

Moment Inequalities and Their Application.

A. Pakes, J. Porter, Kate Ho, and Joy Ishii*

April, 2005.
(First Version, August 2004).

Abstract

This paper provides conditions under which the inequality constraints generated by either single agent optimizing behavior, or by the Nash equilibria of multiple agent problems, can be used as a basis for estimation and inference. We then add to the econometric literature on inference on the parameters of models defined by inequality constraints by providing a new, easy to use, specification test and method of constructing confidence intervals. The paper concludes with two applications which illustrate how the use of inequality constraints simplify the problem of obtaining estimators from complex behavioral models.

1 Introduction

This paper provides conditions under which the inequality constraints generated by single agent optimizing behavior, or by the Nash equilibria of multiple agent problems, can be used as a basis for estimation and inference. The conditions allow for discrete and/or bounded choice sets, endogenous regressors, and though they do restrict the relationship of the disturbances to the choices made, they do not require the researcher to specify a parametric form for the disturbance distribution. We then add to the econometric

*The authors are from Harvard University and the National Bureau of Economic Research, the University of Wisconsin, Harvard University, and Harvard University, respectively. We thank Jeremy Fox, Oliver Hart, and Ali Hortascu for valuable comments. Pakes and Porter thank their respective NSF grants for financial support.

literature on inference on the parameters of models which generate inequality constraints by providing a new specification test and method for constructing confidence intervals. Both the test and the confidence intervals are easy to construct. We conclude with two empirical applications. The applications illustrate how the use of inequality constraints can simplify the problem of obtaining estimators from the constraints generated by complex behavioral models. The examples also allow us to provide detail on the properties of our inferential procedures.

We begin with a description of the problem we consider. The econometrician observes a set of choices made by various agents and is willing to assume that the agents expected the choices they made to lead to returns that were higher (or at least not “too much” lower) than the returns the agents would have earned had they made different choices from a known set of alternatives. We do not restrict the set of alternatives (so we could be considering discrete choice, ordered choice, continuous but bounded choices, ...), and we do not assume the agents know all the determinants of returns when they make their decision (so we allow for uncertainty).

We assume that we can calculate the returns from the actual choice and at least one alternative up to a parameter vector of interest and an additive disturbance. The returns of one agent can be affected by the decisions of other agents, and we do not assume the observed determinants of returns to be orthogonal to the disturbance; so we can allow for interacting agents, discrete choice, and endogenous regressors. Moreover when there are interacting agents we do not assume that there is a unique set of decisions that simultaneously satisfy our “best response” condition, and we need not specify precisely what information each agent has about the determinants of the returns of its competitors. Finally we do not make a functional form assumption on the distribution of the disturbances.

Given these conditions we consider a set of estimators constructed roughly as follows. Compute the sample average of the difference between the observable part of the actual realized returns and the observable part of returns that would have been earned had the alternative choice been made for different values of the parameter vector and accept any value that makes that difference non-negative. More precisely we interact certain linear combinations of these average differences in observable profits with positive functions of our instruments, and search for values of the parameter vector that make this vector of (weighted) profitability differences positive. This approach is a modified method of moments algorithm (Hansen,1982); the modification

being that at the true value of the parameter vector the moments conditions hold as inequalities (rather than as equalities, as in Hansen).

Section 2 of the paper provides conditions under which our vector of inequalities have positive expectation when evaluated at the true value of the parameter vector. Section 3 assumes this inequality condition and provides methods of inference for that parameter vector. Section 4 applies these techniques to two empirical examples that are of substantive interest and could not have been analyzed using more traditional techniques (at least not without further assumptions).

The functional form and stochastic assumptions that lead to our inequality conditions distinguish between unobservable determinants of profits that the decision could not have been a function of, and those that it could have. Unobservable determinants of profits that do not effect decisions include both (i) expectational errors caused by either realizations of random variables that were not known at the time decisions are made or by asymmetric information known to a proper subset of a group of interacting agents, and (ii) measurement errors in either the variables of interest or in the construction of the return function (say due to simulation). We assume that unobservable determinants of profits that can affect the agent's decisions (or our "structural disturbances") enter the return function additively and are mean independent of a known subset of observables (our "instruments").

In problems where structural disturbances can be ignored, any inequality formed from the difference between the profits at the actual choice and the profits at an alternative feasible choice should, under standard regularity conditions, have positive expectation at the true value of the parameter vector. This is the inequality analogue of the case considered in Hansen and Singleton (1982), and what the weakening of their equality constraints allows us to do is to analyze problems with more complex choice sets, interacting agents, and (as will be explained below) agents that are not always able to optimize precisely.

We provide a sufficient condition for obtaining profitability differences that have positive expectation at the true value of the parameter vector when both structural and non-structural disturbances are present. The condition assumes that we can find a linear combination of profitability differences that is additive in the structural disturbance *no matter* the actual decisions made. This allows us to form "unconditional" covariances of the structural disturbance and the observables that are orthogonal to it, and by our stochastic assumptions these differences have expectation zero. The examples show

that this logic can be used in: (i) ordered choice problems, (ii) in contracting problems when the expected transfers between agents that result from the contract have a structural disturbance, and (iii) when we observe multiple decisions by the same agent (and/or involving the same choice) and the structural unobservables are agent (or choice) specific (then we can form difference in difference *inequalities* that have positive expectation conditional on our instruments).

Section 3 builds on recently developed econometric methods for estimation subject to inequality constraints (Andrews, Berry, and Jia 2004, Chernozhukov, Hong, and Tamer 2003). We provide two new methods for confidence region construction, and a new specification test of the model. The first method of confidence interval construction is computationally simple and general enough to be applied to any problem fitting our framework. However it provides conservative inference. As we will show in our examples, just how conservative differs with particular properties of the data and can generally be judged a priori. This method also leads to a test of the inequality constraints per se, but the test is often also conservative.

Specification testing is likely to be important in our context. We expect users to want to be able to test for the importance of allowing for structural disturbances, and inequality tests are likely to be more robust to small deviations in modelling assumptions than tests of a point null hypothesis. Consequently we develop a relatively easy to use alternative test which should be quite a bit more powerful. The alternative test statistic is obtained by adjusting the logic of traditional tests of overidentifying restrictions in method of moment models for the presence of inequalities.

The second method of confidence interval construction is currently tailored to a leading special case: models which are linear in their parameters. This method simulates from the estimated limiting distribution of the data moments and uses the estimates formed from the simulated moments to generate an approximation to the joint distribution of the estimator.

The payoff to using inequalities is an ability to analyze new problems and to provide a better understanding of the dependence of previous results on their assumptions. The estimator is also extremely easy to obtain, so there is no computational cost to using it (and there may be a benefit). However there is likely to be a cost in terms of the precision of inference and the extent of that cost will be problