1)} = \beta^{(2)} = 0$, the model has a unique likelihood function as discussed in Section 3.3.1. The (restricted)

[13] When $\hat{\delta}_n$ is the restricted maximum likelihood estimator, the second term on the right hand side of (3.44) becomes asymptotically negligible making $\hat{S}_n$ asymptotically equivalent to the version without the Neyman orthogonalization (Kocherlakota and Kocherlakota, 1991).

maximum likelihood estimator $\hat{\delta}_n$ maximizes

$$\mathbb{M}_n(\delta) = \sum_{i=1}^{n} \Big(1\{Y_i = (0,0)\} \ln[(1-\Phi_{1,i})(1-\Phi_{2,i})] + 1\{Y_i = (0,1)\} \ln[(1-\Phi_{1,i})\Phi_{2,i}] \\ + 1\{Y_i = (1,0)\} \ln[\Phi_{1,i}(1-\Phi_{2,i})] + 1\{Y_i = (1,1)\} \ln[\Phi_{1,i}\Phi_{2,i}] \Big),$$

where $\Phi_{j,i} = \Phi(X_i^{(j)\prime}\delta^{(j)}), j = 1, 2$.

Alternatively, one can estimate δ by only using features of the model that are uniquely predicted. This strategy used earlier in the literature would maximize a likelihood function based on the empirical frequency of no entry ($Y_i = (0,0)$), monopoly ($Y_i = (0,1)$ or $(1,0)$), and duopoly ($Y_i = (1,1)$) (Bresnahan and Reiss, 1991b; Berry, 1992).[14] If one uses this estimator, we recommend using the orthogonalized score to construct the test statistic (see Remark 2).

Example 3 (Triangular Model with an Incomplete Control Function)**.** When $\beta = 0$, there is no correlation between the errors in the outcome and selection equations. The outcome equation reduces to a binary choice model with exogenous covariates. If we assume $u_i \sim N(0,1)$, the conditional probability of $y_i = 1$ is

$$q_{\theta_0}(1|d_i, w_i, z_i) = \Phi(\alpha d_i + w_i'\eta), \tag{3.47}$$

which can be used to estimate $\delta = (\alpha, \eta)$ by the restricted MLE. This can be done by any software that may estimate probit models. Similarly, the selection equation is another binary choice model. One can estimate the coefficients on the instruments Z using, for example, a probit model.

Example 4 (Dynamic Discrete Choice)**.** A random effects probit model assumes α_i is independent of x_i and follows $N(0, \gamma^2)$, and $\epsilon_{i1}, \ldots, \epsilon_{iT}$ are independent standard normal random variables. This yields the following conditional density function:

$$q_{\beta_0,\delta}(y_i|x_i) = \int \prod_{t=1}^{T} \Phi\big[(2y_{it} - 1)(x_{it}'\lambda + \gamma a)\big]\phi(a)da. \tag{3.48}$$

One can construct a likelihood function based on (3.48) to obtain a point estimator of $\delta = (\lambda, \gamma)$. To approximate the integral, one can, for example, use the simulated maximum likelihood method (Train, 2009).

[14] If this alternative estimator is used, T_n is not asymptotically equivalent to Rao's score statistic in general.

3.3.5 Asymptotic Properties

We collect results on the asymptotic properties of our test. Throughout, we assume that $u^n = (u_1, \ldots, u_n)$ is an independent and identically distributed (i.i.d.) sample drawn from F_θ, and $X^n = (X_1, \ldots, X_n)$ is also an i.i.d. sample drawn from a distribution q_X^n. The joint distribution of the outcome sequence $Y^n = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$ conditional on $x^n = (x_1, \ldots, x_n)$ is not uniquely determined due to the potential incompleteness of the model, and it belongs to the following set:

$$\mathcal{Q}_\theta^n = \Big\{ Q : Q(A|x^n) = \int_{U^n} p(A|u^n, x^n) dF_\theta^n(u),\ \forall A \subseteq \mathcal{Y}^n,$$
$$\text{for some } p \in \Delta_{Y^n|X^n,u^n} \text{ such that } p(G^n(u^n|x^n;\theta)|u^n, x^n) = 1, a.s. \Big\}, \quad (3.49)$$

where F_θ^n denotes the joint law of u^n, and $G^n(u^n|x^n;\theta) = \prod_{i=1}^n G(u_i|x_i;\theta)$ is the Cartesian product of the set-valued predictions.[15] We then let $\mathcal{P}_\theta^n$ collect joint laws of (Y^n, X^n); each element P^n of $\mathcal{P}_\theta^n$ is such that the conditional law of Y^n given X^n belongs to $\mathcal{Q}_\theta^n$, and the law of X^n is q_X^n.

We start with conditions that ensure the $\sqrt{n}$-consistency of $\hat{\delta}_n$. They are mainly regularity conditions on the restricted log-likelihood function. Fixing β to its null value, one can view $q_{\beta_0,\delta}$ as the conditional density of y in a regular parametric model, in which δ is the only unknown parameter. As such, the conditions below parallel the ones in the literature.

Below, let $\delta_0 \in \Theta_\delta$ denote the true value of the nuisance component vector. For each $\delta \in \Theta_\delta$, let $\mathbb{M}(\delta) \equiv E[\ln q_{\beta_0,\delta}]$, where expectation is taken with respect to the conditional density q_{β_0,δ_0} and the distribution of X. Let $\mathbb{M}_n(\delta) \equiv n^{-1} \sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i)$ be the sample counterpart of $\mathbb{M}$ and let $\mathbb{G}_n(\delta) \equiv \sqrt{n}(\mathbb{M}_n(\delta) - \mathbb{M}(\delta))$ be an empirical

[15] Assuming u^n and X^n are i.i.d. does not imply Y^n is i.i.d. The set $\mathcal{Q}_\theta^n$ in general contains dependent and heterogeneous laws because the behavior of the selection mechanism across experiments is unrestricted (see Epstein et al., 2016). This does not create an issue for the size properties of our test because $\mathcal{Q}_\theta^n$ reduces to a single i.i.d. law under the null hypothesis.

process indexed by δ.

Assumption 3. *(i-a) There is a continuous function $M : \Theta_\delta \to \mathbb{R}_+$ such that*

$$\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}} |\ln q_{\beta_0,\delta}(y|x)| \leq M(\delta).$$

(i-b) $\delta \mapsto \ln q_{\beta_0,\delta}(y|x)$ is Lipschitz continuous uniformly in (y,x). That is,

$$\sup_{(y,x)\in\mathcal{Y}\times\mathcal{X}} \left|\ln q_{\beta_0,\delta}(y|x) - \ln q_{\beta_0,\delta'}(y|x)\right| \lesssim \|\delta - \delta'\| \ \forall \delta, \delta' \in \Theta_\delta. \tag{3.50}$$

(i-c) $\delta \neq \delta_0 \Rightarrow q_{\beta_0,\delta}(y|x) \neq q_{\beta_0,\delta_0}(y|x)$ with positive probability.
(ii) Θ_δ is a nonempty compact set.
(iii) $\hat{\delta}_n$ is such that

$$\mathbb{M}_n(\hat{\delta}_n) \geq \inf_{\delta\in\Theta_\delta} \mathbb{M}_n(\delta) + r_n,$$

where, for any h and $\epsilon > 0$, $\sup_{P^n\in\mathcal{P}^n_{\theta_0+h/\sqrt{n}}} P^n(|r_n| > \epsilon) \to 0$.

We also assume F_θ belongs a smooth parametric family in the following sense.

Assumption 4. *For each $\theta \in \Theta$, F_θ is absolutely continuous with respect to a σ-finite measure ζ on U. The Radon-Nikodym density $f_\theta = dF_\theta/d\zeta$ satisfies*

$$\|f_\theta - f_{\theta'}\|_{L^1_\zeta} \leq C\|\theta - \theta'\|, \ \forall \theta, \theta' \in \Theta, \tag{3.51}$$

for some $C > 0$.

Finally, the following condition requires that the population objective function is locally well behaved so that its value is informative about δ_0, and the supremum of an empirical log-likelihood process can be controlled over a neighborhood of δ_0 (see van der Vaart and Wellner, 1996, Sec. 3.2.2).

Assumption 5. *For every δ in a neighborhood of δ_0,*

$$\mathbb{M}(\delta) - \mathbb{M}(\delta_0) \lesssim -\|\delta - \delta_0\|^2. \tag{3.52}$$

Furthermore,

$$\sup_{P^n\in\mathcal{P}^n_{\theta_0+h/\sqrt{n}}} E^*_{P^n} \sup_{\delta\in B_\zeta(\delta_0)} |\mathbb{G}_n(\delta) - \mathbb{G}_n(\delta_0)| \lesssim \zeta, \tag{3.53}$$

where $B_\zeta(\delta_0) = \{\delta : \|\delta - \delta_0\| < \zeta\}$.

Under thse assumptions, the restricted MLE $\hat{\delta}_n$ is $\sqrt{n}$-consistent.

Proposition 3.3.1. *Suppose Assumptions 1-5 hold. Then,*

$$\sqrt{n}\|\hat{\delta}_n - \delta_0\| = O_{P^n}(1), \tag{3.54}$$

uniformly in $P^n \in \mathcal{P}^n_{\theta_0 + h/\sqrt{n}}$.

Below, let $P^n_0 \in \mathcal{P}^n_{\theta_0}$ be the unique joint law of (Y^n, X^n) under the null hypothesis. Let $s_{\theta,j}$ be the j-th component of s_θ. Define

$$\Xi = \big\{\xi : \mathcal{Y} \times \mathcal{X} \to \mathbb{R} | \xi(y,x) = s_{\theta,j}(y|x;\beta_0,\delta) s_{\theta,k}(y|x;\beta_0,\delta),\ 1 \le j,k \le d, \delta \in \Theta_\delta\big\}. \tag{3.55}$$

We assume the elements of Ξ obey a uniform law of large numbers, i.e. Ξ is a Glivenko-Cantelli class.

Assumption 6.

$$\sup_{\xi \in \Xi} \Big| \frac{1}{n} \sum_{i=1}^n \xi(Y_i, X_i) - E_{P_0}[\xi(Y_i, X_i)] \Big| = o_{P_0^n}(1). \tag{3.56}$$

Suppose $\hat{S}_n$ as defined in (3.41) or it is constructed from the orthogonalized score using an $\sqrt{n}$-consistent estimator that is not necessarily the restricted MLE. The following theorem shows that the test controls its size.

Theorem 3.3.1. *Suppose Assumptions 1-6 hold. Let* c_α *be defined by* (3.42)*. Then, for any* $\alpha \in (0,1)$*,*

$$\lim_{n \to \infty} P_0^n(\hat{S}_n > c_\alpha) = \alpha. \tag{3.57}$$

3.3.6 Inference on Parameters

In some applications, the ultimate goal may be to make inference on the underlying parameter, for example, to construct confidence intervals for components of θ. While

this is not our focus, we discuss a possible way to achieve this and leave its formal analysis to future work.

Consider constructing confidence intervals for a component or linear combination $p'\delta_0$ of δ_0.[16] According to Proposition 3.3.1, $\hat{\delta}_n$ is a $\sqrt{n}$-consistent estimator of δ_0 as long as the true value of β is in a neighborhood of β_0 whose radius is of order $n^{-1/2}$. It would be natural to use such an estimator to construct a confidence interval for δ_0 if the complete model is selected. A well-known challenge for such a post-model selection inference procedure is that a naive asymptotic approximation that disregards the model selection step may not be valid uniformly over a large class of DGPs (Leeb and Pötscher, 2005). Given this, we consider a hybrid method, which is summarized in the following box:

Algorithm: A Hybrid Testing Procedure

---

(1) Compute S_n and $c_n = \min(\kappa_n, 1) \times c_\alpha$, where κ_n is a sequence that tends to 0 slowly, e.g. $\kappa_n = (\ln n)^{-1/2}$.

(2) If $S_n > c_n$, reject $H_0 : \beta = \beta_0$:

- construct a *robust confidence interval* by Kaido et al. (2019) for $p'\delta$.

If $S_n \leq c_n$, do not reject $H_0 : \beta = \beta_0$:

- construct the *Wald confidence interval*:

$$[p'\hat{\delta}_n - z_{\alpha/2} \times SE(p'\hat{\delta}_n),\ p'\hat{\delta}_n + z_{\alpha/2} \times SE(p'\hat{\delta}_n)],$$

where $SE(\cdot)$ is the (estimated) standard error of its argument, and z_α is the $(1-\alpha)$ quantile of the standard normal distribution.

The heuristic behind this procedure is as follows. First, we compare S_n to a critical value c_n that tends to 0 slowly. For DGPs with β well separated from its null value's neighborhood, we cannot ensure the asymptotic validity of the Wald confidence interval. In such a case, the pre-test that rejects H_0 with a high probability

[16] Since β's value is pinned down by the null hypothesis, it is natural to consider inference on the parameters that are estimated under both hypotheses.

should prescribe a robust confidence interval, which controls the asymptotic coverage probability regardless of β's value. Since the critical value c_n is shrunk toward zero, we use the Wald confidence interval only if β is in a small local neighborhood of β_0 against which the score test has little power. The shrinkage factor κ_n, therefore, introduces a conservative distortion, which is expected to make the resulting confidence interval's coverage probability above its nominal over a wide range of β values and other features of the DGP.

3.4 Empirical Illustrations

We illustrate the score test through two empirical applications.

3.4.1 Testing Strategic Interaction Effects

The first application revisits the analysis of the airline industry by Kline and Tamer (2016). We test the presence of strategic interaction effects between two types of firms: low-cost carriers (LCC) and other airlines (OA). Below, we briefly summarize the setup and refer to Kline and Tamer (2016) for details. A market is defined as trips between airports regardless of intermediate stops. The two types of firms, LCC and OA, decide whether or not to serve each market. The binary variable $y_i^{(\ell)}$ takes value 1 if airline $\ell \in \{\text{LCC, OA}\}$ serves market i. Airline ℓ's payoff in market i equals

$$y_i^{(\ell)}(\delta_\ell^{cons} + \delta_\ell^{size} X_{i,size} + \delta_\ell^{pres} X_{i,pres}^{(\ell)} + \beta_\ell y_i^{(-\ell)} + u_i^{(\ell)}),$$

where β_ℓ captures the impact of the competitor's entry decision, $y_i^{(-\ell)}$. Each firm's payoff is determined by the airline-specific intercepts and observable covariates. The covariates include the *market size* $X_{i,size}$ and the *market presence* $X_{i,pres}^{(\ell)}$. The market size $X_{i,size}$ is defined as the population at the endpoints of each trip. The latter variable $X_{i,pres}^{(\ell)}$ measures the presence of firm ℓ in market i (see Kline and Tamer,

2016, p.356 for its definition). This is an airline-and-market-specific variable and shows up only in firm ℓ's payoff. The data come from the second quarter of the 2010 Airline Origin and Destination Survey (DB1B) and contain 7882 markets.[17]

Our hypothesis of interest is whether the LCCs and OAs compete in a strategic way, which can be formulated as a one-sided test. The null hypothesis is $H_0 : \beta_{LCC} = \beta_{OA} = 0$, and the alternative hypothesis is $H_1 : \beta_\ell < 0, \ell \in \{\text{LCC,OA}\}$. We also examine whether discretizing covariates has any impacts on the conclusion. In applications of inference methods for partially identified models, it is common to discretize the covariates. Kline and Tamer (2016) transform each of the covariates into binary variables based on whether they realized above or below their respective median. In one of our specifications, we follow their approach to discretize the covariates. In another specification, we treat treat $X_{i,size}$ and $X^{(\ell)}_{i,pres}$ as continuous variables normalized to the unit interval.[18] Finally, the vector of coefficients $\delta = (\delta^{cons}_{LCC}, \delta^{size}_{LCC}, \delta^{pres}_{LCC}, \delta^{cons}_{OA}, \delta^{size}_{OA}, \delta^{pres}_{OA})$ is the nuisance parameter in this model. We estimate δ by the restricted MLE under the null hypothesis.

The results of the score test are reported in Table (D.9). When the covariates are discretized, the test rejects the null hypothesis at the 5% level, which is consistent with the finding of Kline and Tamer (2016) whose credible sets for the strategic interaction effects $\beta_\ell, \ell \in \{\text{LCC,OA}\}$ do not contain the origin. For comparison, we also consider a specification without discretization since the robust score test can accommodate continuous regressors as well. Under this specification, the result changes drastically; we do not reject the presence of strategic substitution even at the 10% level. This suggests that a model without any strategic interaction effects can potentially explain the observed market entry decisions once we take into account the rich variation of the continuous covariates.

[17] The data are available on Brendan Kline's website.

[18] The scale of the two variables without discretization differ significantly, and thus we re-scale each variable to be between 0 and 1.

3.4.2 Testing the Endogeneity of Catholic School Attendance

The second application concerns the causal effect of Catholic school attendance on academic achievements studied by Altonji et al. (2005). The question of whether Catholic schools provide better education than public ones is important for education policies, but the analysis is complicated by the concern that selection into Catholic schools is nonrandom. Using the framework in Example 3, we formally test the endogeneity of Catholic school attendance by testing if the coefficient on the control function is zero.

The data source is a subset of the National Educational Longitudinal Survey of 1988 (NELS:88). We use a version of the data available from Wooldridge (2019). We refer to Altonji et al. (2005) for a detailed discussion of the data. The dependent variable y_i is a binary variable indicating whether the student graduated from high school by year 1994. The binary treatment d_i indicates whether the student attended a Catholic high school. The vector of exogenous control variables w_i includes each parent's years of education and log family income. The instrument variable is a dummy variable indicating whether a parent was reported to be Catholic. The sample size n is 5970 after we remove missing observations on y_i.

Table (D.11) reports the point estimates of nuisance parameters δ under the null hypothesis $H_0 : \beta = 0$. The value of the test statistic is 154.848. The 5% and 1% critical values are 2.755 and 5.201, respectively. We, therefore, reject the null hypothesis at both levels. Our test provides strong evidence supporting the students' selection into Catholic schools based on their unobservable characteristics. This result is in line with the concern expressed in Altonji et al. (2005).

3.5 Monte Carlo Experiments

3.5.1 Size and Power of the Score Test

We examine the size and power properties of the score test through simulations. The DGP is based on Example 1 and is motivated by the empirical illustration in the previous section. There are player-specific covariates $x_i = (x_i^{(1)}, x_i^{(2)})'$, each of which is generated as an independent Rademacher random variable taking values on $\{-1, 1\}$. We then generate $u_i = (u_i^{(1)}, u_i^{(2)})$ from the bivariate standard normal distribution. For each u_i and x_i, we determine the predicted set of outcomes $G(u_i|x_i;\theta)$ based on the payoff functions with $\delta_0 = (\delta_0^{(1)}, \delta_0^{(2)}) = (2, 1.5)'$. We test

$$H_0 : \beta^{(1)} = \beta^{(2)} = 0, \quad v.s. \quad H_1 : \beta^{(1)} < 0, \beta^{(2)} < 0. \tag{3.58}$$

As discussed earlier, the model is complete under H_0. We estimate δ_0 using the restricted MLE. The sample size is set to 2500, 5000, or 7500. This choice is motivated by the sample size used in the empirical application.

The size of the score test is reported in Table (D.12). The size of the test is controlled properly with $n = 7500$, but there are small size distortions when the sample size is smaller possibly due to estimation errors associated with components of the estimated information matrix.

Under alternative hypotheses, multiple equilibria may be predicted. If this is the case, we select an outcome according to one of the following selection mechanisms. The first design uses a selection mechanism, which selects $(1, 0)$ out of $G(u_i|x_i;\theta) = \{(1, 0), (0, 1)\}$ if an i.i.d. Bernoulli random variable ν_i takes 1. In the second design, we generate data from the least favorable distribution, which draws an independent outcome sequence from the least favorable distribution $Q_{\theta_1} \in \mathcal{Q}_{\theta_1}$.

The power of the score test is calculated against local alternatives with $\beta_1^{(j)} = -h/\sqrt{n}, h > 0$ for $j = 1, 2$. For this exercise, we introduce a grid of values for h and

generate data described as above. We then compare the rejection frequency of our test to that of the moment-based testing procedure by Bugni et al. (2017). Their test checks if a hypothesized value $(\beta^{(1)}, \beta^{(2)})' = (0,0)'$ is compatible with a set of moment restrictions. Their statistic and bootstrap critical value are calculated using a sample analog of the following moment inequality and equality restrictions

$$
\begin{aligned}
&P(Y=(1,0)|X=x) \geq (1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta_1 + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)})],\\
&P(Y=(1,0)|X=x) \leq (1-\Phi(x^{(2)\prime}\delta_2 + \beta^{(2)}))\Phi_1,\\
&P(Y=(0,0)|X=x) = (1-\Phi_1)(1-\Phi_2),\\
&P(Y=(1,1)|X=x) = \Phi(x^{(1)\prime}\delta_1 + \beta^{(1)})\Phi(x^{(2)\prime}\delta_2 + \beta^{(2)}),
\end{aligned}
$$

which are the sharp identifying restrictions that characterize $\mathfrak{q}_\theta$ in (3.20).[19]

Figures (C·8) and (C·9) show the rejection frequencies of the score and the moment-based tests by Bugni et al. (2017). The results are similar across the two designs. In each design, the score test outperforms the moment-based test in terms of power by a large margin. This may be due to the fact that the score is based on an approximation to the optimal likelihood ratio test.

3.6 Conclusion

This paper proposes a novel score-based test of model completeness. Our test is attractive in settings where the model involves nuisance parameters, which is common in applications. The score test only requires estimation of nuisance parameters within the restricted model. We utilize a point estimator of the nuisance parameters whereby we avoid evaluations of the test statistic over many parameter values.

The results of Monte Carlo experiments suggest the score test has an advantage

[19] Since the example resembles the specification used in their Monte Carlo experiments, we added minimal changes to their replication code posted on the repository of *Quantitative Economics* to implement their procedure.

in terms of power over an existing method. An avenue for future research includes a unified theory for the uniform validity of inference for post-model selection procedures that are based on our score test.

Appendix A

Proofs of Chapter One

A.1 Auxiliary Results

A.1.1 Proof of Lemma A.1.1

Lemma A.1.1 (Uniform Convergence of Sample Criterion Function using Simulated Data)**.**

$$\max_{1\leq i\leq n}\sup_{(\beta,\gamma)}\left|\widehat{G}^h_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\right|\xrightarrow{p}0,$$

where

$$\widehat{G}^h_{(i)}(\beta,\gamma)=\frac{1}{T}\sum_{t=1}^{T}\ln f(y^h_{it}(\theta,\widehat{\alpha}_i)\mid x_{it};\beta,\gamma);$$

$$G_{(i)}(\beta,\gamma)=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\ln f(y_{it}(\theta,\alpha_{i0})\mid x_{it};\beta,\gamma).$$

Proof. The proof consists of two main steps. The first step deals with $\widehat{\alpha}_i$'s in data simulation and shows that $\widehat{G}^h_{(i)}$ is uniformly close to a criterion that uses α_{i0} to simulate the data, i.e.,

$$\widetilde{G}_{(i)}(\beta,\gamma)=\frac{1}{T}\sum_{t=1}^{T}\int_U\ln f(y_{it}(\theta,\alpha_{i0})\mid x_{it};\beta,\gamma)dF(u).$$

The second step is a uniform law of large number results showing that $\widetilde{G}_{(i)}(\beta,\gamma)$ uniformly converges to $G_{(i)}(\beta,\gamma)$.

Step 1: Given θ and a scalar τ, note that

$$\frac{1}{T}\sum_{t=1}^{T}\ln f(y_{it}^{h}(\theta,\tau)\mid x_{it};\beta,\gamma) := \frac{1}{T}\sum_{t=1}^{T} y_{it}^{h}(\theta,\tau)\ln\Phi(x_{it}'\beta+\gamma) + (1-y_{it}^{h}(\theta,\tau))\ln(1-\Phi(x_{it}'\beta+\gamma))$$

consists of two components: (1) an indicator function of scalar τ and (2) a smooth, bounded and monotone function of (β,γ). The indicator function $y_{it}^{h}(\theta,\widehat{\alpha}_i)$ belongs to type I class of Andrews (1994), which satisfies Pollard's entropy condition. The second component belongs to a class of functions satisfying bracketing entropy condition (van der Vaart and Wellner, 1996, Section 2.7.2).

Because $\frac{1}{T}\sum_{t=1}^{T}\ln f(y_{it}^{h}(\theta,\tau)\mid x_{it};\beta,\gamma)$ is an additive and multiplicative combination of the two classes of components, its function class also satisfies the entropy condition (Andrews, 1994), which is the primitive condition for stochastic equicontinuity. More specifically, define the following empirical process:

$$\nu_T(\tau)=\frac{1}{T}\sum_{t=1}^{T}\Big[\ln f(y_{it}^{h}(\theta,\tau)\mid x_{it};\beta,\gamma)-\int_{U}\ln f(y_{it}^{h}(\theta,\tau)\mid x_{it};\beta,\gamma)dF_u\Big],$$

where the integration is over the known distribution of simulation draws. By one of the equivalent definitions of stochastic equicontinuity (i.e., Andrews, 1994, p.2252), the following condition holds: for every sequence of constants $\{\delta_T\}$ that converges to zero,

$$\sup_{(\beta,\gamma)\in\mathcal{B}\times\Gamma_\gamma,|\tau_1-\tau_2|\le\delta_T}\sqrt{T}|\nu_T(\tau_1)-\nu_T(\tau_2)|\xrightarrow{p}0. \tag{A.1.1}$$

A first–order Taylor expansion on $\int_U\ln f(y_{it}^{h}(\theta,\alpha_{i0})\mid x_{it};\beta,\gamma)$ with respect to α_{i0} around $\widehat{\alpha}_i$ yields

$$\int_{U}\ln f(y_{it}^{h}(\theta,\alpha_{i0})\mid x_{it};\beta,\gamma)dF_u=\int_{U}\ln f(y_{it}^{h}(\theta,\widehat{\alpha}_i)\mid x_{it};\beta,\gamma)dF_u + \frac{\partial\int_U\ln f(y_{it}^{h}(\theta,\overline{\alpha}_i)\mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}(\widehat{\alpha}_i-\alpha_{i0}).$$

Combined with condition (A.1.1),

$$\begin{aligned}
&\sqrt{T}\Big|\frac{1}{T}\sum_{t=1}^{T}\Big[\ln f(y_{it}^h(\theta,\widehat{\alpha}_i)\mid x_{it};\beta,\gamma)-\int_U \ln f(y_{it}^h(\theta,\alpha_{i0})\mid x_{it};\beta,\gamma)dF_u\Big]\Big| \\
&\quad=\sqrt{T}\Big|\nu_T(\widehat{\alpha}_i)-\frac{\partial\int_U \ln f(y_{it}^h(\theta,\overline{\alpha}_i)\mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}(\widehat{\alpha}_i-\alpha_{i0})]\Big| \\
&\quad\leq\sqrt{T}|\nu_T(\widehat{\alpha}_i)|+\sqrt{T}\Big|\frac{1}{T}\sum_{t=1}^{T}\frac{\partial\int_U \ln f(y_{it}^h(\theta,\overline{\alpha}_i)\mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}(\widehat{\alpha}_i-\alpha_{i0})\Big| \\
&\quad=\sqrt{T}|\nu_T(\alpha_{i0})+\nu_T(\widehat{\alpha}_i)-\nu_T(\alpha_{i0})| \\
&\quad+\sqrt{T}\Big|\frac{1}{T}\sum_{t=1}^{T}\frac{\partial\int_U \ln f(y_{it}^h(\theta,\overline{\alpha}_i)\mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}(\widehat{\alpha}_i-\alpha_{i0})\Big| \\
&\quad\leq\sqrt{T}|\nu_T(\alpha_{i0})|+\sqrt{T}|\nu_T(\widehat{\alpha}_i)-\nu_T(\alpha_{i0})| \\
&\quad+\sqrt{T}\Big|\frac{1}{T}\sum_{t=1}^{T}\frac{\partial\int_U \ln f(y_{it}^h(\theta,\overline{\alpha}_i)\mid x_{it};\beta,\gamma)dF_u}{\partial\alpha_i}\Big|\cdot|\widehat{\alpha}_i-\alpha_{i0}|.
\end{aligned}$$

where the third and last lines are due to triangular inequality. Because $\nu_T(\alpha_{i0})$ is a normalized sum of mean zero random variables, $\nu_T(\alpha_{i0}) \xrightarrow{p} 0$ by LLN. The second term is the stochastic equicontinuity condition in Eq. (A.1.1). Because the derivative is bounded by Assumption 4 and $\max_{1\leq i\leq n}|\widehat{\alpha}_i-\alpha_{i0}|=o_p(1)$ (Hahn and Kuersteiner, 2011, Theorem 4), the third term is thus $o_p(1)$. Therefore

$$\sup_{(\beta,\gamma)\in\mathcal{B}\times\Gamma_\gamma}\Big|\frac{1}{T}\sum_{t=1}^{T}[\ln f(y_{it}^h(\theta,\widehat{\alpha}_i)\mid x_{it};\beta,\gamma)-\int_U \ln f(y_{it}^h(\theta,\alpha_{i0})\mid x_{it};\beta,\gamma)dF_u]\Big|\xrightarrow{p}0.$$

Step 2: The second part of the proof shows that

$$\max_{1\leq i\leq n}\sup_{(\beta,\gamma)}\Big|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\Big|\xrightarrow{p}0.$$

Following the the proof structure of Lemma 4 in Hahn and Kuersteiner (2011), note that

$$P\Big[\max_{1\leq i\leq n}\sup_{(\beta,\gamma)}\Big|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\Big|\geq\eta\Big]\leq\sum_{i=1}^{n}P\Big[\sup_{(\beta,\gamma)}\Big|\widetilde{G}_{(i)}(\beta,\gamma)-G_{(i)}(\beta,\gamma)\Big|\geq\eta\Big].$$

Since the parameter space is compact, it suffices to show that

$$\sup_{\Gamma_j} \left| \widetilde{G}_{(i)}(\beta, \gamma) - G_{(i)}(\beta, \gamma) \right| \to 0,$$

where Γ_j is a subset of $\mathcal{B} \times \Gamma_\gamma$ such that $\|\beta - \beta'\| \leq \varepsilon$ and $|\gamma - \gamma'| \leq \varepsilon$ for (β, γ) and $(\beta', \gamma') \in \Gamma_j$. By Assumption 4 on $G_{(i)}$,

$$\begin{aligned}
&\left| G_{(i)}(\beta, \gamma) - G^h_{(i)}(\beta', \gamma') \right| \leq \mathbb{E}M(z_{it}) |(\beta, \gamma) - (\beta', \gamma')| < \varepsilon \mathbb{E}M(z_{it}), \\
&\left| \widetilde{G}_{(i)}(\beta, \gamma) - \widetilde{G}_{(i)}(\beta', \gamma') \right| \leq \frac{1}{T} \sum_{t=1}^{T} M(z_{it}) |(\beta, \gamma) - (\beta', \gamma')| < \frac{\varepsilon}{T} \sum_{t=1}^{T} M(z_{it}).
\end{aligned}$$

By the triangular inequality,

$$\begin{aligned}
&\left| \widetilde{G}_{(i)}(\beta, \gamma) - G_{(i)}(\beta, \gamma) \right| - \left| \widetilde{G}_{(i)}(\beta', \gamma') - G_{(i)}(\beta', \gamma') \right| \\
&\qquad \leq \left| \Big(\widetilde{G}_{(i)}(\beta, \gamma) - \widetilde{G}_{(i)}(\beta', \gamma') \Big) - \Big(G_{(i)}(\beta, \gamma) - G_{(i)}(\beta', \gamma') \Big) \right| \\
&\qquad \leq \left| \widetilde{G}_{(i)}(\beta, \gamma) - \widetilde{G}_{(i)}(\beta', \gamma') \right| + \left| G_{(i)}(\beta, \gamma) - G_{(i)}(\beta', \gamma') \right| \\
&\qquad < \varepsilon \mathbb{E}M(z_{it}) + \frac{\varepsilon}{T} \sum_{t=1}^{T} M(z_{it}) \\
&\qquad = \frac{\varepsilon}{T} \Big(\sum_{t=1}^{T} M(z_{it}) - \mathbb{E}M(z_{it}) \Big) + \frac{\varepsilon}{T} \mathbb{E}M(z_{it}) + \varepsilon \mathbb{E}M(z_{it}) \\
&\qquad < \frac{\varepsilon}{T} \Big| \sum_{t=1}^{T} M(z_{it}) - \mathbb{E}M(z_{it}) \Big| + 2\varepsilon \mathbb{E}M(z_{it}).
\end{aligned}$$

Therefore by a rearrangement of the terms,

$$\begin{aligned}
\left| \widetilde{G}_{(i)}(\beta, \gamma) - G_{(i)}(\beta, \gamma) \right| \leq &\left| \widetilde{G}_{(i)}(\beta', \gamma') - G_{(i)}(\beta', \gamma') \right| \\
&+ \frac{\varepsilon}{T} \Big| \sum_{t=1}^{T} M(x_{it}) - \mathbb{E}M(x_{it}) \Big| + 2\varepsilon \mathbb{E}M(x_{it}).
\end{aligned}$$

Let ε be such that $2\varepsilon \max_i \mathbb{E}M(z_{it}) < \frac{\eta}{3}$, then

$$\begin{aligned}
&P\Big[\sup_{\Gamma_j}\Big|\widetilde{G}_{(i)}(\beta,\gamma) - G_{(i)}(\beta,\gamma)\Big| > \eta\Big] \\
\leq& P\Big[\Big|\widetilde{G}_{(i)}(\beta',\gamma') - G_{(i)}(\beta',\gamma')\Big| > \frac{\eta}{3}\Big] + P\Big[\frac{1}{T}\Big|\sum_{t=1}^{T} M(x_{it}) - \mathbb{E}M(x_{it})\Big| > \frac{\eta}{3\varepsilon}\Big] \\
&+ P\Big[2\varepsilon\mathbb{E}M(x_{it}) > \frac{\eta}{3}\Big] \\
=& o(T^{-2}),
\end{aligned}$$

where the last line follows as the first two terms on the right–hand side are $o(T^{-2})$ by Lemma 1 in Hahn and Kuersteiner (2011) and the last term is of probability zero by construction. Since $n = O(T)$,

$$\begin{aligned}
&P\Big[\max_{1\leq i\leq n}\sup_{(\beta,\gamma)}\Big|\widetilde{G}_{(i)}(\beta,\gamma) - G_{(i)}(\beta,\gamma)\Big| \geq \eta\Big] \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{m(\varepsilon)} P\Big[\sup_{\Gamma_j}\Big|\widetilde{G}_{(i)}(\beta,\gamma) - G_{(i)}(\beta,\gamma)\Big| \geq \eta\Big] \\
&= o(T^{-1})
\end{aligned}$$

□

A.1.2 Proof of Lemma A.1.2

Lemma A.1.2 (Pointwise Consistency of Auxiliary Estimator in the Simulation World)**.** $\forall\theta \in \Theta$,

$$\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}}) \xrightarrow{p} \beta(\theta, \boldsymbol{\alpha_0}) = \theta.$$

Proof. The previous lemma shows that using $\widehat{\boldsymbol{\alpha}}$ for data simulation well approximates data simulated using $\boldsymbol{\alpha_0}$, therefore the randomness in the log likelihood function only comes from observed data. The proof structure of this lemma follows from that for Theorem 3 in Hahn and Kuersteiner (2011), with minor modification of notations. Fix $\eta > 0$ and set

$$\varepsilon = \inf_i \Big[G_{(i)}(\theta, \alpha_{i0}) - \sup_{\{(\beta,\gamma):\|(\beta,\gamma)-(\theta,\alpha_{i0})\|>\eta\}} G_{(i)}(\beta,\gamma)\Big] > 0$$

With probability $1 - o(T^{-1})$,

$$\begin{aligned}
\max_{\|\beta-\theta\|>\eta,\gamma_1,\ldots,\gamma_n} \frac{1}{n}\sum_{i=1}^{n} \widehat{G}^h_{(i)}(\beta,\gamma_i) &\leq \max_{\|(\beta,\gamma_i)-(\theta,\alpha_{i0})\|>\eta} \frac{1}{n}\sum_{i=1}^{n} \widehat{G}^h_{(i)}(\beta,\gamma_i) \\
&\leq \max_{\|(\beta,\gamma_i)-(\theta,\alpha_{i0})\|>\eta} \frac{1}{n}\sum_{i=1}^{n} G_{(i)}(\beta,\gamma_i) + \frac{1}{3}\varepsilon \\
&< \frac{1}{n}\sum_{i=1}^{n} G_{(i)}(\theta,\alpha_{i0}) - \frac{2}{3}\varepsilon \\
&< \frac{1}{n}\sum_{i=1}^{n} \widehat{G}^h_{(i)}(\theta,\alpha_{i0}) - \frac{1}{3}\varepsilon,
\end{aligned}$$

where the second and last inequalities are due to Lemma A.1.1. By definition

$$\max_{\beta,\gamma_1,\ldots,\gamma_n} \frac{1}{n}\sum_{i=1}^{n} \widehat{G}^h_{(i)}(\beta,\gamma_i) \geq \frac{1}{n}\sum_{i=1}^{n} G^h_{(i)}(\theta,\alpha_{i0}).$$

Hence

$$P\Big[\|\widehat{\beta}^h(\theta,\widehat{\boldsymbol{\alpha}}) - \beta(\theta,\boldsymbol{\alpha_0})\| \geq \eta\Big] = o(T^{-1}).$$

□

A.2 Proofs of Main Results

A.2.1 Proof of Proposition 1.4.1

Proof. The structure of the proof follows Theorem 1 in Newey (1991), which requires four main pieces. The parameter space Θ is compact by assumption. The limiting function $\beta(\theta,\boldsymbol{\alpha_0})$ is continuous since it is an identity function. Lemma A.1.2 establishes the pointwise convergence result using simulated data: $\forall\theta \in \Theta$, $\widehat{\beta}^h(\theta,\widehat{\boldsymbol{\alpha}}) \xrightarrow{p} \beta(\theta,\boldsymbol{\alpha_0})$. Therefore, it suffices to prove that $\widehat{\beta}^h(\theta,\widehat{\boldsymbol{\alpha}})$ is stochastic equicontinuous. This section uses $\widehat{\beta}^h(\theta)$ to ease the notation.
By Markov inequality, $\forall\eta > 0$,

$$Pr\Big(\sup_{\theta\in\Theta}\|\widehat{\beta}^h(\theta) - \beta(\theta,\boldsymbol{\alpha_0})\| > \eta\Big) \leq \frac{1}{\eta}\mathbb{E}\Big(\sup_{\theta\in\Theta}\|\widehat{\beta}^h(\theta) - \beta(\theta,\boldsymbol{\alpha_0})\|\Big).$$

Combined with the compactness assumption, it suffices to show that

$$\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\leq\delta} \|\widehat{\beta}^h(\theta_1) - \widehat{\beta}^h(\theta_2)\| \Big) \leq C\delta, \tag{A.2.1}$$

where δ denotes a positive scalar that is arbitrarily small and C is a constant. The rest of the proof consists of three parts. Firstly, a representation of $\widehat{\beta}^h(\theta_1) - \widehat{\beta}^h(\theta_2)$ in terms of profiled likelihood is established. Then, the question is transformed to bounding terms related to components of the profiled log likelihood. Lastly, the different pieces are glued together to give an expression of C.

Step 1: Let $\widehat{Q}(\widehat{\beta}^h(\theta);\theta)$ denote the profiled log likelihood function using simulated data h,

$$\begin{aligned}\widehat{Q}(\beta;\theta) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} & y_{it}^h(\theta,\widehat{\alpha}_i)\ln\Big(\Phi(x_{it}'\beta + \widehat{\gamma}_i(\beta)\Big) \\ & + (1-y_{it}^h(\theta,\widehat{\alpha}_i))\ln\Big(1-\Phi(x_{it}'\beta+\widehat{\gamma}_i(\beta))\Big).\end{aligned} \tag{A.2.2}$$

Then by definition, $\widehat{\beta}^h(\theta_1)$ and $\widehat{\beta}^h(\theta_2)$ satisfy the first–order conditions,

$$\frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_1);\theta_1)}{\partial\beta} = 0, \quad \frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_2)}{\partial\beta} = 0.$$

A first–order Taylor expansion yields

$$\frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_1);\theta_1)}{\partial\beta} = 0 = \frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_1)}{\partial\beta} + \frac{\partial^2\widehat{Q}(\widetilde{\beta};\theta_1)}{\partial\beta\partial\beta'}(\widehat{\beta}^h(\theta_1)-\widehat{\beta}^h(\theta_2)),$$

where $\widetilde{\beta}$ is between $\widehat{\beta}^h(\theta_1)$ and $\widehat{\beta}^h(\theta_2)$. Therefore,

$$\frac{\partial^2\widehat{Q}(\widetilde{\beta};\theta_1)}{\partial\beta\partial\beta'}(\widehat{\beta}^h(\theta_1)-\widehat{\beta}^h(\theta_2)) = \frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_2)}{\partial\beta} - \frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_1)}{\partial\beta}.$$

Let λ_s denote the smallest eigenvalue of the Hessian of the profiled likelihood, then a quadratic inequality leads to

$$\lambda_s\|\widehat{\beta}^h(\theta_1)-\widehat{\beta}^h(\theta_2)\| \leq \Big|\frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_2)}{\partial\beta} - \frac{\partial\widehat{Q}(\widehat{\beta}^h(\theta_2);\theta_1)}{\partial\beta}\Big|,$$

where $\frac{\partial\widehat{\gamma}_i(\widehat{\beta}(\theta_2))}{\partial\beta} = 0$ by the envelope theorem. For binary response panel probit

models, some algebra leads to the following expression of the right–hand–side term in the absolute sign,

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\Big(y_{it}^{h}(\theta_1,\widehat{\alpha}_i)-y_{it}^{h}(\theta_2,\widehat{\alpha}_i)\Big) \times\Big(\frac{\phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))]}\Big), \tag{A.2.3}$$

where $y_{it}^{h}(\theta)=\mathbf{1}\{x_{it}'\theta+\widehat{\alpha}_i\geq u_{it}^{h}\}$ and u_{it}^{h} is from the standard normal distribution. Therefore, to establish Condition (A.2.1), it suffices to focus on Eq. (A.2.3).

Step 2: By the Cauchy–Schwarz inequality,

$$\begin{aligned}&\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\leq\delta}\Big|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\Big(y_{it}^{h}(\theta_1,\widehat{\alpha}_i)-y_{it}^{h}(\theta_2,\widehat{\alpha}_i)\Big)\\ &\qquad\times\Big(\frac{\phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))]}\Big)\Big|\Big)\\ &\qquad\leq\sqrt{\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\leq\delta}\Big|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}y_{it}^{h}(\theta_1,\widehat{\alpha}_i)-y_{it}^{h}(\theta_2,\widehat{\alpha}_i)\Big|^2\Big)}\times\\ &\qquad\sqrt{\mathbb{E}\Big(\Big|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))]}\Big|^2\Big)}.\end{aligned}$$

For each i and t, the following two L^2–smoothness conditions hold:

$$\sqrt{\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\leq\delta}|y_{it}^{h}(\theta_1,\widehat{\alpha}_i)-y_{it}^{h}(\theta_2,\widehat{\alpha}_i)|^2\Big)}\leq\sqrt{\frac{\mathbb{E}\|x_{it}\|_2}{\sqrt{2\pi}}}\sqrt{\delta}, \tag{A.2.4}$$

$$\sqrt{\mathbb{E}\Big(\Big|\frac{\phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^{h}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^{h}(\theta_2))]}\Big|^2\Big)}\leq K_2, \tag{A.2.5}$$

where $\|x\|_2$ denotes the L_2-norm $|x'x|^{1/2}$. This corresponds to type IV class in Andrews (1994).

Proving condition (A.2.4): Denote $\Delta\theta:=\theta_2-\theta_1$ and note that

$$\sup_{\|\theta_1-\theta_2\|\leq\delta}|y_{it}^{h}(\theta_1)-y_{it}^{h}(\theta_2)|=\sup_{\|\Delta\theta\|\leq\delta}|\mathbf{1}\{x_{it}'\theta_1+\widehat{\alpha}_i\geq u_{it}^{h}\}-\mathbf{1}\{x_{it}'(\theta_1+\Delta\theta)+\widehat{\alpha}_i\geq u_{it}^{h}\}|.$$

The direction that obtains the supremum is given by

$$\Delta\theta = \pm\frac{\delta}{\|x_{it}\|_2}x_{it}.$$

Therefore

$$\mathbb{E}\Big[\sup_{\|\theta_1-\theta_2\|\leq\delta}|y_{it}^h(\theta_1)-y_{it}^h(\theta_2)|\Big] \leq \mathbb{E}\Big(\mathbf{1}\{x_{it}'\theta_1+\widehat{\alpha}_i \geq u_{it}^h\}-\mathbf{1}\{x_{it}'\theta_1-\|x_{it}\|_2\delta+\widehat{\alpha}_i \geq u_{it}^h\}\Big). \tag{A.2.6}$$

Because δ is a scalar, a proof strategy à la Chen et al. (2003) is employed to bound the right–hand–side term in Equation (A.2.6). More specifically, note that

$$\mathbf{1}\{x_{it}'\theta_1+\widehat{\alpha}_i \geq u_{it}^h\}-\mathbf{1}\{x_{it}'\theta_1-\|x_{it}\|_2\delta+\widehat{\alpha}_i \geq u_{it}^h\}$$

takes value either 1 or 0, and the expectation is the probability that the following event occurs:

$$x_{it}'\theta_1+\widehat{\alpha}_i \geq u_{it}^h \geq x_{it}'\theta_1-\|x_{it}\|_2\delta+\widehat{\alpha}_i.$$

Applying law of iterated expectation on the right–hand–side term and first–order Taylor expansion around δ,

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}\Big(\mathbf{1}\{x_{it}'\theta_1+\widehat{\alpha}_i \geq u_{it}^h\}-\mathbf{1}\{x_{it}'\theta_1-\|x_{it}\|_2\delta+\widehat{\alpha}_i \geq u_{it}^h\} \mid x_{it}',\widehat{\alpha}_i\Big)\Big]\\
=&\mathbb{E}\Big[\Phi(x_{it}'\theta_1+\widehat{\alpha}_i)-\Phi(x_{it}'\theta_1-\|x_{it}\|_2\delta+\widehat{\alpha}_i)\Big]\\
=&\mathbb{E}\Big[\phi(x_{it}'\theta+\widehat{\alpha}_i)\|x_{it}\|_2\Big]\delta
\end{aligned}$$

Therefore,

$$\mathbb{E}\Big[\sup_{\|\theta_1-\theta_2\|\leq\delta}|y_{it}^h(\theta_1)-y_{it}^h(\theta_2)|\Big] \leq \mathbb{E}\Big[\phi(x_{it}'\theta+\widehat{\alpha}_i)\|x_{it}\|_2\Big]\delta \leq \frac{\mathbb{E}\|x_{it}\|_2}{\sqrt{2\pi}}\delta,$$

where the last inequality uses the fact that $\phi(\cdot)\leq\frac{1}{\sqrt{2\pi}}$.

Proving condition (A.2.5): Note that

$$\sqrt{\mathbb{E}\Big(\Big|\frac{\phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}^h(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}^h(\theta_2))]}\Big|^2\Big)}$$

is no greater than

$$\sqrt{\mathbb{E}\Big(\sup_{(\beta,\gamma)\in\mathcal{B}\times\Gamma_\gamma}\Big|\frac{\phi(x_{it}'\beta+\gamma)x_{it}}{\Phi(x_{it}'\beta+\gamma)[1-\Phi(x_{it}'\beta+\gamma)]}\Big|^2\Big)},$$

which is bounded based on Lipschitz condition.

Step 3: Because the supremum of sum is no greater than sum of the supremum,

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}\Big|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}y_{it}^h(\theta_1,\widehat{\alpha}_i)-y_{it}^h(\theta_2,\widehat{\alpha}_i)\Big|^2\Big)\\
&\le\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}|y_{it}^h(\theta_1,\widehat{\alpha}_i)-y_{it}^h(\theta_2,\widehat{\alpha}_i)|^2\Big)\\
&\le\frac{\delta}{\sqrt{2\pi}}\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\|x_{it}\|_2,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\Big(\Big|\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\phi(x_{it}'\widehat{\beta}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}(\theta_2))x_{it}}{\Phi(x_{it}'\widehat{\beta}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}(\theta_2))[1-\Phi(x_{it}'\widehat{\beta}(\theta_2)+\widehat{\gamma}_i(\widehat{\beta}(\theta_2))]}\Big|^2\Big)\\
&\le\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}K_{it}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\Big(\sup_{\|\theta_1-\theta_2\|\le\delta}\|\widehat{\beta}(\theta_1)-\widehat{\beta}(\theta_2)\|\Big)\le\frac{\sqrt{\delta}}{(2\pi)^{1/4}}\sqrt{\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\|x_{it}\|_2K_{it}}.$$

This verifies condition (A.2.1) and hence establishes the stochastic equicontinuity condition.

Step 4: By Theorem 1 in Newey (1991), $\widehat{\beta}^h(\theta,\widehat{\boldsymbol{\alpha}})$ converges to $\beta(\theta,\boldsymbol{\alpha_0})$ uniformly over $\theta\in\Theta$. □

A.2.2 Proof of Theorem 1.4.1

Proof. Following the argument as in Appendix 1 of Gouriéroux et al. (1993), consistency of $\widetilde{\theta}^H$ requires the following three conditions to hold:

1. the function $\beta(\theta, \boldsymbol{\alpha_0})$ is invertible;
2. $\widehat{\theta}$ converges to $\beta(\theta_0, \boldsymbol{\alpha_0})$ in $\theta_0 \in \Theta$ pointwise;
3. $\widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$ converges to $\beta(\theta, \boldsymbol{\alpha_0})$ uniformly over $\theta \in \Theta$.

The first condition is satisfied because function is an identity. The second condition only involves fixed effect estimator using observed data, and is a standard result in large–T panel literature (e.g, Hahn and Kuersteiner, 2011, Theorem 3). The third condition is verified by Proposition 1.4.1.
Recall that $\widetilde{\theta}^H$ is the solution to the optimization problem:

$$\widetilde{\theta}^H = \arg\min_{\theta \in \Theta} [\widehat{\theta} - \widehat{\beta}_H(\theta, \widehat{\boldsymbol{\alpha}})]' [\widehat{\theta} - \widehat{\beta}_H(\theta, \widehat{\boldsymbol{\alpha}})],$$

where $\widehat{\beta}_H(\theta, \widehat{\boldsymbol{\alpha}}) := \frac{1}{H} \sum_{h=1}^{H} \widehat{\beta}^h(\theta, \widehat{\boldsymbol{\alpha}})$. Therefore. the limit of the optimization problem becomes

$$\min_{\theta \in \Theta} [\theta_0 - \theta]' [\theta_0 - \theta],$$

which has a unique solution θ_0. Therefore,

$$\widetilde{\theta}^H \xrightarrow{p} \theta_0.$$

□

A.2.3 Proof of Theorem 1.4.2

Proof. By Assumption 6 and consistency of $\widetilde{\theta}^H$,

$$\widehat{\theta} = \widehat{\beta}_H(\widetilde{\theta}^H, \widehat{\boldsymbol{\alpha}}) = \widehat{\beta}_H(\theta_0, \widehat{\boldsymbol{\alpha}}) + \mathbb{E}(\widehat{\beta}_H(\widetilde{\theta}^H, \widehat{\boldsymbol{\alpha}}) - \widehat{\beta}_H(\theta_0, \widehat{\boldsymbol{\alpha}})) + o_p\Big(\frac{1}{\sqrt{nT}}\Big).$$

By the mean–value theorem,

$$\mathbb{E}(\widehat{\beta}_H(\widetilde{\theta}^H, \widehat{\boldsymbol{\alpha}}) - \widehat{\beta}_H(\theta_0, \widehat{\boldsymbol{\alpha}})) = \frac{\partial \mathbb{E}\widehat{\beta}_H(\overline{\theta}, \widehat{\boldsymbol{\alpha}})}{\partial \theta} (\widetilde{\theta}^H - \theta_0),$$

where $\overline{\theta}$ is between θ_0 and $\widetilde{\theta}^H$. Therefore,

$$\begin{aligned}\sqrt{nT}(\widetilde{\theta}^H - \theta_0) &= -\Big[\frac{\partial \mathbb{E}\widehat{\beta}_H(\overline{\theta}, \widehat{\boldsymbol{\alpha}})}{\partial \theta}\Big]^{-1}\sqrt{nT}\Big(\widehat{\beta}_H(\theta_0, \widehat{\boldsymbol{\alpha}}) - \widehat{\theta}\Big) \\ &= \sqrt{nT}\Big(\widehat{\theta} - \widehat{\beta}_H(\theta_0, \widehat{\boldsymbol{\alpha}})\Big) + o_p(1),\end{aligned}$$

where the last equality uses the property that $\beta(\theta, \boldsymbol{\alpha_0}) = \theta$. Therefore, it suffices to focus on $\sqrt{nT}(\widehat{\theta} - \widehat{\beta}_H(\theta_0, \widehat{\boldsymbol{\alpha}}))$. Hahn and Kuersteiner (2011) derive the representation of $\widehat{\theta} - \theta_0$ as follows:

$$\widehat{\theta} - \theta_0 = \frac{A(\theta_0, \boldsymbol{\alpha_0})}{\sqrt{nT}} + \frac{B(\theta_0, \boldsymbol{\alpha_0})}{T} + o_p\Big(\frac{1}{T}\Big),$$

where $A(\theta_0, \boldsymbol{\alpha_0})$ and $B(\theta_0, \boldsymbol{\alpha_0})$ are complicated functions of the high–order derivatives of the log likelihood. Because the same regression is run on simulated data,

$$\widehat{\beta}_H(\theta_0) - \beta(\theta_0, \boldsymbol{\alpha_0}) = \frac{A^h(\theta_0, \widehat{\boldsymbol{\alpha}})}{\sqrt{nT}} + \frac{B^h(\theta_0, \widehat{\boldsymbol{\alpha}})}{T} + o_p\Big(\frac{1}{T}\Big),$$

where $\widehat{\boldsymbol{\alpha}} := (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)$. This implies

$$\widehat{\beta}_H(\theta_0, \widehat{\boldsymbol{\alpha}}) = \beta(\theta_0, \boldsymbol{\alpha_0}) + \frac{1}{H}\sum_{h=1}^{H}\frac{A^h(\theta_0, \widehat{\boldsymbol{\alpha}})}{\sqrt{nT}} + \frac{1}{H}\sum_{h=1}^{H}\frac{B^h(\theta_0, \widehat{\boldsymbol{\alpha}})}{T} + o_p\Big(\frac{1}{T}\Big).$$

Combined with $\beta(\theta_0, \boldsymbol{\alpha_0}) = \theta_0$,

$$\begin{aligned}\sqrt{nT}\Big(\widehat{\theta} - \widehat{\beta}_H(\theta_0, \widehat{\boldsymbol{\alpha}})\Big) &= \Big(A(\theta_0, \boldsymbol{\alpha_0}) - \frac{1}{H}\sum_{h=1}^{H}A^h(\theta_0, \widehat{\boldsymbol{\alpha}})\Big) \\ &\quad + \sqrt{\frac{n}{T}}\Big(B(\theta_0, \boldsymbol{\alpha_0}) - \frac{1}{H}\sum_{h=1}^{H}B^h(\theta_0, \widehat{\boldsymbol{\alpha}})\Big) + o_p\Big(\sqrt{\frac{n}{T}}\Big).\end{aligned}$$

The rest of the proof shows that bias term cancels out and the asymptotic normality holds. To simplify notation, the rest of the proof proceeds by setting $H = 1$.

Step 1: Bias correction is established in Appendix A.2.4.

Step 2: The simulation analog of the CLT term is

$$A^h(\theta_0, \widehat{\boldsymbol{\alpha}}) = \Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}^h(\theta, \widehat{\alpha}_i)$$

Note that

$$\begin{aligned}&\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}-\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\\ =&\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\big[\mathcal{I}_i(\theta_0,\alpha_{i0})-\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\big]\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}.\end{aligned}$$

By continuous mapping theorem, $\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\xrightarrow{p}\mathcal{I}_i(\theta_0,\alpha_{i0})$ for each i, and thus

$$\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}-\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\Big|\xrightarrow{p}0.$$

Combined with Assumption 8, $A^h(\theta_0,\widehat{\boldsymbol{\alpha}})$ has the same distribution as $A^h(\theta_0,\boldsymbol{\alpha_0})$, and by Proposition 5 in Gouriéroux et al. (1993),

$$A(\theta_0,\boldsymbol{\alpha_0})-\frac{1}{H}\sum_{h=1}^{H}A^h(\theta_0,\boldsymbol{\alpha_0})\sim\mathcal{N}(0,(1+\frac{1}{H})\mathbb{E}(A(\theta_0,\boldsymbol{\alpha_0})A(\theta_0,\boldsymbol{\alpha_0})')).$$

□

A.2.4 Proof of Proposition 1.4.2

Proof. Consider an infeasible fixed effect estimator $\widehat{\beta}_H(\theta_0,\alpha_0)$ that is obtained from data simulated by (θ_0,α_0). Then the representation of $\widehat{\beta}_H(\theta_0,\alpha_0)-\theta_0$ takes the form

$$\widehat{\beta}_H(\theta_0,\boldsymbol{\alpha_0})-\theta_0=\frac{A^h(\theta_0,\boldsymbol{\alpha_0})}{\sqrt{nT}}+\frac{B^h(\theta_0,\boldsymbol{\alpha_0})}{T}+o\Big(\frac{1}{T}\Big),$$

where the superscript h denotes the fact that the dependent variable in $B^h(\theta_0,\boldsymbol{\alpha_0})$ is $y^h_{it}(\theta_0,\alpha_{i0})$. Because $B(\theta_0,\boldsymbol{\alpha_0})$ and $B^h(\theta_0,\boldsymbol{\alpha_0})$ have the same probability limit, they converge to the same expectation, which is the asymptotic bias. Therefore, it suffices to show that $B^h(\theta_0,\widehat{\boldsymbol{\alpha}})$ uniformly well approximates $B^h(\theta_0,\boldsymbol{\alpha_0})$.

Now prove bias correction of the following form:

$$|B^h(\theta,\widehat{\boldsymbol{\alpha}})-B(\theta,\boldsymbol{\alpha_0})|\xrightarrow{p}0.$$

By Markov inequality, $\forall \eta > 0$,

$$Pr(|B^h(\theta, \widehat{\boldsymbol{\alpha}}) - B(\theta, \boldsymbol{\alpha_0})| \geq \eta) \leq \frac{1}{\eta}\mathbb{E}(|B^h(\theta, \widehat{\boldsymbol{\alpha}}) - B(\theta, \boldsymbol{\alpha_0})|).$$

Therefore it suffices to bound the RHS term. By the triangular inequality,

$$\begin{aligned}
&\mathbb{E}(|B^h(\theta, \widehat{\boldsymbol{\alpha}}) - B(\theta, \boldsymbol{\alpha_0})|) \\
&\leq \mathbb{E}(|B^h(\theta, \widehat{\boldsymbol{\alpha}}) - B^h(\theta, \boldsymbol{\alpha_0})|) + \mathbb{E}(|B^h(\theta, \boldsymbol{\alpha_0}) - B(\theta, \boldsymbol{\alpha_0})|). \qquad \text{(A.2.7)}
\end{aligned}$$

The second RHS term in equation (A.2.7) is $o_p(1)$ because $B^h(\theta, \boldsymbol{\alpha_0})$ and $B(\theta, \boldsymbol{\alpha_0})$ have the same probability limit. Regarding the first RHS term, by the triangular inequality,

$$\begin{aligned}
\mathbb{E}|B^h(\theta, \widehat{\boldsymbol{\alpha}}) - B^h(\theta, \boldsymbol{\alpha_0})| \leq\ & \mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}B_i^h(\theta_0, \widehat{\alpha}_i) \\
&- \Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}B_i^h(\theta_0, \alpha_{i0}) \\
&+ \Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}B_i^h(\theta_0, \alpha_{i0}) \\
&- \Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0, \alpha_{i0})\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}B_i^h(\theta_0, \alpha_{i0})\Big| \\
\leq\ & \mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1}\Big| \\
&\times \Big|\frac{1}{n}\sum_{i=1}^{n}\big[B_i^h(\theta_0, \widehat{\alpha}_i) - B_i^h(\theta_0, \alpha_{i0})\big]\Big| \qquad \text{(A.2.8)} \\
&+ \mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0, \widehat{\alpha}_i)\Big)^{-1} - \Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0, \alpha_{i0})\Big)^{-1}\Big| \times \\
&\Big|\frac{1}{n}\sum_{i=1}^{n}B_i^h(\theta_0, \alpha_{i0})\Big|. \qquad \text{(A.2.9)}
\end{aligned}$$

Therefore, it suffices to focus on bounding terms (A.2.8) and (A.2.9).

For term (A.2.9), note that

$$\begin{aligned}
&\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}-\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\\
=&\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\big[\mathcal{I}_i(\theta_0,\alpha_{i0})-\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\big]\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}.
\end{aligned}$$

By continuous mapping theorem, $\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\xrightarrow{p}\mathcal{I}_i(\theta_0,\alpha_{i0})$ for each i. Therefore,

$$\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}-\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\Big|\cdot\Big|\frac{1}{n}\sum_{i=1}^{n}B_i^h(\theta_0,\alpha_{i0})\Big|\xrightarrow{p}0.$$

For term (A.2.8), note that

$$\begin{aligned}
\Big|\frac{1}{n}\sum_{i=1}^{n}\big[B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})\big]\Big|&\le\frac{1}{n}\sum_{i=1}^{n}|B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})|\\
&\le\max_{1\le i\le n}|B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big|\cdot\Big|\frac{1}{n}\sum_{i=1}^{n}\big[B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})\big]\Big|\\
\le&\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big|\cdot\max_{1\le i\le n}|B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})|\\
\le&\sqrt{\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big|^2}\cdot\sqrt{\mathbb{E}\max_{1\le i\le n}\Big|B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})\Big|^2},
\end{aligned}$$

where the second inequality is due to Cauchy–Schwarz inequality. By continuous mapping theorem, $\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\xrightarrow{p}\mathcal{I}_i(\theta_0,\alpha_{i0})$, and by Slutsky theorem,

$$\sqrt{\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big|^2}\xrightarrow{p}\sqrt{\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\alpha_{i0})\Big)^{-1}\Big|^2}.$$

Combined with Assumption 7,

$$\mathbb{E}\Big|\Big(\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_i(\theta_0,\widehat{\alpha}_i)\Big)^{-1}\Big|\cdot\Big|\frac{1}{n}\sum_{i=1}^{n}\big[B_i^h(\theta_0,\widehat{\alpha}_i)-B_i^h(\theta_0,\alpha_{i0})\big]\Big|\xrightarrow{p}0.$$

□

Appendix B

Proofs of Chapter Three

Let Ω be a compact metric space and let Σ_Ω denote its Borel σ-algebra. Let $\mathcal{K}(\Omega)$ be the set of compact subsets of Ω endowed with the Hausdorff metric. Let $\mathcal{C}(\Omega)$ be the set of continuous functions on Ω. Let $\Delta(\Omega)$ be the set of Borel probability measures on Ω endowed with the weak topology.

A set function ν^* is said to be a *capacity* if ν^* satisfies the following conditions:

(i) $\nu^*(\emptyset) = 0, \nu^*(\Omega) = 1$,

(ii) $A \subset B \Rightarrow \nu^*(A) \leq \nu^*(B)$, for all $A, B \in \Sigma_\Omega$.

(iii) $A_n \uparrow A \Rightarrow \nu^*(A_n) \uparrow \nu^*(A)$, for all $\{A_n, n \geq 1\} \subset \Sigma_\Omega$ and $A \in \Sigma_\Omega$.

(iv) $F_n \downarrow F, F_n$ closed $\Rightarrow \nu^*(F_n) \downarrow \nu^*(F)$.

One may define integral operations with respect to capacities as follows. Let $f : \Omega \to \mathbb{R}$ be a measurable function. The *Choquet integral* of f with respect to ν is defined by

$$\int f d\nu \equiv \int_{-\infty}^{0} (\nu(\{\omega : f(\omega) \geq t\}) - \nu(\Omega))dt + \int_{0}^{\infty} \nu(\{\omega : f(\omega) \geq t\})dt, \qquad \text{(B.0.1)}$$

where the integrals on the right hand side are Riemann integrals. A capacity ν is said to be *monotone of order* k or, for short, *k-monotone* if for any $A_i \subset S, i = 1 \cdots, k$,

$$\nu\big(\cup_{i=1}^{k} A_i\big) \geq \sum_{I \subseteq \{1,\cdots,k\}, I \neq \emptyset} (-1)^{|I|+1} \nu\big(\cap_{i \in I} A_i\big). \qquad \text{(B.0.2)}$$

Conjugate $\nu^*(A) = 1 - \nu(A^c)$ is then called a *k-alternating* capacity. A capacity that satisfies (B.0.2) is called an *infinitely monotone capacity* or a *belief function*. Capacities are used in various areas of statistics (Dempster, 1967; Shafer, 1976; Wasserman, 1990) and economics (Gilboa and Schmeidler, 1989).

The following result, known as Choquet's theorem, states that a random closed set K following a distribution M induces a belief function, and it follows from Theorems 1-3 in Philippe et al. (1999).

Lemma B.0.1. *Let Ω be a Polish space. Let M be a probability measure on $\mathcal{K}(\Omega)$. Let $\mathcal{P} = \{P \in \Delta(\Omega) : P = \int P_K dM(K), P_K \in \Delta(K)\}$. Then, $\nu(\cdot) = \inf_{P \in \mathcal{P}} P(\cdot)$ is a belief function and satisfies*

$$\nu(A) = M(\{K \subset A\}). \tag{B.0.3}$$

In our setting, we apply the lemma above with a random subset of $\mathcal{Y} \times \mathcal{X}$. Namely, we take $K = G(u|X;\theta) \times \{X\}$, and M is the law of K induced by u's conditional distribution F_θ and X's marginal distribution q_x. We then denote the induced belief function by ν_θ and its conjugate ν_θ^* (see (B.1.6)-(B.1.7) below).

This section is organized as follows. In Section B.1, we show $\sqrt{n}$-consistency of $\hat{\delta}_n$ by extending standard arguments for extremum estimators to locally incomplete models. In Section B.2, we use the results in B.1 to show results on the asymptotic size of our score test.

B.1 Consistency of Nuisance Parameter Estimates

Lemma B.1.1. *Suppose Assumption 4 holds. Then for any bounded function $g : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$,*

$$\left| \int g d\nu_\theta^* - \int g d\nu_{\theta'}^* \right| \leq C' \|\theta - \theta'\|, \ \forall \theta, \theta' \in \Theta, \tag{B.1.1}$$

and

$$\Big| \int g d\nu_\theta - \int g d\nu_{\theta'} \Big| \leq C' \|\theta - \theta'\|, \ \forall \theta, \theta' \in \Theta, \tag{B.1.2}$$

for some $C' > 0$.

Proof. Note that

$$\begin{aligned} \int g d\nu_\theta^* &= \int \max_{(y,x) \in G(u|x;\theta) \times \{x\}} g(y,x) dF_\theta(u) \\ &= \int \bar{g}(u) f_\theta(u) du, \end{aligned} \tag{B.1.3}$$

where $\bar{g}(u) = \max_{(y,x) \in G(u|x;\theta) \times \{x\}} g(y,x)$. This, boundedness of g, and Assumption 4 imply

$$\begin{aligned} \Big| \int g d\nu_\theta^* - \int g d\nu_{\theta'}^* \Big| &= \Big| \int \bar{g}(u)(f_\theta(u) - f_{\theta'}(u)) du \Big| \\ &\lesssim \|f_\theta - f_{\theta'}\|_{L^1_\zeta} \lesssim \|\theta - \theta'\|. \end{aligned} \tag{B.1.4}$$

This ensures (B.1.1). Showing (B.1.2) is analogous and is omitted. □

The following proposition shows that the sample log-likelihood converges to its population counterpart uniformly over a set of distributions that are consistent with the null or local alternative hypotheses.

Proposition B.1.1 (ULLN)**.** *Suppose Assumptions 1-3 hold. Let* $h \in \mathcal{V}_1$. *Then, for any* $\theta_0 \in \Theta_0$ *and* $\epsilon > 0$, *there exists* $N_\epsilon \in \mathbb{N}$ *that does not depend on* δ *such that*

$$\sup_{P^n \in \mathcal{Q}^n_{\theta_0 + h/\sqrt{n}}} P^n \Big(\sup_{\delta \in \Theta_\delta} \Big| n^{-1} \sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - E_{Q_0}[\ln q_{\beta_0,\delta}] \Big| \geq \epsilon \Big) < \epsilon, \ \forall n \geq N_\epsilon. \tag{B.1.5}$$

Proof. Below, let ν_θ and ν_θ^* be a belief function and its conjugate induced by the correspondence $(u,x) \mapsto G(u|x;\theta) \times \{x\}$ on $\mathcal{Y} \times \mathcal{X}$. That is, they are set functions

such that

$$\nu_\theta(A) = \int_{\mathcal{X}} \int_{\mathcal{U}} 1\{G(u|x;\theta)) \times \{x\} \subseteq A\} dF_\theta(u) dq_x(x),\ A \subset \mathcal{Y} \times \mathcal{X} \tag{B.1.6}$$

$$\nu^*_\theta(A) = \int_{\mathcal{X}} \int_{\mathcal{U}} 1\{G(u|x;\theta)) \times \{x\} \cap A \neq \emptyset\} dF_\theta(u) dq_x(x).\ A \subset \mathcal{Y} \times \mathcal{X}. \tag{B.1.7}$$

A key observation is that, for any $\theta_0 \in \Theta_0$,

$$\int \ln q_{\beta_0,\delta} d\nu_{\theta_0} = E_{Q_{\theta_0}}[\ln q_{\beta_0,\delta}] = \int \ln q_{\beta_0,\delta}(y|x) d\nu^*_{\theta_0}. \tag{B.1.8}$$

This is because the model is complete under H_0 by Assumption 1 and the fact that the Choquet integrals with respect to ν_{θ_0} and $\nu^*_{\theta_0}$ coincide with each other in such a setting.

Note that one may write the event (i.e. the argument of P^n) in (B.1.5) as the union of the following two events:

$$A^U_n = \Big\{ s^n : \sup_{\delta \in \Theta_\delta} \Big(n^{-1} \sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0} \Big) \geq \epsilon \Big\} \tag{B.1.9}$$

$$A^L_n = \Big\{ s^n : \inf_{\delta \in \Theta_\delta} \Big(n^{-1} \sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu_{\theta_0} \Big) \leq -\epsilon \Big\}. \tag{B.1.10}$$

Let $K^n = \prod_{i=1}^n K_i$ be a random set whose distribution follows the law induced by $m_{\theta_0 + h/\sqrt{n}}$. Below, we simply write $K^n \sim m_{\theta_0+h/\sqrt{n}}$. Note that

$$\begin{aligned}
&\sup_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^U_n \cap A^L_n) \\
&= F_{\theta_0+h/\sqrt{n}}\Big(K^n \cap (A^U_n \cup A^L_n) \neq \emptyset \Big) \\
&\leq F_{\theta_0+h/\sqrt{n}}\big(K^n \cap A^U_n \neq \emptyset \big) + F_{\theta_0+h/\sqrt{n}}\big(K^n \cap A^L_n \neq \emptyset \big) \\
&= \nu^*_{\theta_0+h/\sqrt{n}}\Big(\sup_{\delta \in \Theta_\delta} \big[n^{-1} \sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0} \big] \geq \epsilon \Big) \qquad \text{(B.1.11)} \\
&\quad + \nu^*_{\theta_0+h/\sqrt{n}}\Big(\inf_{\delta \in \Theta_\delta} \big[n^{-1} \sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu_{\theta_0} \big] \leq -\epsilon \Big). \qquad \text{(B.1.12)}
\end{aligned}$$

By Assumption 3 and Lemma B.1.1,

$$\Big| \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0+h/\sqrt{n}} - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0} \Big| \leq \frac{C'|h|}{\sqrt{n}}, \tag{B.1.13}$$

implying there exists $\bar{\eta} > 0$ and $N_{\bar{\eta}}$ such that $\sqrt{n} \sup_{\delta \in \Theta_\delta} \Big(\int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0+h/\sqrt{n}} - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0} \Big) < \bar{\eta}$ for all $n \geq N_{\bar{\eta}}$. Hence, for all $n \geq N_{\bar{\eta}}$, (B.1.11) is bounded by

$$\nu^*_{\theta_0+h/\sqrt{n}} \Big(\sup_{\delta \in \Theta_\delta} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big[\ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu^*_{\theta_0+h/\sqrt{n}} \big] \geq \sqrt{n}\epsilon - \bar{\eta} \Big). \tag{B.1.14}$$

As we show below, we may apply Lemma B.1.2 to this quantity. Similarly, by Assumption 3 and Lemma B.1.1, (B.1.12) is bounded by

$$\nu^*_{\theta_0+h/\sqrt{n}} \Big(\inf_{\delta \in \Theta_\delta} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big[\ln q_{\beta_0,\delta}(s_i) - \int \ln q_{\beta_0,\delta} d\nu_{\theta_0+h/\sqrt{n}} \big] \leq -\sqrt{n}\epsilon + \bar{\eta} \Big). \tag{B.1.15}$$

Now, let $\mathcal{G} \equiv \{g = \ln q_{\beta_0,\delta}, \delta \in \Theta_\delta\}$. Then, by Lemma B.1.3, the induced family of functions $\mathcal{F}_\mathcal{G}$ defined in (B.1.18) consists of uniformly bounded and Lipschitz functions. By Theorem 2.7.11 in van der Vaart and Wellner (1996), it follows that

$$N_{[]}(\epsilon \|F_\mathcal{G}\|_{L^2(M)}, \mathcal{F}_\mathcal{G}, L^2(M)) \leq N(\epsilon/2, \Theta_\delta, \|\cdot\|) \leq (2\text{diam}(\Theta_\delta)/\epsilon)^{d_\delta}. \tag{B.1.16}$$

Therefore, $\mathcal{F}_\mathcal{G}$ satisfies the condition of Lemma B.1.2. Applying the lemma ensures that (B.1.14) is bounded by

$$\Big(C_{\text{diam}(\Theta_\delta)} \frac{\sqrt{n}\epsilon - \bar{\eta}}{\sqrt{d_\delta}} \Big)^{d_\delta} e^{-2(\sqrt{n}\epsilon - \bar{\eta})^2}, \tag{B.1.17}$$

which tends to 0 as $n \to \infty$. (B.1.15) can be handled similarly. This completes the proof. □

Let S be a Euclidean space. Given a family $\mathcal{G}$ of measurable functions on S and a random set $K : \Omega \mapsto \mathcal{K}(S)$, define a family of measurable functions on $\mathcal{K}(S)$ by

$$\mathcal{F}_\mathcal{G} \equiv \Big\{ f : f(K) = \max_{s \in K} g(s), \ g \in \mathcal{G} \Big\}. \tag{B.1.18}$$

We denote the envelope function of $\mathcal{F}_\mathcal{G}$ by $F_\mathcal{G}$. A class $\mathcal{F}$ of uniformly bounded functions is covered by at most $(\frac{D}{\epsilon})^v$ brackets if for positive constants v and D,

$$N_{[]}(\epsilon\|F\|_{L^2(M)}, \mathcal{F}, L^2(M)) \leq \Big(\frac{D}{\epsilon}\Big)^v, \quad 0 < \epsilon < D, \tag{B.1.19}$$

The following lemma gives concentration inequalities for the suprema (and infima) of empirical processes under plausibility functions.

Lemma B.1.2. *Let ν^n be a belief function such that $\nu^n(B) = M^n(K^n \subset A)$ for any $A \in \mathcal{K}(S^n)$. Let $\mathcal{G}$ be a family of uniformly bounded measurable functions on S such that $\mathcal{F}_\mathcal{G}$ in* (B.1.18) *is covered by at most $(\frac{D}{\epsilon})^v$ brackets. Then, for all $t > 0$*

$$\nu^{*,n}\Big(\sup_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[g(s_i) - \int g d\nu^*\big] \geq t\Big) \leq \Big(C_D \frac{t}{\sqrt{v}}\Big)^v e^{-2t^2}, \tag{B.1.20}$$

$$\nu^{*,n}\Big(\inf_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[g(s_i) - \int g d\nu\big] \leq -t\Big) \leq \Big(C_D \frac{t}{\sqrt{v}}\Big)^v e^{-2t^2}. \tag{B.1.21}$$

for some C_D that depends on D only.

Proof. Define the following events

$$B_n^U = \Big\{s^n : \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[g(s_i) - \int g d\nu^*\big] \geq t\Big\} \tag{B.1.22}$$

$$B_n^L = \Big\{s^n : \inf_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[g(s_i) - \int g d\nu\big] \leq -t\Big\}. \tag{B.1.23}$$

Observe that

$$K^n \cap B_n^U \neq \emptyset \;\Leftrightarrow\; \sup_{s^n\in K^n} \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[g(s_i) - \int g d\nu^*\big] \geq t \tag{B.1.24}$$

$$\Leftrightarrow\; \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[\sup_{s_i\in K_i} g(s_i) - \int g d\nu^*\big] \geq t \tag{B.1.25}$$

$$\Leftrightarrow\; \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[\sup_{s_i\in K_i} g(s_i) - \int \sup_{s\in K} g(s) dM(K)\big] \geq t \tag{B.1.26}$$

$$\Leftrightarrow\; \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[f(K_i) - E_M[f(K)]\big] \geq t. \tag{B.1.27}$$

Therefore,

$$\begin{aligned}
&\nu^{*,n}\Big(\sup_{g\in\mathcal{G}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[g(s_i)-\int g d\nu^*\big]\geq t\Big)\\
&=M^n(K^n\cap B_n^U\neq\emptyset)\\
&=M^n\Big(\sup_{g\in\mathcal{G}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[f(K_i)-E_M[f(K)]\big]\geq t\Big),
\end{aligned}\tag{B.1.28}$$

By Theorem 1.3 (ii) in Talagrand (1994), for all $t>0$,

$$\begin{aligned}
M^n\Big(\sup_{g\in\mathcal{G}}\big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f(K_i)-E_M[f(K)]\big)\geq t\Big)&\leq M^n\Big(\|\mathbb{G}_n f\|_{\mathcal{F}}\geq t\Big)\\
&\leq\Big(C_D\frac{t}{\sqrt{v}}\Big)^v e^{-2t^2}.
\end{aligned}\tag{B.1.29}$$

A similar argument can be applied to B_n^L as well. □

Let K be a subset of $\mathcal{S}=\mathcal{Y}\times\mathcal{X}$. The following lemma shows that $f_\delta(K)\equiv\max_{(y,x)\in K}\ln q_{\beta_0,\delta}(y|x)$ is uniformly bounded and Lipschitz, which provides a control of the covering number.

Lemma B.1.3. *Suppose Assumption 3 holds. Then, (i) f_δ is uniformly bounded; and (ii) for any $\delta,\delta'\in\Theta_\delta$,*

$$|f_\delta(K)-f_{\delta'}(K)|\lesssim\|\delta-\delta'\|.\tag{B.1.30}$$

Proof. (i) follows from the map $\delta\mapsto M(\delta)$ being continuous by Assumption 3 and hence achieving a finite maximum on the compact set Θ_δ.

(ii) Let $s=(y,x)$ and let $g(\delta,s)=\ln q_{\beta_0,\delta}(y|x)$. By Assumption 3,

$$\begin{aligned}
f_{\delta'}(K)&=\max_{(y,x)\in K}\Big(g(\delta',s)-g(\delta,s)+g(\delta,s)\Big)\\
&\leq\max_{s\in K}\Big(\|\delta-\delta'\|+g(\delta,s)\Big)=f_\delta(K)+\|\delta-\delta'\|.
\end{aligned}\tag{B.1.31}$$

Similarly,

$$f_{\delta'}(K) = \max_{s\in K}\Big(g(\delta',s) - g(\delta,s) + g(\delta,s)\Big)$$
$$\geq \max_{s\in K}\Big(- \|\delta-\delta'\| + g(\delta,s)\Big) = f_{\delta}(K) - \|\delta-\delta'\|. \quad \text{(B.1.32)}$$

Combining the two inequalities above yields (B.1.30). □

Below, we write $X_n = O_{P^n}(a_n)$ uniformly in $P^n \in \mathcal{F}_n$ if for any $\epsilon > 0$, there exist finite $M > 0$ and $N > 0$ such that $\sup_{P^n\in\mathcal{F}_n} P^n(|X_n/a_n| > M) < \epsilon$ for all $n > N$.

Theorem B.1.1. *Suppose Assumptions 1-3 hold. Then, for any $\eta > 0$,*

$$\lim_{n\to\infty} \inf_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n\Big(\|\hat{\delta}_n - \delta_0\| < \eta\Big) = 1 \quad \text{(B.1.33)}$$

and $\mathbb{M}_n(\hat{\delta}_n) \geq \mathbb{M}_n(\delta_0) - O_{P^n}(r_n^{-2})$ uniformly in $P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}$.

Proof. The proof is based on the standard argument for the consistency of extremum estimators (see e.g. Newey and McFadden, 1994). A slight difference is that one needs a uniform law of large numbers under any sequence $P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}$, which is established by Proposition B.1.1. For each $\delta \in \Theta_\delta$, recall that $\mathbb{M}(\delta) \equiv E_{Q_0}[\ln q_{\beta_0,\delta}]$ and let $\mathbb{M}_n(\delta) \equiv n^{-1}\sum_{i=1}^n \ln q_{\beta_0,\delta}(s_i)$. Given any neighborhood V of δ_0, we want to show that $\hat{\delta}_n \in V$, wp$\to$ 1 uniformly over $\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}$. For this, it suffices to show that $\inf_{P^n\in\mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(\mathbb{M}(\hat{\delta}_n) < \inf_{\delta\in\Theta\cap V^c}\mathbb{M}(\delta)) \to 1$. Let $\epsilon \equiv \inf_{\delta\in\Theta\cap V^c}\mathbb{M}(\delta) - \mathbb{M}(\delta_0)$. This constant is well-defined since $\inf_{\Theta\cap V^c}\mathbb{M}(\delta) = \mathbb{M}(\delta^*) > \mathbb{M}(\delta_0)$ for some $\delta^* \in \Theta\cap V^c$ by Assumption 3 and the compactness of Θ_δ.

Let $A_{1n} \equiv \{\omega : \mathbb{M}(\hat{\delta}_n) < \mathbb{M}_n(\hat{\delta}_n)+\epsilon/3\}$, $A_{2n} \equiv \{\omega : \mathbb{M}_n(\hat{\delta}_n) < \mathbb{M}_n(\delta_0)+\epsilon/3\}$, $A_{3n} \equiv \{\omega : \mathbb{M}_n(\delta_0) < \mathbb{M}(\delta_0)+\epsilon/3\}$. For any $\omega \in A_{1n}\cap A_{2n}\cap A_{3n}$,

$$\begin{aligned}\mathbb{M}(\hat{\delta}_n) &< \mathbb{M}_n(\hat{\delta}_n) + \epsilon/3\\ &< \mathbb{M}_n(\delta_0) + 2\epsilon/3\\ &< \mathbb{M}(\delta_0) + \epsilon.\end{aligned}$$

Therefore,

$$\begin{aligned}
&\inf_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(\mathbb{M}(\hat{\delta}_n) < \mathbb{M}(\delta_0) + \epsilon) \\
&\geq \inf_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A_{1n} \cap A_{2n} \cap A_{3n}) && \text{(B.1.34)} \\
&\geq \inf_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} \Big(1 - P^n(A^c_{1n}) - P(A^c_{2n}) - P(A^c_{3n})\Big) && \text{(B.1.35)} \\
&\geq 1 - \sum_{j=1}^{3} \sup_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^c_{jn}). && \text{(B.1.36)}
\end{aligned}$$

Note that, for any h,

$$\sup_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^c_{1n}) \to 0, \qquad \sup_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^c_{3n}) \to 0$$

by Proposition B.1.1. Also note that by the construction of $\hat{\delta}_n$,

$$\sup_{P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}} P^n(A^c_{2n}) \to 0.$$

□

Proof of Proposition 3.3.1. The result follows immediately from Theorem 3.2.5 in van der Vaart and Wellner (1996) with $\phi_n(\zeta) = \zeta$, $r_n = \sqrt{n}$, and applying their argument uniformly over $P^n \in \mathcal{Q}^n_{\theta_0+h/\sqrt{n}}$. □

B.2 Size Control

Proof of Theorem 3.3.1. We show the result for the general setting, in which the orthogonalized score is used to construct $\hat{S}_n$. To simplify the exposition below, we assume I_{θ_0} is known for now. Let

$$g^*_n(\beta_0) = C^*_{\beta_0,n} - I_{\beta,\delta} I_\delta^{-1} C^*_{\delta,n}, \tag{B.2.1}$$

where

$$C^*_{\beta_0,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\beta(Y_i|X_i;\beta_0,\delta_0), \quad C^*_{\delta,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\delta(Y_i|X_i;\beta_0,\delta_0). \tag{B.2.2}$$

By Assumption 1, $P_0^n = Q_{\theta_0}^n \times q_X^n$ for some unique product measure. By Assumption 2 and arguing as in Theorem 7.2 in van der Vaart (2000), we have $E[s_\beta(Y_i|X_i;\beta_0,\delta_0)] = E[s_\delta(Y_i|X_i;\beta_0,\delta_0)] = 0$, where expectation is with respect to P_0. By the square integrability of s_β and s_δ ensured by Assumption 2, the central limit theorem for i.i.d. sequences ensures

$$C_n^* = \begin{bmatrix} C_{\beta_0,n}^* \\ C_{\delta,n}^* \end{bmatrix} \overset{P_0^n}{\rightsquigarrow} N(0, I_{\theta_0}). \tag{B.2.3}$$

Observing that $\sqrt{n} g_n^*(\beta_0) = [I_{d_\beta}, -I_{\beta,\delta} I_\delta^{-1}] C_n^*$ and applying the continuous mapping theorem, we obtain

$$\sqrt{n} g_n^*(\beta_0) \overset{P_0^n}{\rightsquigarrow} N(0, V_0), \tag{B.2.4}$$

where $V_0 = I_\beta - I_{\beta,\delta} I_\delta^{-1} I_{\delta,\beta}$. Define

$$S_n^* = n g_n^*(\beta_0)' V_0^{-1} g_n^*(\beta_0) - \inf_{h \in \mathcal{V}_1} n (g_n^*(\beta_0) - h)' V_0^{-1} (g_n^*(\beta_0) - h). \tag{B.2.5}$$

By (B.2.4), it then follows that

$$S_n^* \overset{P_0^n}{\rightsquigarrow} S, \tag{B.2.6}$$

where S is as in (3.43).

For the desired result, it remains to show S_n is asymptotically equivalent to S_n^* under P_0^n. For each δ, let $\mathbb{G}_n s_\beta(\delta) \equiv \frac{1}{n} \sum_{i=1}^n s_\beta(Y_i|X_i;\beta_0,\delta) - E[s_\beta(Y_i|X_i;\beta_0,\delta)]$. We may then write

$$\begin{aligned}
\sqrt{n} g_n(\beta_0) - \sqrt{n} g_n^*(\beta_0) &= \mathbb{G}_n s_\beta(\hat{\delta}_n) - \mathbb{G}_n s_\beta(\delta_0) \\
&\quad - \sqrt{n}(E[s_\beta(Y_i|X_i;\beta_0,\hat{\delta}_n)] - E[s_\beta(Y_i|X_i;\beta_0,\delta_0)]) \\
&\quad - I_{\beta,\delta} I_\delta^{-1} C_{\delta,n} + I_{\beta,\delta} I_\delta^{-1} C_{\delta,n}^* \\
&= \mathbb{G}_n s_\beta(\hat{\delta}_n) - \mathbb{G}_n s_\beta(\delta_0) - I_{\beta,\delta} I_\delta^{-1} C_{\delta,n}^* + o_{P^n}(1) \\
&\quad - I_{\beta,\delta} I_\delta^{-1} C_{\delta,n} + I_{\beta,\delta} I_\delta^{-1} C_{\delta,n}^* \\
&= o_{P^n}(1),
\end{aligned} \tag{B.2.7}$$

where the last equality follows from the stochastic equicontinuity of $\mathbb{G}_n s_\beta$, $\sqrt{n}$-consistency of $\hat{\delta}_n$, and $C_{\delta,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n s_\delta(Y_i|X_i;\beta_0,\hat{\delta}_n) = o_{P^n}(1)$ by the first-order

condition for the RMLE.

Let $\varphi(x) = x'V_0^{-1}x - \inf_{h\in\mathcal{V}_1}(x-h)'V_0^{-1}(x-h)$. Note that $x \mapsto \inf_{h\in\mathcal{V}_1}(x-h)'V_0^{-1}(x-h)$ is continuous due to Berge's maximum theorem (Aliprantis and Border, 2006, Theorem 17.31). Hence, φ is continuous. By (B.2.7) and the continuous mapping theorem,

$$\hat{S}_n - S_n^* = \varphi(\sqrt{n}g_n(\beta_0)) - \varphi(\sqrt{n}g_n^*(\beta_0)) = o_{P^n}(1). \tag{B.2.8}$$

By (B.2.6) and (B.2.8),

$$\lim_{n\to\infty} P_0^n(\hat{S}_n > c_\alpha) = \alpha. \tag{B.2.9}$$

This establishes the claim of the theorem.

Note that we assumed I_{θ_0} was known. In general, it can be consistently estimated by $\hat{I}_n = n^{-1}\sum_{i=1}^n s_\theta(Y_i|X_i;\beta_0,\hat{\delta}_n)s_\theta(Y_i|X_i;\beta_0,\hat{\delta}_n)'$. To see this, let $s_{\theta,j}$ be the j-th component of s_θ. For each j and k, define

$$\xi_{j,k}(Y_i,X_i;\delta) \equiv s_{\theta,j}(Y_i|X_i;\beta_0,\delta)s_{\theta,k}(Y_i|X_i;\beta_0,\delta). \tag{B.2.10}$$

For the (j,k)-th component of $\hat{I}_n$, we then have

$$\begin{aligned}
[\hat{I}_n]_{j,k} - [I_{\theta_0}]_{j,k} &= \frac{1}{n}\sum_{i=1}^n \xi_{j,k}(Y_i,X_i;\hat{\delta}_n) - E_{q_{\theta_0}}[\xi_{j,k}(Y_i,X_i;\delta_0)] \\
&= \big(\frac{1}{n}\sum_{i=1}^n \xi_{j,k}(Y_i,X_i;\hat{\delta}_n) - E_{q_{\theta_0}}[\xi_{j,k}(Y_i,X_i;\hat{\delta}_n)\big) \\
&\quad + \big(E_{q_{\theta_0}}[\xi_{j,k}(Y_i,X_i;\hat{\delta}_n) - E_{q_{\theta_0}}[\xi_{j,k}(Y_i,X_i;\delta_0)\big) = o_{P^n}(1),
\end{aligned}$$

where the last equality follows because

$$\sup_\delta |\frac{1}{n}\sum_{i=1}^n \xi_{j,k}(Y_i,X_i;\delta) - E_{q_{\theta_0}}[\xi_{j,k}(Y_i,X_i;\delta)| = o_{P^n}(1)$$

by Assumption 6, $\hat{\delta}_n$'s consistency by Theorem B.1.1, and the continuity of $\delta \mapsto E[\xi_{j,k}(Y_i,X_i;\delta)]$. Given this, showing the claim of the theorem with the estimated I_{θ_0} is straightforward by applying Slutsky's theorem.

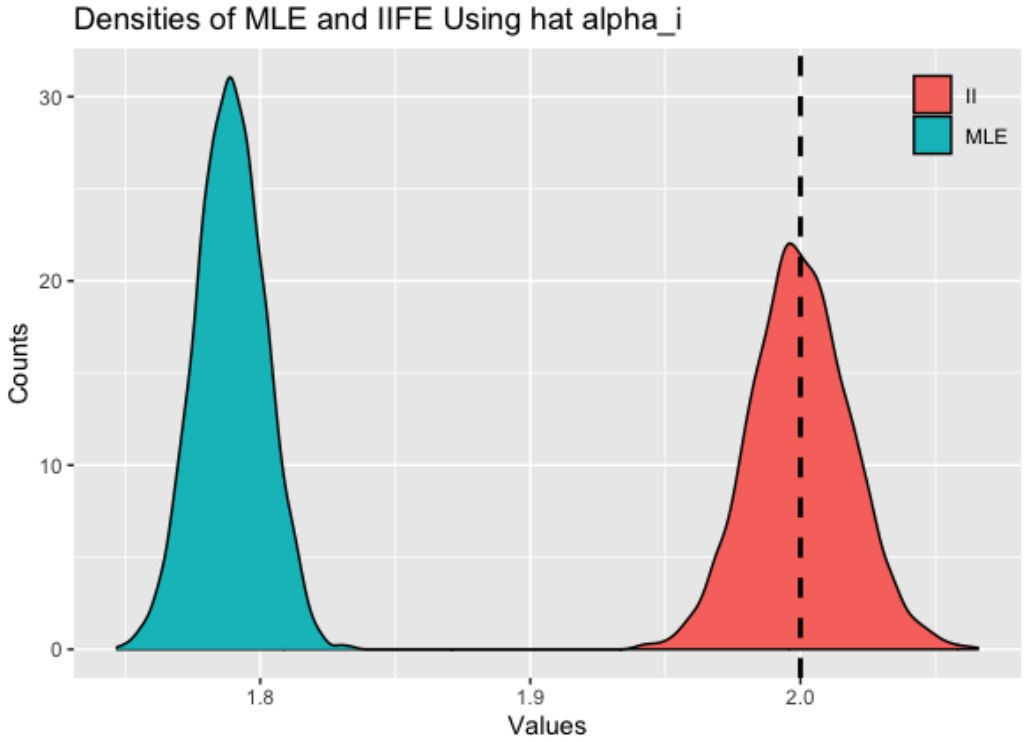
□

Appendix C

Figures

C.1 Figures of Chapter One

Figure C·1: A Comparison between FE and IFE



Note: Density plots of fixed effects and indirect fixed effect estimator for θ_0. The DGP is $y_{it} = \alpha_{i0} + \sqrt{\theta_0} u_{it}$, where $u_{it} \sim \mathcal{N}(0,1)$. The true value $\theta_0 = 2$ is depicted by the dashed line and $\alpha_{i0} = i$ for $i = 1, \ldots, n$. The sample size is $n = 2500, T = 5$ and number of simulation H is set to be 1. The simulations are conducted 5000 times.

C.2 Figures of Chapter Two

Figure C·2: A Graphical Illustration of Split–Panel Splitting

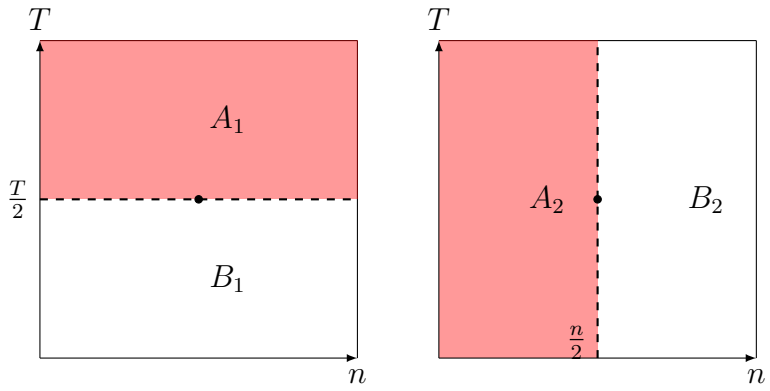


Figure C·3: A Graphical Illustration of Crossover Splitting

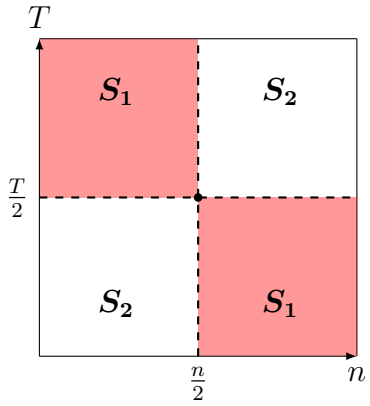
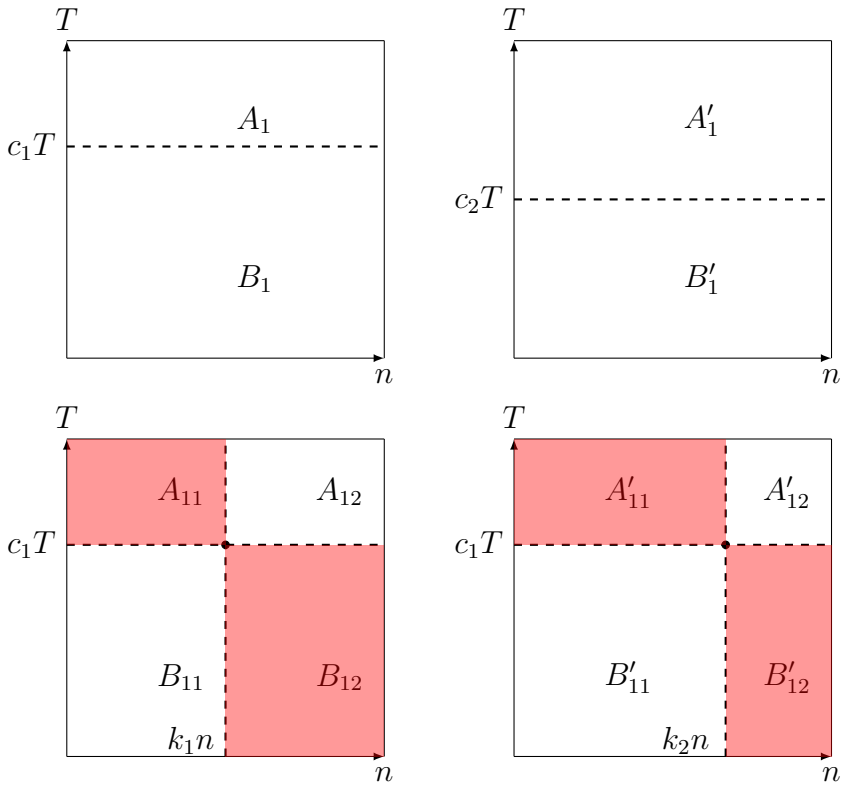
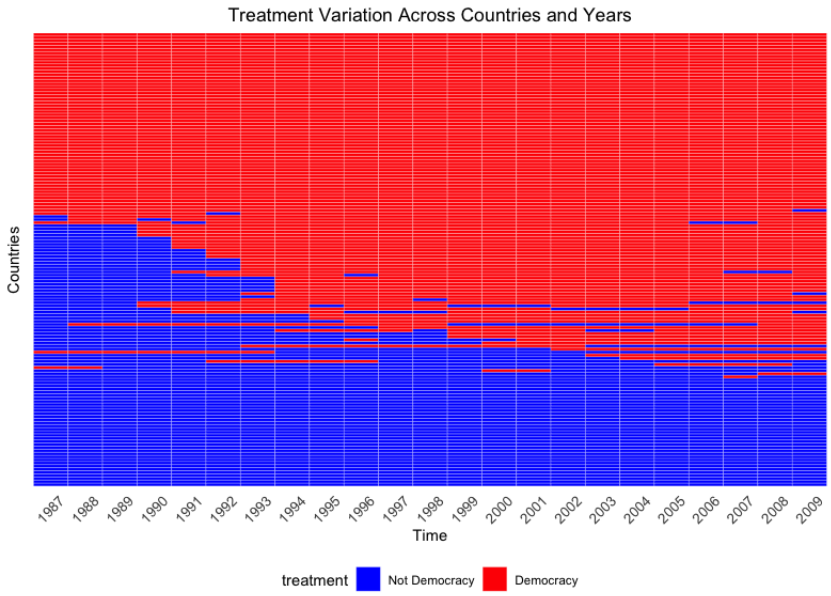


Figure C·4: A Graphical Illustration of Sub–panel Asymptotics under the Crossover Splitting



Note: The upper two graphs illustrate two possible splits along the time series. The constants c_1 and c_2 are both between 0 and 1. The bottom two graphs illustrate two possible splits along the cross section, conditional on the same split along the time series. The constants k_1 and k_2 are both between 0 and 1.

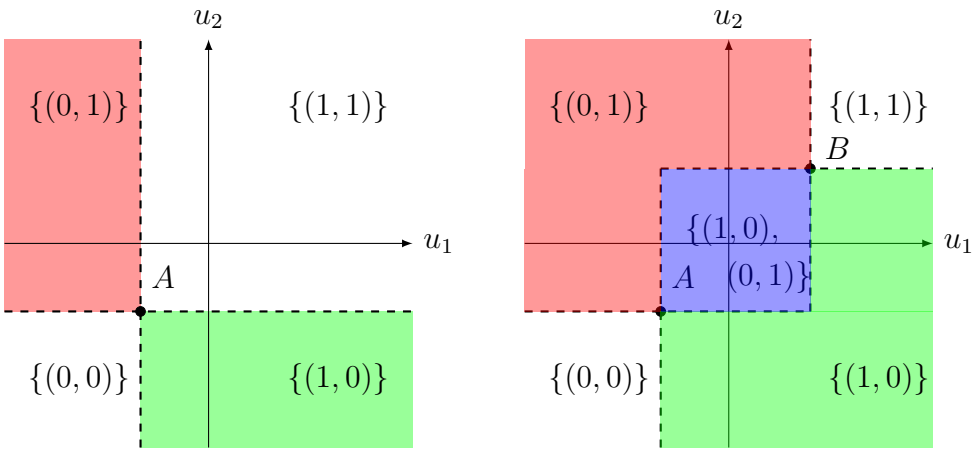
Figure C·5: Non–stationarity of the Treatment Variable



Note: The x–axis denotes years and the y–axis denotes the cross sectional units (countries). A red rectangle denotes that the country has the democracy treatment variable being 1 in that year, while a blue rectangle means that the treatment variable takes value 0. The graph is produced using the R package by Imai et al. (2021).

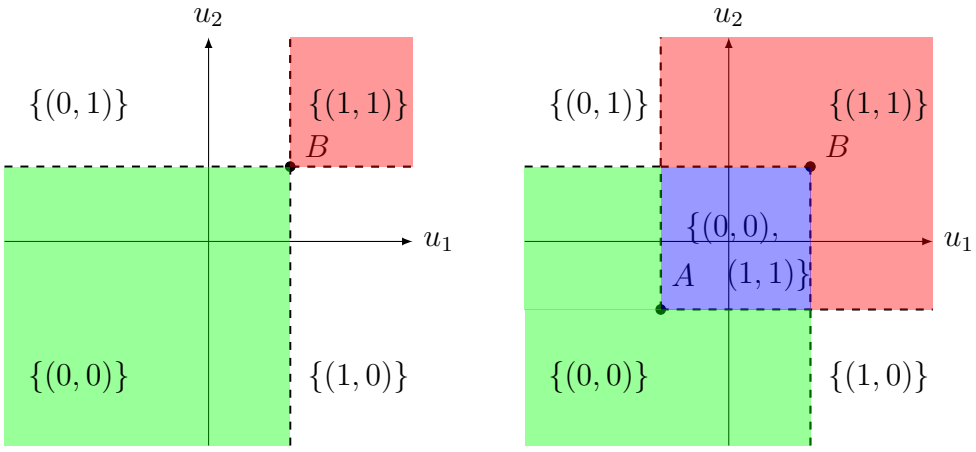
C.3 Figures of Chapter Three

Figure C·6: Level Sets of $u \mapsto G(u|x;\theta)$ (Example 1)



Note: $A = (-x^{(1)\prime}\delta^{(1)}, -x^{(2)\prime}\delta^{(2)})$; $B = (-x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, -x^{(2)\prime}\delta^{(2)} - \beta^{(2)})$. Left Panel: $\beta^{(1)} = \beta^{(2)} = 0$ and the model is complete. Right panel: $\beta^{(1)} < 0$ and $\beta^{(2)} < 0$ and the model is incomplete. U_1 in (3.4) corresponds to the region in green, and similarly U_2 is the region in red. Multiple equiribria $\{(1,0),(0,1)\}$ are predicted in the blue region.

Figure C·7: Level Sets of $u \mapsto G(u|x;\theta)$ (Example 2)



Note: $A = (-x^{(1)\prime}\delta^{(1)} - \beta^{(1)}, -x^{(2)\prime}\delta^{(2)} - \beta^{(2)})$; $B = (-x^{(1)\prime}\delta^{(1)}, -x^{(2)\prime}\delta^{(2)})$. Left Panel: $\beta^{(1)} = \beta^{(2)} = 0$ and the model is complete. Right panel: $\beta^{(1)} > 0$ and $\beta^{(2)} > 0$ and the model is incomplete. U_1 in (3.5) corresponds to the region in green, and similarly U_2 is the region in red. Multiple equiribria $\{(0,0),(1,1)\}$ are predicted in the blue region.

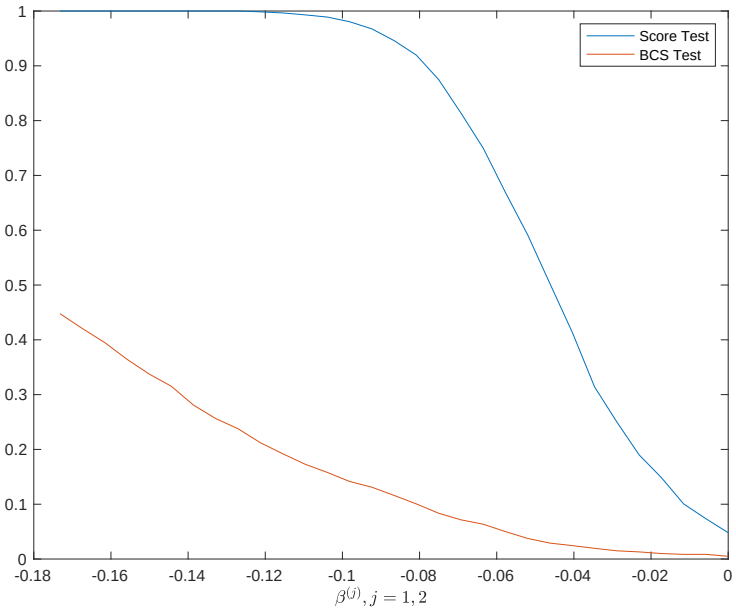


Figure C·8: Power of the Score and BCS Tests (Design 1)

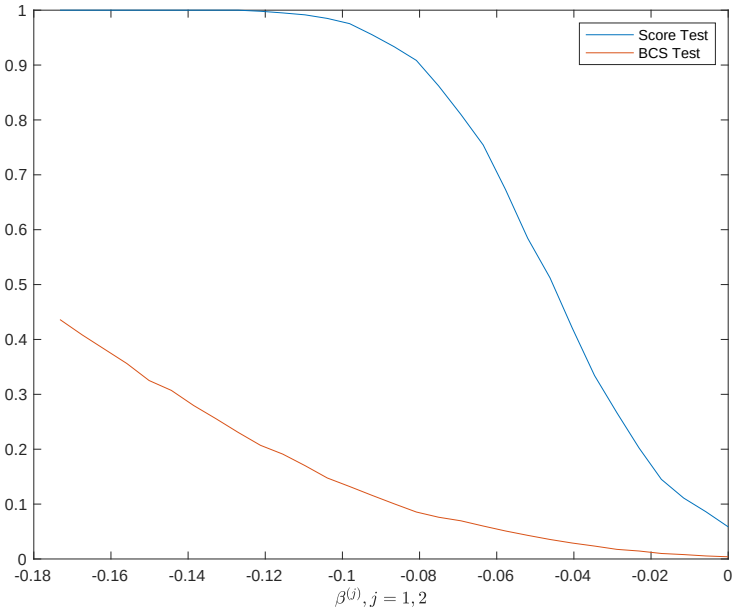


Figure C·9: Power of the Score and BCS Tests (Design 2)

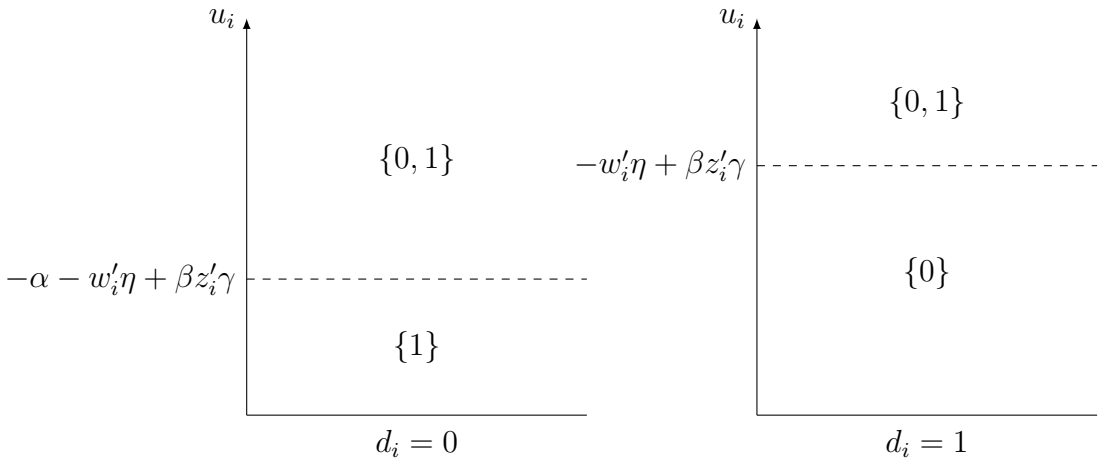
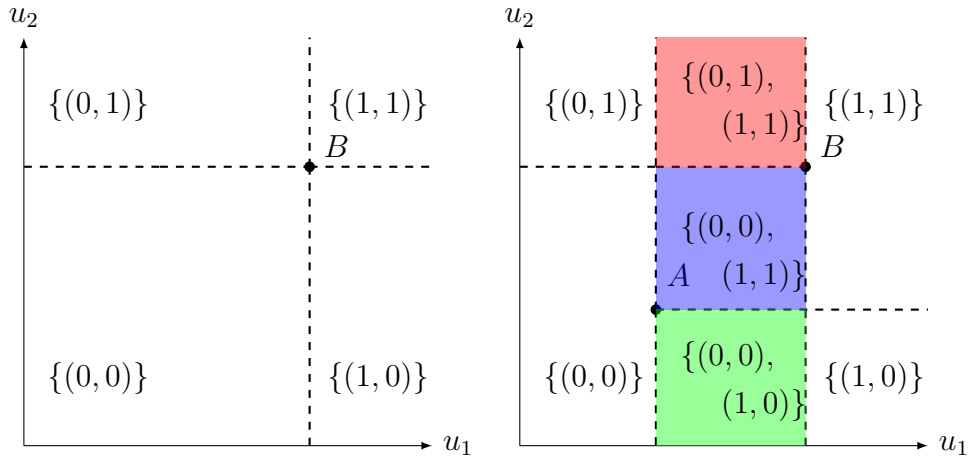


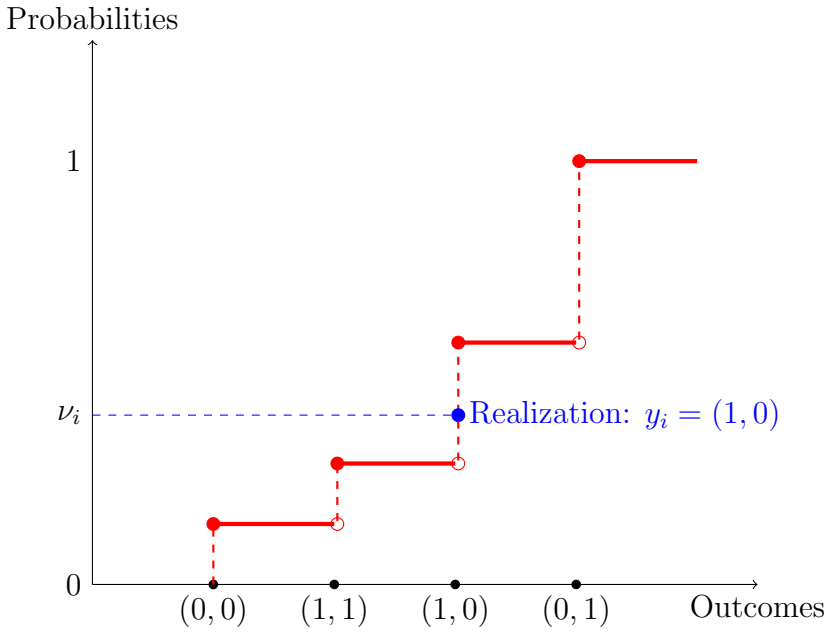
Figure C·10: The Set of Predicted Outcomes $G(u|d_i, w_i, z_i; \theta)$ When $\beta > 0$

Figure C·11: Level Sets of $u \mapsto G(u|x;\theta)$ When $\beta \geq 0$



Note: The level sets of G when $\beta = 0$ (left) and $\beta > 0$ (right). $A = (-x_{i1}'\lambda - \beta, -x_{i2}'\lambda - \beta)$; $B = (-x_{i1}'\lambda, -x_{i2}'\lambda)$.
Multiple outcome values are predicted in each of the red, blue, and green regions.

Figure C·12: CDF of LFP of One Observation



Appendix D

Tables

D.1 Tables of Chapter One

Table D.1: Parameter Estimates for Static Specification

	kids0_2	kids3_5	kids6_17	loghusinc	age	age2
FE	-0.71	-0.41	-0.13	-0.24	2.32	-0.29
	(0.06)	(0.05)	(0.04)	(0.05)	(0.38)	(0.05)
IFE–1	-0.65	-0.36	-0.08	-0.17	2.24	-0.29
	(0.08)	(0.07)	(0.06)	(0.08)	(0.53)	(0.07)
IFE–10	-0.60	-0.32	-0.10	-0.30	2.08	-0.27
	(0.06)	(0.05)	(0.04)	(0.06)	(0.39)	(0.05
IFE–20	-0.60	-0.32	-0.10	-0.30	2.08	-0.27
	(0.06)	(0.05)	(0.04)	(0.06)	(0.38)	(0.05)
ABC	-0.63	-0.37	-0.11	-0.22	2.39	-0.25
	(0.06)	(0.05)	(0.04)	(0.05)	(0.38)	(0.05)
BC–HN	-0.62	-0.36	-0.10	-0.21	1.73	-0.22
	(0.06)	(0.05)	(0.04)	(0.05)	(0.38)	(0.05)
SBC	-0.92	-0.58	-0.26	-0.30	2.28	-0.26
	(0.09)	(0.09)	(0.08)	(0.07)	(0.89)	(0.12)

Notes: Standard errors are reported in the parenthesis and are computed based on the Hessian matrix of profiled log likelihood. The SBC estimates and standard errors computation follows page 1025 in Dhaene and Jochmans (2015).

Table D.2: Simulation Results for Static Specification (Part I)

	FE			IFE–10			IFE–20		
	Bias	Std Dev	Cvge	Bias	Std Dev	Cvge	Bias	Std Dev	Cvge
kids0_2	14.75	9.62	0.79	-4.78	8.18	0.94	-4.21	8.80	0.93
kids3_5	14.74	14.27	0.91	-6.83	13.12	0.95	-6.43	13.23	0.94
kids6_17	14.49	36.58	0.94	-18.06	38.81	0.94	-16.98	39.83	0.93
loghusinc	14.87	25.83	0.94	-3.34	26.76	0.97	-4.66	24.97	0.96
age	13.53	19.34	0.92	0.24	7.79	0.97	-0.87	9.98	0.98
age2	13.47	20.61	0.92	-2.25	22.51	0.96	-4.00	23.00	0.96

Notes: FE denotes fixed effects estimates. IFE–10 and IFE–20 denote indirect fixed effect estiamtes with H being 10 and 20. Cvge denotes the empirical coverage probability. The nominal coverage is 95%. Simulations are conducted 1000 times, and all relative statistics are multiplied by 100. The nominal coverage is 95%.

Table D.3: Simulation Results for Static Specification (Part II)

	ABC			BC–HN			SBC		
	Bias	Std Dev	Cvge	Bias	Std Dev	Cvge	Bias	Std Dev	Cvge
kids0_2	1.18	8.39	0.95	-3.46	8.09	0.95	-5.22	12.47	0.96
kids3_5	1.35	12.59	0.96	-3.33	12.17	0.96	-4.70	21.12	0.98
kids6_17	1.54	32.39	0.95	-3.50	31.12	0.96	-4.16	56.48	0.98
loghusinc	1.55	22.75	0.96	-3.43	21.87	0.96	-6.11	28.27	0.98
age	0.38	27.50	0.97	4.27	16.67	0.96	-4.02	34.81	0.98
age2	0.48	18.41	0.96	-4.36	17.78	0.94	-3.95	37.07	0.98

Notes: ABC denotes analytical bias correction in Fernández-Val (2009). BC–HN denotes leave–one–out jackknife bias correction in Hahn and Newey (2004). SBC denotes split–panel bias correction in Dhaene and Jochmans (2015). Cvge denotes the empirical coverage probability. The nominal coverage is 95%. Simulations are conducted 1000 times, and all relative statistics are multiplied by 100. The nominal coverage is 95%.

Table D.4: Estimates of θ_0

	FE			IFE–1			IFE–10			SBC		
	Bias	Std Dev	Coverage	Bias	Std Dev	Coverage	Bias	Std Dev	Coverage	Bias	Std Dev	Coverage
$n = 100, T = 4$	39.81	39.69	0.87	-3.00	44.06	0.99	-2.98	38.52	0.96	-42.41	80.30	0.94
$n = 100, T = 8$	18.58	14.27	0.86	1.45	17.33	0.98	0.94	12.98	0.96	-7.00	25.09	0.96
$n = 100, T = 12$	12.93	9.89	0.85	0.29	11.17	0.99	0.45	10.12	0.95	-3.50	16.55	0.97
$n = 200, T = 4$	41.21	27.17	0.78	1.67	30.70	0.99	1.17	27.97	0.97	-36.56	53.53	0.92
$n = 200, T = 8$	18.25	10.31	0.76	1.17	12.84	0.96	0.83	10.85	0.96	-6.11	17.94	0.96
$n = 200, T = 12$	13.34	6.92	0.73	1.47	8.29	0.95	1.25	6.93	0.95	-2.38	11.65	0.97

Notes: FE denotes fixed effects estimates. IFE–1 and IFE–10 denote indirect fixed effect estiamtes with H being 1 and 10 respectively. SBC denotes split–sample jackknife bias correction in Dhaene and Jochmans (2015). Cvge denotes the empirical coverage probability. The nominal coverage is 95%. Simulations are conducted 1000 times, and all the statistics are multiplied by 100.

Table D.5: Parameter Estimates for Dynamic Specification

	lfp_lagged	kids0_2	kids3_5	kids6_17	loghusinc	age	age2
FE	0.76	-0.55	-0.28	-0.07	-0.25	2.05	-0.25
	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.38)	(0.05)
IFE–1	0.80	-0.41	-0.25	-0.06	-0.31	2.04	-0.24
	(0.06)	(0.08)	(0.08)	(0.06)	(0.08)	(0.54)	(0.07)
IFE–10	1.09	-0.39	-0.07	-0.04	-0.32	1.78	-0.19
	(0.04)	(0.06)	(0.06)	(0.04)	(0.06)	(0.40)	(0.05)
IFE–20	1.11	-0.48	-0.22	-0.07	-0.28	1.75	-0.23
	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.39)	(0.05)
SBC	1.35	-0.63	-0.34	-0.15	-0.31	1.79	-0.20
	(0.05)	(0.09)	(0.09)	(0.08)	(0.07)	(0.88)	(0.12)
ABC	0.99	-0.48	-0.21	-0.06	-0.23	1.84	-0.22
	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.38)	(0.05)

Notes: Standard errors are stored in the parenthesis and are computed based on the Hessian matrix of profiled log likelihood. For details of the SBC estimates and standard errors computation, refer to page 1025 in Dhaene and Jochmans (2015).

Table D.6: Simulation Results for Dynamic LFP

	FE			IFE–10			SBC		
	Bias	Std Dev	RMSE	Bias	Std Dev	RMSE	Bias	Std Dev	RMSE
lfp_lagged	-53.59	5.84	0.50	3.06	6.22	0.91	-6.43	7.32	0.92
kids0_2	33.45	13.64	0.62	-5.81	9.69	0.96	7.62	17.27	0.97
kids3_5	47.88	24.37	0.70	-8.53	18.65	0.96	24.14	31.89	0.96
kids6_17	53.38	73.44	0.91	-23.29	55.91	0.97	33.74	98.06	0.97
loghusinc	24.08	28.90	0.90	5.29	44.90	0.98	5.70	31.67	0.98
age	29.49	19.34	0.84	1.44	5.54	0.97	-1.46	33.73	0.97
age2	29.07	26.91	0.86	-1.67	20.75	0.98	-1.04	36.54	0.97

Notes: FE denotes fixed effects estimates. IFE–10 denotes indirect fixed effect estiamtes with $H = 10$. SBC denotes split–sample jackknife method. Simulations are conducted 1000 times, and all relative statistics are multiplied by 100. The nominal coverage is 95%.

D.2 Tables of Chapter Two

Table D.7: Calibrated Monte Carlo Dynamic Probit, N = 664, T = 9

	Bias	Std Dev	RMSE	BSE/SD	p.95 (BSE)	Length (BSE)
			Coefficient of lag-lfp			
FE	-53.83	6.20	54.19	0.97	0.00	0.23
SBC	-5.67	7.79	9.63	0.96	0.88	0.29
ABC	-10.54	6.12	12.18	0.97	0.56	0.23
CBC	-6.27	7.78	9.99	0.96	0.87	0.29
			Coefficient of kids02			
FE	33.33	13.14	35.82	1.03	0.30	0.53
SBC	6.51	16.66	17.87	1.03	0.92	0.67
ABC	5.61	10.91	12.26	1.03	0.92	0.44
CBC	4.24	15.68	16.23	1.02	0.93	0.63
			Coefficient of kids35			
FE	46.31	23.40	51.88	1.02	0.52	0.93
SBC	21.65	31.91	38.54	0.99	0.89	1.24
ABC	9.25	19.49	21.55	1.01	0.93	0.77
CBC	18.62	30.67	35.85	1.00	0.89	1.20
			Coefficient of kids617			
FE	55.04	76.72	94.36	0.97	0.88	2.91
SBC	37.70	100.86	107.58	1.00	0.94	3.94
ABC	15.24	64.46	66.17	0.95	0.94	2.41
CBC	31.25	100.24	104.90	0.99	0.94	3.88
			Coefficient of log husband income			
FE	26.99	26.83	38.04	1.02	0.84	1.08
SBC	6.00	31.17	31.71	1.00	0.94	1.22
ABC	6.15	23.30	24.07	1.02	0.95	0.93
CBC	6.20	30.90	31.48	0.99	0.94	1.19
			Coefficient of age			
FE	28.02	32.72	43.05	1.04	0.87	1.34
SBC	-2.54	38.88	38.93	1.05	0.96	1.60
ABC	3.04	28.33	28.47	1.03	0.94	1.14
CBC	0.95	38.35	38.32	1.05	0.96	1.57
			Coefficient of age2			
FE	31.76	25.88	40.95	1.08	0.82	1.09
SBC	0.58	35.49	35.46	1.06	0.95	1.48
ABC	5.30	21.37	22.00	1.06	0.96	0.89
CBC	3.61	36.04	36.18	1.04	0.95	1.47

[1] The BSE denotes nonparametric bootstrap standard error with 200 simulations. FE denotes fixed effect estimators, SBC denotes split–sample bias correction (Fernández-Val and Weidner, 2016), ABC denotes analytical bias correction (Hahn and Kuersteiner, 2011), CBC denotes crossover jackknife.

Table D.8: Calibrated Dynamic Linear (NB), N = 147, T = 19

	Bias	Std Dev	RMSE	BSE/SD	p.95 (BSE)	Length (BSE)
			Coefficient of dem			
FE	-6.83	23.93	24.86	1.02	0.93	0.95
AB	-13.23	75.86	76.93	1.18	0.97	3.50
SBC	-2.43	30.77	30.83	1.02	0.93	1.23
ABC	-10.06	24.31	26.29	1.00	0.92	0.95
CBC	-1.65	29.68	29.69	1.00	0.94	1.17
			Coefficient of l1lgdp			
FE	-5.17	1.69	5.44	1.00	0.13	0.07
AB	-11.61	4.02	12.28	1.39	0.44	0.22
SBC	3.01	1.93	3.57	1.02	0.69	0.08
ABC	-1.27	2.14	2.49	1.03	0.92	0.09
CBC	0.91	1.90	2.10	1.02	0.92	0.08
			Coefficient of l2lgdp			
FE	-14.19	25.43	29.10	0.93	0.88	0.92
AB	-35.82	33.79	49.22	1.58	0.98	2.09
SBC	3.69	27.25	27.47	0.95	0.93	1.02
ABC	-5.37	26.37	26.88	0.93	0.92	0.97
CBC	-0.97	27.32	27.31	0.95	0.94	1.01
			Coefficient of l3lgdp			
FE	-19.94	41.02	45.58	0.94	0.89	1.51
AB	-31.60	54.38	62.85	1.58	0.99	3.37
SBC	-5.17	44.14	44.40	0.97	0.94	1.68
ABC	-13.07	43.37	45.25	0.95	0.92	1.61
CBC	-5.69	45.18	45.49	0.94	0.93	1.66
			Coefficient of l4lgdp			
FE	12.42	20.89	24.28	1.02	0.90	0.83
AB	23.60	29.80	38.00	1.67	0.99	1.95
SBC	4.58	26.24	26.61	1.02	0.95	1.05
ABC	25.98	22.24	34.18	1.04	0.81	0.91
CBC	3.23	24.72	24.90	1.00	0.94	0.97
			Coefficient of long run effect			
FE	-29.88	18.47	35.12	1.02	0.64	0.74
AB	-49.05	46.13	67.30	0.97	0.77	1.75
SBC	-4.82	26.81	27.21	1.02	0.92	1.08
ABC	-23.87	23.04	33.16	1.00	0.79	0.90
CBC	-9.15	25.61	27.17	1.01	0.91	1.02

[1] The BSE denotes nonparametric bootstrap standard error with 200 simulations. FE denotes fixed effect estimators, AB denotes Arellano–Bond, SBC denotes split–sample bias correction (Fernández-Val and Weidner, 2016), ABC denotes analytical bias correction (Hahn and Kuersteiner, 2011), CBC denotes crossover jackknife.

D.3 Tables of Chapter Three

Table D.9: p-Values of the Score Test

	p-value
Discretized	0.0002
Not discretized	0.6262

Table D.10: Estimated Values of δ Under H_0 (First Application)

	$\hat{\delta}^{pres}_{LCC}$	$\hat{\delta}^{size}_{LCC}$	$\hat{\delta}^{cons}_{LCC}$	$\hat{\delta}^{pres}_{OA}$	$\hat{\delta}^{size}_{OA}$	$\hat{\delta}^{cons}_{OA}$
Discretized	1.643	0.795	-2.084	0.388	0.440	0.338
Not discretized	7.102	0.453	-4.111	4.690	1.224	-2.656

Notes: $X^{(\ell)}_{i,size}$ and $X^{(\ell)}_{i,pres}$ are treated as continuous variables on the unit interval when they are not discretized; $X^{(\ell)}_{i,size}$ and $X^{(\ell)}_{i,pres}$ are binary indicators of whether the original variables are above their median or not when they are discretized.

Table D.11: Estimated Values of δ Under H_0 (Second Application)

$\widehat{\alpha}$	$\widehat{\eta}_{\text{motheduc}}$	$\widehat{\eta}_{\text{fatheduc}}$	$\widehat{\eta}_{\text{lfaminc}}$	$\widehat{\gamma}_{\text{parcath}}$	$\widehat{\gamma}_{\text{motheduc}}$	$\widehat{\gamma}_{\text{fatheduc}}$	$\widehat{\gamma}_{\text{lfaminc}}$
0.630	0.029	0.062	0.028	1.220	0.003	0.082	-0.319

Notes: Constants are not included in either equations (2.5) or (2.6). In the table, motheduc, fatheduc, lfaminc and parcath respectively denote mother's years of education, father's years of education, log family income and whether one of the parents is reported being Catholic. The estimation was conducted using the *glmfit* in MATLAB.

Table D.12: Size of the Score Test

Sample size	2500	5000	7500
Size	0.065	0.057	0.048

Table D.13: The Upper and Lower Probability Bounds in the Entry Game with Nuisance Parameters

Event A	$\nu_\theta(A) = \min P(A)$	$\nu^*_\theta(A) = \max P(A)$
$\{(0,\,0)\}$	$(1-\Phi_1)(1-\Phi_2)$	$(1-\Phi_1)(1-\Phi_2)$
$\{(1,\,1)\}$	$\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})$	$\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})$
$\{(1,\,0)\}$	$(1-\Phi_2)\Phi_1+\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})[\Phi_2-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]$	$(1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}))\Phi_1$
$\{(0,\,1)\}$	$(1-\Phi_1)\Phi_2+\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})[\Phi_1-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})]$	$(1-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)}))\Phi_2$

Appendix E

Computation Details of Chapter One

E.1 Computation Appendix

E.1.1 Calibration Procedures

Simulation procedures for the labor force participation application.

1. Run the regression on the LFP data to obtain $\widetilde{\theta}$ and $\widetilde{\alpha}_i$'s. These are treated as true coefficients for the calibration exercise.

2. For each simulation $s = 1, \ldots, S$, create a synthetic panel data based on the equation

$$y_{it}^s = \mathbf{1}\{X_{it}\widetilde{\theta} + \widetilde{\alpha}_i > u_{it}^s\},$$

   where $u_{it}^s \sim iid\mathcal{N}(0,1)$. The data $\{(y_{it}^s, X_{it})\}$ are considered as the observed data for simulation s.

3. **Implementing the estimation:**

   (a) Run Probit regression on $\{(y_{it}^s, X_{it})\}$ and obtain $\widehat{\theta}^s$ and $\widehat{\alpha}_i^s$. This denotes the fixed effect estimators using observed data.

   (b) **Data simulation**:

      i. Given a set of parameter θ, simulate dependent variable using

$$y_{it}^h(\theta) = \mathbf{1}\{X_{it}'\theta + \widehat{\alpha}_i^s > \varepsilon_{it}^h\}, \quad \varepsilon_{it}^h \sim iid\mathcal{N}(0,1)$$

Run Probit regression on $\{y_{it}^h(\theta), X_{it}\}$to obtain $\widehat{\beta}^h(\theta)$.

ii. Repeat step (i) for $H = 10$ times and compute

$$\widehat{\beta}^H(\theta) = \frac{1}{H}\sum_{h=1}^{H}\widehat{\beta}^h(\theta).$$

iii. Compute the indirect inference estimator $\widetilde{\theta}^H$ by solving the following equation

$$\widehat{\theta}^s = \widehat{\beta}^H(\widetilde{\theta}^H).$$

4. Repeat steps 2 and 3 for $S = 1000$ times.

E.1.2 Simulations for Dynamic Labor Force Participation

This subsection introduces dynamics into the specification and compare the performance of indirect fixed effect estimators with other estimators.

Positive serial correlation observed in employment outcomes motivates the question of identifying state dependence, i.e., the causal impact of past employment on future employment for married women. However, the positive correlation can also be driven by individual–specific unobserved heterogeneity such as willingness to work. Therefore, an important question of interest is to distinguish between state dependence and persistent unobserved heterogeneity.

Following the empirical specification in Fernández-Val (2009), this paper controls for time–invariant unobserved heterogeneity by adding individual fixed effects,

$$y_{it} = \mathbf{1}\{X_{it}'\theta + \alpha_i \geq u_{it}\}, \quad u_{it} \sim \mathcal{N}(0,1), \tag{E.1.1}$$

where the vector of pre–determined covariates $X_{it} := (x_{it}, y_{i,t-1})$ now contains an extra variable: $y_{i,t-1}$, which denotes the lagged dependent variable. The first year of the sample is excluded for use as the initial condition in the dynamic model. In the data

simulation step, the dependent variable at time t has the following representation:

$$y_{it}^h(\theta, \widehat{\alpha}_i) = \mathbf{1}\{\theta_1 y_{i,t-1}^h(\theta, \widehat{\alpha}_i) + x_{it}'\theta_{-1} + \widehat{\alpha}_i \geq u_{it}^h\}, \quad u_{it}^h \sim \mathcal{N}(0,1). \tag{E.1.2}$$

where θ_{-1} denotes parameters other than the one for $y_{i,t-1}^h$.

Table (D.5) reports the coefficients estimates using different methods. The analytical bias correction (ABC) corresponds to the method proposed by Hahn and Kuersteiner (2011) and serves as a benchmark. The JBC method by Hahn and Newey (2004) is no longer applicable due to dynamics in the specification. The results are similar to the static case. When $H = 20$, the indirect inference estimator produces bias correction results close to the ABC. On the other hand, the SBC estimate of lagged LFP is larger. Regarding the standard errors, the indirect fixed effect estimator does not inflate the errors when $H = 20$, but SBC has larger standard errors across all variables.

Table (D.6) reports the results of the Monte Carlo simulations. Compared to the static case in Table (D.2), adding dynamics into the regression further deteriorates fixed effect estimators of strictly exogenous covariates, which are comparable with the standard deviations. On the other hand, indirect fixed effect estimators correct the bias significantly. Compared to SBC, the reduction of bias is comparable but the standard deviation is smaller, which is consistent with the theory: by construction SBC does not use the whole sample for bias correction and thus inflates the variance.

Appendix F

More Details of Chapter Two

F.1 Calibration Procedures

This section summarizes the main steps of the calibrated Monte Carlo simulations.

1. Run the panel Probit or dynamic linear regression using real panel data;
2. Simulate $\{y_{it}\}$ using coefficients and the parametric specification in step one;
3. Construct simulated panel data using synthetic $\{y_{it}\}$ and real covariates;
4. Use simulated panel data to run panel Probit or dynamic linear regression. Compute uncorrected FE, ABC, SBC and CBC. Compute bootstrap standard errors;
5. Repeat steps two–four 500 times;
6. Compute diagnostic statistics.

Appendix G

More Details of Chapter Three

G.1 Details on the Examples

G.1.1 Discrete Games of Complete Information

We focus on Example 1 below, but the analysis of Example 2 is similar.

Model restrictions and Assumption 1

The upper and lower probabilities of all singleton events are tabulated in Table D.13. In this example, they constitute the sharp identifying restrictions (Galichon and Henry, 2011).

As argued in Section 3.3.2, the model's prediction reduces to (3.32) when $\beta^{(1)} = \beta^{(2)} = 0$, which implies a unique density in (3.30). Therefore, Assumption 1 (i) holds. For Assumption 1 (ii), it suffices to show that $\mathcal{Q}_{\theta_0}$ and $\mathcal{Q}_{\theta_1}$ are disjoint. For this, consider the event $\{(1,1)\}$. Table D.13 suggests

$$\nu_{\theta_0}(\{(1,1)\}|x) = \Phi(x^{(1)\prime}\delta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}), \tag{G.1.1}$$

whereas

$$\nu^*_{\theta_1}(\{(1,1)\}|x) = \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}). \tag{G.1.2}$$

This means $\nu^*_{\theta_1}(\{(1,1)\}|x) < \nu_{\theta_0}(\{(1,1)\}|x)$ whenever $\beta^{(j)} < 0, j = 1,2$. Hence, $\mathcal{Q}_{\theta_0}$ and $\mathcal{Q}_{\theta_1}$ are disjoint.

Computing the LFP

The LF density q_θ is given by

$$q_\theta(0,0|x) = (1-\Phi_1)(1-\Phi_2) \tag{G.1.3}$$

$$q_\theta(1,1|x) = \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}) \tag{G.1.4}$$

$$q_\theta(1,0|x) = \begin{cases} \Phi_1(1-\Phi_2) + \frac{\Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}{\Phi_1+\Phi_2-2\Phi_1\Phi_2} & \theta \in \Theta_1(x) \\ \Phi_1(1-\Phi_2) + \Phi(x^{(1)\prime}\delta_1 + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)})] & \theta \in \Theta_2(x) \cdot \\ \Phi_1(1-\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})) & \theta \in \Theta_3(x) \end{cases} \tag{G.1.5}$$

The parameter subsets, $\Theta_j(x), j = 1,2,3$, are given by

$$\Theta_1(x) = \big\{\theta : \Phi_1(1-\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})) \geq \frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2}, \tag{G.1.6}$$

$$\frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2} \geq \Phi_1(1-\Phi_2) + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]\big\}, \tag{G.1.7}$$

$$\Theta_2(x) = \big\{\theta : \frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2} < \Phi_1(1-\Phi_2) + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]\big\} \tag{G.1.8}$$

$$\Theta_3(x) = \big\{\theta : \Phi_1(1-\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})) < \frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2}\big\}, \tag{G.1.9}$$

where

$$z_1 = \Phi_1(1-\Phi_2)$$

$$z_2 = \Phi_2(1-\Phi_1) + \Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}).$$

Below, we outline how to obtain this density from the convex program in (3.22)-(3.24).

As discussed in the text, q_{θ_0} is determined by the four equality restrictions (3.26)-

(3.29). Therefore, it remains to solve the convex program in (3.22)-(3.24) for q_1. For this, we can reduce the number of control variables. First, Table D.13 implies

$$q_{\theta_1}(0,0|x) = (1-\Phi_1)(1-\Phi_2) \tag{G.1.10}$$

$$q_{\theta_1}(1,1|x) = \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}). \tag{G.1.11}$$

Hence, the remaining free components of q_1 are $q_1(1,0|x)$ and $q_1(0,1|x)$. Let $\omega = q_1(1,0|x)$. We may then express the other component as

$$\begin{aligned} q_1(0,1|x) &= 1 - q_1(0,0|x) - q_1(1,1|x) - \omega \\ &= \Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}) - \omega. \end{aligned}$$

Hence, to solve (3.22)-(3.24), it suffices to choose $\omega = q_1(1,0|x)$ optimally in the following problem:

$$\min_{\omega\in[0,1]} -\ln\Big(\frac{z_1}{z_1+\omega}\Big)(z_1+\omega) - \ln\Big(\frac{(1-\Phi_1)\Phi_2}{z_2-\omega}\Big)(z_2-\omega) \tag{G.1.12}$$

$$s.t. \omega - (1 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}))\Phi_1 \leq 0 \tag{G.1.13}$$

$$(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})] - \omega \leq 0. \tag{G.1.14}$$

Let the Lagrangian be

$$\begin{aligned} \mathcal{L}(\omega,\lambda) = &-\ln\Big(\frac{z_1}{z_1+\omega}\Big)(z_1+\omega) - \ln\Big(\frac{(1-\Phi_1)\Phi_2}{z_2-\omega}\Big)(z_2-\omega) \\ &- \lambda_1((1 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}))\Phi_1 - \omega) \\ &- \lambda_2(\omega - (1-\Phi_2)\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]). \end{aligned}$$

The Karush-Kuhn-Tucker (KKT) conditions are

$$-\ln\left(\frac{z_1}{z_1+\omega}\right)+\ln\left(\frac{\Phi_2(1-\Phi_1)}{z_2-\omega}\right)+\lambda_1-\lambda_2=0 \tag{G.1.15}$$

$$\lambda_1\Big(\Phi_1(1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}))-\omega\Big)\geq 0 \tag{G.1.16}$$

$$\lambda_2\Big(\omega-(1-\Phi_2)\Phi_1-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})[\Phi_2-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]\Big)\geq 0 \tag{G.1.17}$$

$$\lambda_1,\lambda_2\geq 0. \tag{G.1.18}$$

Below, we consider three subcases depending on the value of the Lagrange multipliers.

Case 1 ($\lambda_1=\lambda_2=0$): The FOC in (G.1.15) with $\lambda_1=\lambda_2=0$ identifies the solution $q_{\theta_1}(1,0|x)$ as follows:

$$\begin{aligned}\omega=q_{\theta_1}(1,0|x)&=\frac{z_1z_2-\Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2}\\&=\frac{\Phi_1(1-\Phi_2)[\Phi_1+\Phi_2-\Phi_1\Phi_2-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]}{\Phi_2+\Phi_1-2\Phi_1\Phi_2}.\end{aligned} \tag{G.1.19}$$

This implies

$$q_{\theta_1}(0,1|x)=\Phi_1+\Phi_2-\Phi_1\Phi_2-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})-\omega \tag{G.1.20}$$

$$=\frac{\Phi_2(1-\Phi_1)[\Phi_1+\Phi_2-\Phi_1\Phi_2-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]}{\Phi_2+\Phi_1-2\Phi_1\Phi_2}. \tag{G.1.21}$$

Substituting the value of ω into its bounds, we obtain the following restrictions:

$$\Phi_1(1-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)}))-\frac{z_1z_2-\Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2}\geq 0 \tag{G.1.22}$$

$$\frac{z_1z_2-\Phi_2(1-\Phi_1)z_1}{\Phi_2+\Phi_1-2\Phi_1\Phi_2}-\Phi_1(1-\Phi_2)-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})[\Phi_2-\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})]\geq 0. \tag{G.1.23}$$

We let $\Theta_1(x)$ denote the set of parameter values that satisfy (G.1.22)-(G.1.23).

Case 2 ($\lambda_1 = 0$, $\lambda_2 > 0$): By $\lambda_2 > 0$ and (G.1.17), we obtain

$$\omega = q_{\theta_1}(1,0|x) = (1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)})],$$

and $q_{\theta_1}(0,1|x) = (1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)}))\Phi_2$. Note that $\lambda_2 > 0$ iff

$$\frac{z_1}{z_1+\omega} < \frac{\Phi_2(1-\Phi_1)}{z_2-\omega},$$

which is equivalent to

$$(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)})] > \frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2}. \tag{G.1.24}$$

We let $\Theta_2(x)$ denote the set of parameter values that satisfy (G.1.24).

Case 3 ($\lambda_1 > 0$, $\lambda_2 = 0$): By $\lambda_1 > 0$ and (G.1.16), we obtain

$$\omega = q_{\theta_1}(1,0|x) = (1 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)}))\Phi_1,$$

and hence $q_{\theta_1}(0,1|x) = (1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta_1 + \beta^{(1)})]$. Note that $\lambda_1 > 0$ iff

$$\frac{z_1}{z_1+\omega} > \frac{\Phi_2(1-\Phi_1)}{z_2-\omega},$$

which is equivalent to

$$(1 - \Phi(x^{(2)\prime}\delta_2 + \beta^{(2)}))\Phi_1 < \frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2}. \tag{G.1.25}$$

We let $\Theta_3(x)$ denote the set of parameter values that satisfy (G.1.25).

Score:

We let $s_\theta = (s_{\beta^{(1)}}, s_{\beta^{(2)}}, s_{\delta^{(1)}}, s_{\delta^{(2)}})'$. Each component of s_θ takes the following form:

$$s_\vartheta(y|x) = \sum_{\bar{y}\in\mathcal{Y}} 1\{y = \bar{y}\} z_\vartheta(\bar{y}|x),\ \vartheta \in \{\beta^{(1)}, \beta^{(2)}, \delta^{(1)}, \delta^{(2)}\}, \tag{G.1.26}$$

where $z_\vartheta(\bar{y}|x)$ is the partial derivative of $\ln p_\theta(\bar{y}|x)$ with respect to ϑ, which is well-defined if θ is in $\Theta_2(x)$, $\Theta_3(x)$, or in the interior of $\Theta_1(x)$. Let

$$r_h(y,x) \equiv (\sqrt{q_{\theta+h}(y|x)} - \sqrt{q_\theta(y|x)} - \frac{1}{2}h' s_\theta(y|x)\sqrt{q_\theta(y|x)})^2. \tag{G.1.27}$$

Suppose $\theta \in \Theta_2(x)$. By (G.1.8), $\theta + h \in \Theta_2(x)$ for $\|h\|$ small enough. Then, pointwise, $r_h(y,x) = o(\|h^2\|)$ because $s_\theta(y|x) = 2\frac{1}{\sqrt{q_\theta(y|x)}}\frac{\partial}{\partial\theta}\sqrt{q_\theta(y|x)} = \frac{\partial}{\partial\theta}\ln q_\theta(y|x)$. The same argument applies when $\theta \in \Theta_3(x)$ or $\theta \in \text{int}(\Theta_1(x))$. The only case this argument does not apply is when θ is on the boundary between $\Theta_2(x)$ and $\Theta_1(x)$ (or between $\Theta_3(x)$ and $\Theta_1(x)$). For example, suppose θ is on the boundary between $\Theta_2(x)$ and $\Theta_1(x)$. Then, we may have $\theta + h \in \Theta_2(x)$ for all h with $\|h\| > 0$ but $\theta \in \Theta_1$. Then, the pointwise argument above does not apply. However, θ being on the boundary between the two sets means

$$\frac{z_1 z_2 - \Phi_2(1-\Phi_1)z_1}{\Phi_2 + \Phi_1 - 2\Phi_1\Phi_2} = \Phi_1(1-\Phi_2) + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})].$$

If x contains a continuous component (e.g. distance from headquarters/distribution center), the set of x's satisfying above has measure 0, $r_h(y,x)$ is bounded on the set, and hence it does not affect the integral in Assumption 2. Hence, Assumption 2 holds.

For completeness, the functional form of $z_\vartheta(\bar{y}|x)$ is derived below for each $\bar{y} \in \mathcal{Y}$ and $\vartheta \in \{\beta^{(1)}, \beta^{(2)}, \delta^{(1)}, \delta^{(2)}\}$. Across all subcases analyzed in the previous section, the form of $q_\theta(0,0|x)$ and $q_\theta(1,1|x)$ remains the same. We calculate score functions first

by taking the pointwise derivative of $\ln q_\theta(0,0|x)$ and $\ln q_\theta(1,1|x)$. This yields

$$
\begin{aligned}
&z_{\beta^{(1)}}(0,0|x) = 0, \quad z_{\beta^{(2)}}(0,0|x) = 0, \\
&z_{\delta^{(1)}}(0,0|x) = -\frac{\phi_1 x^{(1)\prime}}{(1-\Phi_1)}, \quad z_{\delta^{(2)}}(0,0|x) = -\frac{\phi_2 x^{(2)\prime}}{(1-\Phi_2)} \\
&z_{\beta^{(1)}}(1,1|x) = \frac{\phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})}{\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})}, \quad z_{\beta^{(2)}}(1,1|x) = \frac{\phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}{\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})} \\
&z_{\delta^{(1)}}(1,1|x) = \frac{\phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})x^{(1)}}{\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})}, \quad z_{\delta^{(2)}}(1,1|x) = \frac{\phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})x^{(2)}}{\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})},
\end{aligned}
$$

where $\phi_j = \phi(x^{(j)\prime}\delta^{(j)}), j = 1,2$. Next, we derive $z_\vartheta(1,0|x)$ and $z_\vartheta(0,1|x)$.

Case 1: Suppose $\theta \in \Theta_1(x)$. By taking the pointwise derivative of $\ln q_\theta$ in (G.1.5), one can obtain

$$
\begin{aligned}
z_{\delta^{(1)}}(1,0|x) = &\frac{\phi_1 x^{(1)}}{\Phi_1} + \frac{\phi_1 x^{(1)} - \phi_1\Phi_2 x^{(1)} - \phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})x^{(1)}}{\Phi_1+\Phi_2-\Phi_1\Phi_2-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})} \\
&- \frac{\phi_1 x^{(1)}(1-2\Phi_2)}{\Phi_1+\Phi_2-2\Phi_1\Phi_2}
\end{aligned}
$$

$$
\begin{aligned}
z_{\delta^{(2)}}(1,0|x) = &\frac{-\phi_2 x^{(2)}}{1-\Phi_2} + \frac{\phi_2 x^{(2)} - \phi_2\Phi_1 x^{(2)} - \phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})x^{(2)}}{\Phi_1+\Phi_2-\Phi_1\Phi_2-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})} \\
&- \frac{\phi_2 x^{(2)}(1-2\Phi_1)}{\Phi_1+\Phi_2-2\Phi_1\Phi_2}
\end{aligned}
$$

$$
z_{\beta^{(1)}}(1,0|x) = \frac{-\phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}{\Phi_1+\Phi_2-\Phi_1\Phi_2-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}
$$

$$
z_{\beta^{(2)}}(1,0|x) = \frac{-\phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})}{\Phi_1+\Phi_2-\Phi_1\Phi_2-\Phi(x^{(1)\prime}\delta^{(1)}+\beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)}+\beta^{(2)})}.
$$

Similarly,

$$z_{\delta^{(1)}}(0,1|x) = \frac{-\phi_1 x^{(1)}}{1-\Phi_1} + \frac{\phi_1 x^{(1)} - \phi_1\Phi_2 x^{(1)} - \phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})x^{(1)}}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})} - \frac{\phi_1 x^{(1)}(1-2\Phi_2)}{\Phi_1 + \Phi_2 - 2\Phi_1\Phi_2}$$

$$z_{\delta^{(2)}}(0,1|x) = \frac{\phi_2 x^{(2)}}{\Phi_2} + \frac{\phi_2 x^{(2)} - \phi_2\Phi_1 x^{(2)} - \phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})x^{(2)}}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})} - \frac{\phi_2 x^{(2)}(1-2\Phi_1)}{\Phi_1 + \Phi_2 - 2\Phi_1\Phi_2}$$

$$z_{\beta^{(1)}}(0,1|x) = \frac{-\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}$$

$$z_{\beta^{(2)}}(0,1|x) = \frac{-\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{\Phi_1 + \Phi_2 - \Phi_1\Phi_2 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}.$$

Case 2: Suppose $\theta \in \Theta_2(x)$. Similarly to the analysis in Case 1, we may obtain

$$z_{\delta^{(1)}}(1,0|x) = \frac{x^{(1)}(1-\Phi_2)\phi_1 + x^{(1)}\Phi_2\phi(x^{(1)\prime}\delta^{(1)} + \beta_1)}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]} - \frac{x^{(1)}\Phi(x^{(2)\prime}\delta^{(2)} + \beta_2)\phi(x^{(1)\prime}\delta^{(1)} + \beta_1)}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]}$$

$$z_{\delta^{(2)}}(1,0|x) = \frac{-x^{(2)}\Phi_1\phi_2 + x^{(2)}\phi_2\Phi(x^{(1)\prime}\delta^{(1)} + \beta_1)}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]} - \frac{x^{(2)}\Phi(x^{(1)\prime}\delta^{(1)} + \beta_1)\phi(x^{(2)\prime}\delta^{(2)} + \beta_2)}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]}$$

$$z_{\beta^{(1)}}(1,0|x) = \frac{(\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)}))\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]}$$

$$z_{\beta^{(2)}}(1,0|x) = \frac{\Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}{(1-\Phi_2)\Phi_1 + \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})[\Phi_2 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})]},$$

and

$$z_{\delta^{(1)}}(0,1|x) = -\frac{x^{(1)}\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}, \quad z_{\delta^{(2)}}(0,1|x) = \frac{x^{(2)}\phi_2}{\Phi_2}$$

$$z_{\beta^{(1)}}(0,1|x) = -\frac{\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}, \quad z_{\beta^{(2)}}(0,1|x) = 0.$$

Case 3: Suppose $\theta \in \Theta_3(x)$. Similarly to the previous two cases, we may obtain

$$z_{\delta^{(1)}}(1,0|x) = x^{(1)}\phi_1/\Phi_1, \quad z_{\delta^{(2)}}(1,0|x) = \frac{-x^{(2)}\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}{1 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}$$

$$z_{\beta^{(1)}}(1,0|x) = 0, \quad z_{\beta^{(2)}}(1,0|x) = \frac{-\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}{1 - \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})},$$

and

$$z_{\delta^{(1)}}(0,1|x) = \frac{-x^{(1)}\Phi_2\phi_1 + \Phi(x^{(2)\prime} + \delta^{(2)})x^{(1)}(\phi_1 - \phi(x^{(1)\prime}\delta^{(1)} + \delta^{(1)}))}{(1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})]}$$

$$z_{\delta^{(2)}}(0,1|x) = \frac{x^{(2)}(1-\Phi_1)\phi_2 + x^{(2)}(\Phi_1 - \Phi(x^{(1)\prime}\delta_1 + \beta_1))\phi(x^{(2)\prime}\delta_2 + \beta_2)}{(1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})]}$$

$$z_{\beta^{(1)}}(0,1|x) = \frac{-\Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})\phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})}{(1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})]}$$

$$z_{\beta^{(2)}}(0,1|x) = \frac{(\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)}))\phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})}{(1-\Phi_1)\Phi_2 + \Phi(x^{(2)\prime}\delta^{(2)} + \beta^{(2)})[\Phi_1 - \Phi(x^{(1)\prime}\delta^{(1)} + \beta^{(1)})]}.$$

G.1.2 Triangular Model with an Incomplete Control Function

We simplify G in Example 3 below. Let $\mathcal{U} = \mathbb{R}$. We assume $\beta > 0$ throughout, but a similar analysis can be done by assuming $\beta < 0$. Suppose $d_i = 1$ first. By (3.6)-(3.7), $y_i = 0$ if

$$u_i < -\alpha - w_i'\eta - \beta v_i, \quad \text{for some } v_i \in [-z_i'\gamma, \infty). \tag{G.1.28}$$

Then, by $\beta > 0$, we can write this event as $u_i \in \bigcup_{v \in [-z_i'\gamma, \infty)} (-\infty, -\alpha - w_i'\eta - \beta v) = (-\infty, -\alpha - w_i'\eta + \beta z_i'\gamma)$. By (3.6)-(3.7) again, $y_i = 1$ if

$$u_i \geq -\alpha - w_i'\eta - \beta v_i, \quad \text{for some } v_i \in [-z_i'\gamma, \infty). \tag{G.1.29}$$

This means that $y_i = 1$ is consistent with the model whenever $u_i \in \bigcup_{v \in [-z_i'\gamma, \infty)} [-\alpha - w_i'\eta - \beta v, \infty) = \mathbb{R}$. These predictions can be summarized as

$$G(u_i|1, w_i, z_i; \theta) = \begin{cases} \{1\} & u_i \geq -\alpha - w_i'\eta + \beta z_i'\gamma \\ \{0, 1\} & u_i < -\alpha - w_i'\eta + \beta z_i'\gamma. \end{cases} \tag{G.1.30}$$

Now suppose $d_i = 0$ implying $v_i \in (-\infty, -z_i'\gamma)$. Repeating a similar analysis yields the following correspondence

$$G(u_i|0, w_i, z_i; \theta) = \begin{cases} \{0\} & u_i \leq -w_i'\eta + \beta z_i'\gamma \\ \{0, 1\} & u_i > -w_i'\eta + \beta z_i'\gamma. \end{cases} \tag{G.1.31}$$

These predictions are summarized in Figure C·10.

Model restrictions and Assumption 1

Assumption 1 (i) holds because, as argued in Section 3.3.2, the model's prediction under the null hypothesis is complete and is characterized by the reduced form function:

$$g(u_i|d_i, w_i, z_i; \theta) = 1\{\alpha d_i + w_i'\eta + u_i \geq 0\}. \tag{G.1.32}$$

This structure induces a unique conditional density for y_i. Suppose $u_i \sim N(0, 1)$.[1] Assumption 1 (ii) holds as long as z_i's support is rich enough so that $z_i'\gamma < 0$ with

[1] Here, we normalize the scale by setting the variance of u_i to 1. Other choices of normalization are also possible.

positive probability. For this, we demonstrate that there exists an event A such that $\nu_{\theta_1}(A|x) > \nu^*_{\theta_0}(A|x)$ for some value of $x = (d, w, z)$. For this, take $A = \{1\}$ and suppose $d_i = 0$. Under the null hypothesis, the conditional probability of $y_i = 1$ is uniquely determined as $\nu^*_{\theta_0}(1|d_i = 0, w_i, z_i) = \Phi(w_i'\eta)$. When $\beta > 0$, (G.1.30) implies

$$\nu_{\theta_1}(\{1\}|d_i = 0, w_i, z_i) = \Phi(w_i'\eta - \beta z_i'\gamma), \tag{G.1.33}$$

which is greater than $\nu^*_{\theta_0}(1|d_i = 0, w_i, z_i)$ for values of z_i such that $z_i'\gamma < 0$. Hence, $\mathfrak{q}_{\theta_0}$ and $\mathfrak{q}_{\theta_1}$ are disjoint.

Computing the LFP and score

For any $\theta = (\beta, \delta)$ with $\beta > 0$, the set $\mathfrak{q}_\theta$ of densities compatible with θ is then characterized by the following inequalities

$$q(0|d = 0, w, z) \geq \Phi(-w_i'\eta + \beta z_i'\gamma) \tag{G.1.34}$$

$$q(1|d = 0, w, z) \geq 0, \tag{G.1.35}$$

and

$$q(0|d = 1, w, z) \geq 0 \tag{G.1.36}$$

$$q(1|d = 1, w, z) \geq 1 - \Phi(-\alpha - w_i'\eta + \beta z_i'\gamma). \tag{G.1.37}$$

Suppose $d = 0$. Let $z = q_1(0|d = 0, w, z)$. Then, the convex program in (3.22)-

(3.24) can be written as

$$\min_{(q_0,q_1)} \ln\Big(\frac{q_0(0|x)+z}{q_0(0|x)}\Big)(q_0(0|x)+z) + \Big(\frac{1-q_0(0|x)+1-z}{1-q_0(0|x)}\Big)(q_0(1|x)+1-z) \tag{G.1.38}$$

$$s.t.\ q_0(y|x) = 1-\Phi(w'\eta) \tag{G.1.39}$$

$$z \geq \Phi(-w'\eta+\beta z'\gamma) \tag{G.1.40}$$

$$1-z \geq 0, \tag{G.1.41}$$

where (G.1.39) is due to the completeness of the model under the null hypothesis (see (3.47)) and $d=0$. Note that (G.1.41) is redundant since q_1 being in the probability simplex is implicitly assumed. The KKT conditions associated with the program is therefore

$$\ln\frac{q_0(0|x)+z}{q_0(0|x)} - \ln\frac{2-q_0(0|x)-z}{1-q_0(0|x)} - \lambda = 0 \tag{G.1.42}$$

$$\lambda(\Phi(-w'\eta+\beta z'\gamma)-z) \tag{G.1.43}$$

$$\lambda \geq 0 \tag{G.1.44}$$

where $q_0(y|x) = 1-\Phi(w'\eta)$. There are two cases to consider.

Case 1 ($\lambda = 0$): When $\lambda = 0$, (G.1.42) implies $z = q_0(y|x) = 1-\Phi(w'\eta)$. This holds when $z = 1-\Phi(w'\eta) \geq \Phi(-w'\eta+\beta z'\gamma)$.

Case 2 ($\lambda > 0$): When $\lambda > 0$, $z = \Phi(-w'\eta+\beta z'\gamma)$ by (G.1.40). This occurs when

$$\lambda = \ln\frac{q_0(0|x)+z}{q_0(0|x)} - \ln\frac{2-q_0(0|x)-z}{1-q_0(0|x)} > 0, \tag{G.1.45}$$

which is equivalent to

$$\Phi(-w'\eta+\beta z'\gamma) > q_0(y|x) = 1-\Phi(w'\eta). \tag{G.1.46}$$

In sum, we have

$$q_{\theta_1}(0|d=0,w,z) = \begin{cases} 1-\Phi(w'\eta) & \text{if } \Phi(-w'\eta+\beta z'\gamma) \leq 1-\Phi(w'\eta), \\ \Phi(-w'\eta+\beta z'\gamma) & \text{if } \Phi(-w'\eta+\beta z'\gamma) > 1-\Phi(w'\eta), \end{cases} \tag{G.1.47}$$

and $q_{\theta_1}(1|d=0,w,z) = 1 - q_{\theta_1}(0|d=0,w,z)$. Repeating a similar analysis for $d=1$ yields

$$\begin{aligned} &q_{\theta_1}(1|d=1,w,z) \\ &= \begin{cases} \Phi(\alpha+w'\eta) & \text{if } 1-\Phi(-\alpha-w'\eta+\beta z'\gamma) \leq \Phi(\alpha+w'\eta), \\ 1-\Phi(-\alpha-w'\eta+\beta z'\gamma) & \text{if } 1-\Phi(-\alpha-w'\eta+\beta z'\gamma) > \Phi(\alpha+w'\eta), \end{cases} \end{aligned} \tag{G.1.48}$$

and $q_{\theta_1}(0|d=1,w,z) = 1 - q_{\theta_1}(1|d=1,w,z)$.

Recalling $\beta > 0$ and Φ is strictly increasing, we may summarize (G.1.47)-(G.1.48) as follows

$$q_{\theta_1}(0|d=0,w,z) = \begin{cases} \Phi(-w'\eta) & \text{if } z'\gamma \leq 0, \\ \Phi(-w'\eta+\beta z'\gamma) & \text{if } z'\gamma > 0, \end{cases} \tag{G.1.49}$$

$$q_{\theta_1}(1|d=0,w,z) = \begin{cases} 1-\Phi(-w'\eta) & \text{if } z'\gamma \leq 0, \\ 1-\Phi(-w'\eta+\beta z'\gamma) & \text{if } z'\gamma > 0, \end{cases} \tag{G.1.50}$$

and

$$q_{\theta_1}(0|d=1,w,z) = \begin{cases} \Phi(-\alpha - w'\eta) & \text{if } z'\gamma \geq 0, \\ \Phi(-\alpha - w'\eta + \beta z'\gamma) & \text{if } z'\gamma < 0, \end{cases} \tag{G.1.51}$$

$$q_{\theta_1}(1|d=1,w,z) = \begin{cases} 1 - \Phi(-\alpha - w'\eta) & \text{if } z'\gamma \geq 0, \\ 1 - \Phi(-\alpha - w'\eta + \beta z'\gamma) & \text{if } z'\gamma < 0. \end{cases} \tag{G.1.52}$$

The corresponding score function with respect to β is

$$s_\beta(0|d=0,w,z) = \begin{cases} 0 & \text{if } z'\gamma \leq 0, \\ \frac{\phi(-w'\eta+\beta z'\gamma)}{\Phi(-w'\eta+\beta z'\gamma)} z'\gamma & \text{if } z'\gamma > 0, \end{cases} \tag{G.1.53}$$

$$s_\beta(1|d=0,w,z) = \begin{cases} 0 & \text{if } z'\gamma \leq 0, \\ -\frac{\phi(-w'\eta+\beta z'\gamma)}{\Phi(-w'\eta+\beta z'\gamma)} z'\gamma & \text{if } z'\gamma > 0, \end{cases} \tag{G.1.54}$$

and

$$s_\beta(0|d=1,w,z) = \begin{cases} 0 & \text{if } z'\gamma \geq 0, \\ \frac{\phi(-\alpha-w'\eta+\beta z'\gamma)}{1-\Phi(-\alpha-w'\eta+\beta z'\gamma)} z'\gamma & \text{if } z'\gamma < 0, \end{cases} \tag{G.1.55}$$

$$s_\beta(1|d=1,w,z) = \begin{cases} 0 & \text{if } z'\gamma \geq 0, \\ -\frac{\phi(-\alpha-w'\eta+\beta z'\gamma)}{1-\Phi(-\alpha-w'\eta+\beta z'\gamma)} z'\gamma & \text{if } z'\gamma < 0. \end{cases} \tag{G.1.56}$$

G.1.3 Panel Dynamic Discrete Choice Models

For each t, let $u_{it} = \alpha_i + \epsilon_{it}$. We explicitly derive a form of G below. Note that, $y_i = (y_{i1}, y_{i2}) = (0,0)$ occurs if

$$u_{i1} < -x'_{i1}\lambda, \; u_{i2} < -x'_{i2}\lambda, \tag{G.1.57}$$

which follows from (3.11)-(3.12) or

$$u_{i1} < -x'_{i1}\lambda - \beta,\ u_{i2} < -x'_{i2}\lambda, \tag{G.1.58}$$

which follows from (3.13)-(3.14). When $\beta \geq 0$, the union of the two events reduces to (G.1.57).

Similarly, $y = (0, 1)$ occurs if

$$u_{i1} < -x'_{i1}\lambda,\ u_{i2} \geq -x'_{i2}\lambda, \tag{G.1.59}$$

or

$$u_{i1} < -x'_{i1}\lambda - \beta,\ u_{i2} \geq -x'_{i2}\lambda. \tag{G.1.60}$$

When $\beta \geq 0$, the union of the two events reduces (G.1.59).

The outcome $y = (1, 0)$ occurs if

$$u_{i1} \geq -x'_{i1}\lambda,\ u_{i2} < -x'_{i2}\lambda - \beta, \tag{G.1.61}$$

or

$$u_{i1} \geq -x'_{i1}\lambda - \beta,\ u_{i2} < -x'_{i2}\lambda - \beta, \tag{G.1.62}$$

When $\beta \geq 0$, the union of the two events reduces to (G.1.62).

The outcome $y = (1, 1)$ occurs if

$$u_{i1} \geq -x'_{i1}\lambda,\ u_{i2} \geq -x'_{i2}\lambda - \beta, \tag{G.1.63}$$

or

$$u_{i1} \geq -x'_{i1}\lambda - \beta,\ u_{i2} \geq -x'_{i2}\lambda - \beta, \tag{G.1.64}$$

When $\beta \geq 0$, the union of the two events reduces to (G.1.64). These predictions are summarized in Figure C·11.

The correspondence can therefore be written as

$$G(u_i|x_i;\theta) = \begin{cases} \{(0,0)\} & u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} < -x_{i2}'\lambda, \\ \{(0,1)\} & u_{i1} < -x_{i1}'\lambda - \beta,\ u_{i2} \geq -x_{i2}'\lambda, \\ \{(1,0)\} & u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} < -x_{i2}'\lambda - \beta, \\ \{(1,1)\} & u_{i1} \geq -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda - \beta, \\ \{(0,0),(1,0)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ u_{i2} \leq -x_{i2}'\lambda - \beta, \\ \{(0,0),(1,1)\} & -x_{it}'\lambda - \beta \leq u_{it} < -x_{it}'\lambda,\ t \in \{1,2\}, \\ \{(0,1),(1,1)\} & -x_{i1}'\lambda - \beta \leq u_{i1} < -x_{i1}'\lambda,\ u_{i2} \geq -x_{i2}'\lambda. \end{cases} \tag{G.1.65}$$

A similar analysis can be done for the setting with $\beta \leq 0$, which we omit for brevity.

Model restrictions and Assumption 1

Assumption 1 (i) holds because, as argued in (3.34), the model makes a complete prediction with the following reduced-form function when $\beta = 0$:

$$g(u_i|x_i;\theta) = \begin{bmatrix} 1\{x_{i1}'\lambda + \alpha_i + \epsilon_{i1} \geq 0\} \\ 1\{x_{i2}'\lambda + \alpha_i + \epsilon_{i2} \geq 0\} \end{bmatrix}. \tag{G.1.66}$$

Assumption 1 (ii) holds if u follows a distribution F that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$. We show below, when $\beta > 0$, there exists an event A such that $\nu_{\theta_1}(A) > \nu^*_{\theta_0}(A)$ for all $\theta_1 \in \Theta_1$. For example, take $A = \{(1,1)\}$. As shown on the left panel of Figure C·11, the probability of $\{(1,1)\}$ is uniquely

determined when $\beta = 0$. Therefore, the upper bound on the probability of $\{(1,1)\}$ is

$$\nu^*_{\theta_0}(\{(1,1)\}|x) = F(u_{i1} \geq -x'_{i1}\lambda,\ u_{i2} \geq -x'_{i2}\lambda). \tag{G.1.67}$$

When $\beta > 0$, the lower bound on the probability of the same event is

$$\nu_{\theta_1}(\{(1,1)\}|x) = F(u_{i1} \geq -x'_{i1}\lambda,\ u_{i2} \geq -x'_{i2}\lambda - \beta), \tag{G.1.68}$$

which exceeds $\nu^*_{\theta_0}(\{(1,1)\}|x)$ as long as F is absolutely continuous. This means $\mathfrak{q}_{\theta_1}$ and $\mathfrak{q}_{\theta_0}$ are disjoint.

The analysis of the LFP and score is similar to that of discrete games. For brevity, we omit details.

G.2 Details on Monte Carlo Experiments

This section provides details on the selection mechanisms in DGP under the alternatives. Multiple equilibria exist for some values of u, so we select $s \in G(u|\beta; X, \delta)$ according to one of the three choices: (1) i.i.d., (2) non i.i.d. and (3) LFP.

G.2.1 IID Selection

The i.i.d. selection mechanism selects $(1,0)$ out of $G(u|\beta; X, \delta) = \{(1,0),(0,1)\}$ if an i.i.d. Bernoulli random variable ν_i takes one.

G.2.2 Non–Ergodic Selection

Let N^*_k be an increasing sequence of integers. More specifically, we specify the sequence to be $N^*_k = \left\{2^{2^{\left\lceil \frac{ln(\ln(k)/\ln(2))}{\ln(2)} \right\rceil}}, k = 1, 2, ...\right\}$, where $\lceil \cdot \rceil$ denotes the ceiling

function. For each i, let $h(i) = N_k^*$ where $N_{k-1}^* < i \leq N_k^*$, define

$$\tilde{\nu}_i = \begin{cases} 1 & \Psi_{h(i)}^G(u) > \Lambda_{h(i)}, \\ 0 & \Psi_{h(i)}^G(u) \leq \Lambda_{h(i)}, \end{cases}$$

where

$$\Psi_{h(i)}^G(u) = \frac{\sum_{i=1}^{h(i)} 1[G(u_i|\beta; X, \delta) = \{(1,0)\}]}{\sum_{i=1}^{h(i)} 1[G(u_i|\beta; X, \delta) = \{(1,0),(0,1)\}]}$$

is the ratio of the frequencies of $(1,0)$ being predicted as a unique PSNE relative to the frequencies of $(1,0)$ or $(0,1)$ being unique PSNE up until the last observation of cluster N_k^*. If the ratio is above threshold $\Lambda_{h(i)}$, select $(1,0)$ all the time in the region of multiplicity; otherwise select $(0,1)$. The intuition is that if $(1,0)$ is already occurring relatively more frequently (as a unique PSNE) up to N_k^*, the selection mechanism adds more $(1,0)$ from the region of multiplicity, which makes data heavily dependent on the path up to N_k^*. However, once leaving the k-th cluster, the DGP switches to another selection path, and the sequence N_k^* is constructed in such a way that the dependence doesn't vanish asymptotically. We specify $\Lambda_{h(i)}$ as follows. Conditional on X, we can compute the conditional probabilities of $(1,0)$ and $(0,1)$ being the unique equilibrium. These are the respective lower probabilities of the two events, $\nu_{\beta,\delta|X}((1,0))$ and $\nu_{\beta,\delta|X}((0,1))$. Let N_c denote the number of occurrences of X's configuration c within N_k^*, i.e., $\sum_c N_c = h(i)$, we calculate the empirical weighted sum of the two events and define $\Lambda_{h(i)}$ as follows,

$$\Lambda_{h(i)} = \frac{\sum_{c\in\{(1,1),(1,-1),(-1,1),(-1,-1)\}} N_c \nu_{\beta,\delta|c}((1,0))}{\sum_{c\in\{(1,1),(1,-1),(-1,1),(-1,-1)\}} N_c \Big(\nu_{\beta,\delta|c}((1,0)) + \nu_{\beta,\delta|c}((0,1))\Big)}.$$

G.2.3 DGP under the LFP

For each observation in the Monte Carlo experiment the LFP is represented by its own cumulative distribution function. We take a draw ν_i from uniformly distributed

$[0, 1]$ and use the inverse CDF mapping to generate each of the outcomes. We repeat this for every observation and thus get the DGP. See figure C·12 for a graphical illustration for one observation.

References

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2019). Democracy does cause growth. *Journal of Political Economy*, 127(1):47–100.

Aghion, P., Bloom, N., Blundell, R., Griffith, R., and Howitt, P. (2005). Competition and innovation: an inverted-u relationship. *The Quarterly Journal of Economics*, 120(2):701–728.

Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *Journal of Human Resources*, 40(4):791–821.

Altonji, J. G., Smith, A. A., and Vidangos, I. (2013). Modeling earnings dynamics. *Econometrica*, 81(4):1395–1454.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum–likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):283–301.

Andrews, D. W. (1994). Empirical process methods in econometrics. In Heckman, J. J. and Leamer, E., editors, *Handbook of Econometrics*, volume 4, chapter 37, pages 2247–2294. Elsevier.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67(6):1341–1383.

Andrews, D. W. K. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157.

Andrews, I., Roth, J., and Pakes, A. (2019). Inference for linear conditional moment inequalities. Working Paper.

Arcidiacono, P. and Miller, R. A. (2011). Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica*, 79(6):1823–1867.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297.

Arellano, M. and Hahn, J. (2007). Understanding bias in nonlinear panel models: Some recent developments. In Blundell, R., Newey, W., and Persson, T., editors, *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress*, chapter 12, pages 381–409. Cambridge University Press.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *Annals of Statistics*, 49(1):486–507.

Barseghyan, L., Coughlin, M., Molinari, F., and Teitelbaum, J. C. (2021). Heterogeneous choice sets and preferences. *Econometrica*, 89(5):2015–2048.

Bera, A. K. and Bilias, Y. (2001). Rao's score, neyman's $c(\alpha)$ and silvey's lm tests: An essay on historical developments and some new results. *Journal of Statistical Planning and Inference*, 97(1):9–44.

Beresteanu, A., Molchanov, I., and Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. *Econometrica*, 79(6):1785–1821.

Berger, D. and Vavra, J. (2019). Shocks versus responsiveness: What drives time-varying dispersion? *Journal of Political Economy*, 127(5):2104–2142.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica*, 60(4):889–917.

Bresnahan, T. F. and Reiss, P. C. (1991a). Empirical models of discrete games. *Journal of Econometrics*, 48(1):57–81.

Bresnahan, T. F. and Reiss, P. C. (1991b). Entry and competition in concentrated markets. *Journal of Political Economy*, 99(5):977–1009.

Bruins, M., Duffy, J. A., Keane, M. P., and Smith, A. A. (2018). Generalized indirect inference for discrete choice models. *Journal of Econometrics*, 205(3):177–203.

Bugni, F., Canay, I., and Shi, X. (2017). Inference for subvectors and other functions of partially identified parameters in moment inequality models. *Quantitative Economics*, 8(1):1–38.

Caetano, C., Callaway, B., Payne, S., and Rodrigues, H. S. (2022). Difference in differences with time-varying covariates. Working Paper.

Callaway, B. and Li, T. (2021). Policy evaluation during a pandemic. Working Paper.

Canay, I. A. and Shaikh, A. M. (2017). Practical and theoretical advances in inference for partially identified models. In Honoré, B., Pakes, A., Piazzesi, M., and Samuelson, L., editors, *Advances in Economics and Econometrics: Eleventh World Congress*, volume 2 of *Econometric Society Monographs*, pages 271–306. Cambridge University Press.

Chamberlain, G. (1984). Panel data. In Griliches, J. and Intriligator, M., editors, *Handbook of Econometrics*, volume 2, chapter 22, pages 1247–1318. Elsevier.

Chen, M., Fernández-Val, I., and Weidner, M. (2021). Nonlinear factor models for network and panel data. *Journal of Econometrics*, 220(2):296–324.

Chen, X., Christensen, T. M., and Tamer, E. (2018). Monte carlo confidence sets for identified sets. *Econometrica*, 86(6):1965–2018.

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):1094–1111.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580.

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

Chesher, A. and Rosen, A. M. (2017). Generalized instrumental variable models. *Econometrica*, 85(3):959–989.

Chudik, A., Pesaran, M. H., and Yang, J.-C. (2018). Half-panel jackknife fixed-effects estimation of linear panels with weakly exogenous regressors. *Journal of Applied Econometrics*, 33(6):1094–1111.

Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828.

Collard-Wexler, A. (2013). Demand fluctuations in the ready-mix concrete industry. *Econometrica*, 81(3):1003–1037.

Cox, G. and Shi, X. (2020). Simple adaptive size-exact testing for full-vector and subvector inference in moment inequality models. Working Paper.

de Paula, Á., Richards-Shubik, S., and Tamer, E. (2018). Identifying preferences in networks with bounded degree. *Econometrica*, 86(1):263–288.

de Paula, Á. and Tang, X. (2012). Inference of signs of interaction effects in simultaneous games with incomplete information. *Econometrica*, 80(1):143–172.

Dedecker, J. and Louhichi, S. (2002). Maximal inequalities and empirical central limit theorem. In Dehling, H., Mikosch, T., and Sørensen, M., editors, *Empirical Process Techniques for Dependent Data*, chapter 3, pages 137–160. Springer.

Dempster, A. (1967). Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38(2):325–339.

Dhaene, G. and Jochmans, K. (2015). Split–panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3):991—-1030.

Dridi, R., Guay, A., and Renault, E. (2007). Indirect inference and calibration of dynamic stochastic general equilibrium models. *Journal of Econometrics*, 136(2):397–430.

Dridi, R. and Renault, E. (2000). Semi–parametric indirect inference. Working Paper.

Duffie, D. and Singleton, K. J. (1993). Simulated moments estimation of markov models of asset prices. *Econometrica*, 61(4):929–952.

Eizenberg, A. (2014). Upstream innovation and product variety in the u.s. home pc market. *The Review of Economic Studies*, 81(3):1003–1045.

Epstein, L., Kaido, H., and Seo, K. (2016). Robust confidence regions for incomplete models. *Econometrica*, 84(5):1799–1838.

Fack, G., Grenet, J., and He, Y. (2019). Beyond truth-telling: Preference estimation with centralized school choice and college admissions. *American Economic Review*, 109(4):1486–1529.

Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, 150(1):71–85.

Fernández-Val, I. and Weidner, M. (2016). Individual and time effects in nonlinear panel models with large n, t. *Journal of Economics*, 192(1):291–312.

Fernández-Val, I. and Weidner, M. (2018). Fixed effects estimation of large–t panel data models. *Annual Review of Economics*, 10:109–138.

Forneron, J.-J. (2020). A sieve–smm estimator for dynamic models. Working Paper.

Frazier, D. T., Oka, T., and Zhu, D. (2019). Indirect inference with a non–smooth criterion function. *Journal of Econometrics*, 212(2):623–645.

Freyaldenhoven, S., Hansen, C., Pérez, J. P., and Shapiro, J. M. (2021). Visualization, identification, and estimation in the linear panel event-study design. *Advances in Economics and Econometrics, Theory and Applications, Twelfth World Congress*, forthcoming.

Galichon, A. and Henry, M. (2011). Set identification in models with multiple equilibria. *The Review of Economic Studies*, 78(4):1264–1298.

Gallant, R. A. and Tauchen, G. (1996). Which moments to match? *Econometric Theory*, 12(4):657–681.

Galvao, A. F. and Kato, K. (2018). Quantile regression methods for longitudinal data. In Koenker, R., Chernozhukov, V., He, X., and Peng, L., editors, *Handbook of Quantile Regression*, chapter 19, pages 363–380. Chapman and Hall/CRC.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153.

Glasserman, P. and Yao, D. D. (1992). Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908.

Gonçlaves, S. and Kaffo, M. (2015). Bootstrap inference for linear dynamic panel data models with individual fixed effects. *Journal of Econometrics*, 186(5):407–426.

Gospodinov, N., Komunjer, I., and Ng, S. (2017). Simulated minimum distance estimation of dynamic models with errors-in-variables. *Journal of Econometrics*, 200(2):181–193.

Gouriéroux, C. and Monfort, A. (1997). *Simulation-based Econometric Methods*. Oxford University Press.

Gouriéroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8(1):85–118.

Gouriéroux, C., Phillips, P., and Yu, J. (2010). Indirect inference for dynamic panel models. *Journal of Econometrics*, 157(1):68–77.

Guvenen, F. and Smith, A. A. (2014). Inferring labor income risk and partial insurance from economic choices. *Econometrica*, 82(6):2085–2129.

Hahn, J. and Kuersteiner, G. (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica*, 70(4):1639–1657.

Hahn, J. and Kuersteiner, G. (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory*, 27(6):1152 – 1191.

Hahn, J. and Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72(4):1295–1319.

Haile, P. A. and Tamer, E. (2003). Inference with an incomplete model of english auctions. *Journal of Political Economy*, 111(1):1–51.

Heckman, J. J. (1978). Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence. *Annales de INSEE*, 1(30/31):227–269.

Heckman, J. J. (1987). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process and some monte carlo evidence. In Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data With Econometric Applications*. MIT Press.

Helpman, E., Melitz, M., and Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics*, 123(2):441–487.

Henry, M., Meango, R., and Mourifié, I. (2020). Revealing gender-specific costs of stem in an extended roy model of major choice. Working Paper.

Honoré, B. E. and Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, 74(3):611–629.

Horowitz, J. L. (2001). The bootstrap. In Heckman, J. J. and Leamer, E., editors, *Handbook of Econometrics*, volume 5, chapter 52, pages 3159–3228. Elsevier.

Horowitz, J. L. (2019). Bootstrap methods in econometrics. *Annual Review of Economics*, 11(3):193–224.

Huber, P. and Strassen, V. (1973). Minimax tests and neyman–pearson lemma for capacities. *The Annals of Statistics*, 1(2):251–263.

Hughes, D. W. and Hahn, J. (2020). The higher–order efficiency of jackknife bias corrections in panel models. Working Paper.

Hyslop, D. R. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica*, 67(6):1255–1294.

Imai, K., Kim, I. S., and Wang, E. H. (2021). Matching methods for causal inference with time–series cross–sectional data. *American Journal of Political Science*.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Kaido, H., Molinari, F., and Stoye, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica*, 87(4):1397–1432.

Kaido, H. and Zhang, Y. (2019). Robust likelihood-ratio tests for incomplete economic models. Working Paper.

Kasahara, H. and Shimotsu, K. (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175.

Kato, K., Galvao Jr., A. F., and Montes-Rojas, G. V. (2012). Asymptotics for panel quantile regression models with individual effects. *Journal of Econometrics*, 170(1):76–91.

Kawai, K. and Watanabe, Y. (2013). Inferring strategic voting. *American Economic Review*, 103(2):624–62.

Kim, M. S. and Sun, Y. (2016). Bootstrap and k-step bootstrap bias corrections for the fixed effects estimator in nonlinear panel data models. *Econometric Theory*, 32(6):1523–1568.

Kline, B. and Tamer, E. (2016). Bayesian inference in a class of partially identified models. *Quantitative Economics*, 7(2):329–366.

Kocherlakota, S. and Kocherlakota, K. (1991). Neyman's $c(\alpha)$ test and rao's efficient score test for composite hypotheses. *Statistics & Probability Letters*, 11:491–493.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413.

Lee, B.-S. and Ingram, B. F. (1991). Simulation estimation of time-series models. *Journal of Econometrics*, 47(2):197–205.

Leeb, H. and Pötscher (2005). Model selection and inference: Facts and fictions. *Econometric Thoery*, 51:21 – 59.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lewbel, A. (2007). Coherency and completeness of structural models containing a dummy endogenous variable. *International Economic Review*, 48(4):1379–1392.

Manski, C. and Lerman, S. (1981). On the use of simulated frequencies to approximate choice probabilities. In Manski, C. and McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications*, chapter 7, pages 2–50. MIT Press.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026.

Molinari, F. (2020). Microeconometrics with partial identification. In Durlauf, S. N., Hansen, L. P., Heckman, J. J., and Matzkin, R. L., editors, *Handbook of Econometrics*, volume 7, pages 355–486. Elsevier.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–1167.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Heckman, J. J. and Leamer, E., editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier.

Neyman, J. (1959). Asymptotic tests of composite statistical hypotheses. In *Probability and Statistics (The Harald Cramér Volume)*, pages 213–234. Almquist and Wiksells, Uppsala, Sweden.

Neyman, J. (1979). C(α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, 41(1):1–21.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 5(55):999–1033.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49(6):874–897.

Otsu, T., Pesendorfer, M., and Takahashi, Y. (2016). Pooling data across markets in dynamic markov games. *Quantitative Economics*, 7(2):523–559.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57(5):1027–1057.

Pelican, A. and Graham, B. S. (2021). An optimal test for strategic interaction in social and economic network formation between heterogeneous agents. Working Paper.

Philippe, F., Debs, G., and Jaffray, J.-Y. (1999). Decision making with monotone lower probabilities of infinite order. *Mathematics of Operations Research*, 24(3):767–784.

Quenouille, M. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 20(3):355–375.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika*, 43(1):353–360.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2014). A practical two-step method for testing moment inequalities. *Econometrica*, 82(5):1979–2002.

Roth, J., Sant'Anna, P., Bilinski, A., and Poe, J. (2022). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. Working Paper.

Schennach, S. M. (2014). Entropic latent variable integration via simulation. *Econometrica*, 82(1):345–385.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Shaikh, A. M. and Vytlacil, E. J. (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica*, 79(3):949–955.

Sheng, S. (2020). A structural econometric analysis of network formation games through subnetworks. *Econometrica*, 88(5):1829–1858.

Silvapulle, M. J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association*, 90(429):342–349.

Smith, A. A. (1993). Estimating nonlinear time–series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8(1):63–84.

Sun, L. and Shapiro, J. M. (2022). A linear panel model with heterogeneous coefficients and variation in exposure. *Journal of Economic Perspectives*, forthcoming.

Taber, C. and Sauer, R. M. (2021). Understanding women's wage growth using indirect inference with importance sampling. *Journal of Applied Econometrics*, 36(4):453–473.

Talagrand, M. (1994). Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, 70(1):147–165.

Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series of Statistics.

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1):1625–1651.

Wasserman, L. A. (1990). Belief functions and statistical inference. *Canadian Journal of Statistics*, 18(3):183–196.

White, H. (2000). *Asymptotic Theory for Econometricians: Revised Edition*. Emerald Publishing Limited.

Wollmann, T. G. (2018). Trucks without bailouts: Equilibrium product characteristics for commercial vehicles. *American Economic Review*, 108(6):1364–1406.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20(1):39–54.

Wooldridge, J. M. (2014). Quasi-Maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables. *Journal of Econometrics*, 182(1):226–234.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445.

Wooldridge, J. M. (2019). *Introductory Econometrics: A Modern Approach*. Cengage Learning, 7 edition.

CURRICULUM VITAE

Shuowen Chen

Department of Economics, Boston University
270 Bay State Road, Office B03B
Boston MA 02215
Phone: (585) 319-6057
Email: swchen@bu.edu
Website: https://shuowencs.github.io/
Citizenship: China

Education

Ph.D., Economics, Boston University, Boston MA, 2022
M.A., Economics, The University of Texas at Austin, Austin TX, 2016
B.A., Economics (Honors & Highest Distinction) and Mathematics (High Distinction), *Magna Cum Laude*, University of Rochester, Rochester NY, 2015

Fields of Interest

Econometrics, Empirical Industrial Organization, Economics of Innovation

Publications

SortedEffects: Sorted Causal Effects in R
(*with* Victor Chernozhukov, Iván Fernández-Val and Ye Luo)
The R Journal, 12(1): 131–146, 2020

Mastering Panel Metrics: Causal Impact of Democracy on Growth
(*with* Victor Chernozhukov and Iván Fernández-Val)
American Economic Association Papers and Proceedings, 109: 77–82, 2019

Working Papers

Indirect Inference for Nonlinear Panel Models with Fixed Effects

Robust Tests of Model Incompleteness in the Presence of Nuisance Parameters
(*with* Hiroaki Kaido)

R&D Heterogeneity and Countercyclical Productivity Dispersion
(*with* Yang Ming)

Crossover Jackknife Bias Correction for Non–Stationary Nonlinear Panel
(*with* Victor Chernozhukov, Iván Fernández-Val, Hiroyuki Kasahara and Paul Schrimpf)

Work In Progress

Dynamic Discrete Choice Models with Fixed Effects

Sensitivity Analysis of Estimation with Discretization

Presentations

Bates White LLC	2022
The Chinese University of Hong Kong	2022
BU–BC Green Line Econometrics Meeting	2021
Econometrics Seminar, Boston University	2018, 2020, 2021

Honors and Awards

Best Second Year Paper Award, Boston University	2019
Dean's Fellowship and Assistantship, Boston University	2016 – 2021
Merit–Based Scholarship, The University of Texas at Austin	2015
William Morse Hastings Essay Prize, University of Rochester	2015
Citation of Special Achievement in Economics, University of Rochester	2015
Dean's Scholarship, University of Rochester	2011 – 2015

Working Experience

Research Assistant to Hiroaki Kaido	Fall 2021 – Spring 2022
Research Assistant to Iván Fernández-Val	Fall 2020 – Spring 2021
Research Assistant to Hiroaki Kaido	Summer 2019
Research Assistant to Iván Fernández-Val	Fall 2018 – Spring 2019
Research Assistant to Pierre Perron	Summer 2018

Teaching Experience

Graduate Teaching Fellow at Boston University

EC 708 PhD Econometrics I	Spring 2020
EC 102 Principles of Macroeconomics	Fall 2017 – Spring 2018

Undergraduate Teaching Assistant at University of Rochester

ECO 217 Contract Theory	Fall 2014
ECO 211 Money, Credits, and Banking	Spring 2014
MTH 161 Calculus I	Fall 2013

Miscellaneous

Languages: English, Chinese
Programming: Julia, Mathematica, Matlab, Python, R, Stata/Mata
Organizer: BU Econometrics Reading Group (Spring 2020 – Spring 2021)

References

Professor Iván Fernández-Val
Department of Economics
Boston University
Phone: (617) 353-9670
Email: ivanf@bu.edu

Professor Hiroaki Kaido
Department of Economics
Boston University
Phone: (617) 358-5924
Email: hkaido@bu.edu

Professor Jean-Jacques Forneron
Department of Economics
Boston University
Phone: (617) 353-4824
Email: jjmf@bu.edu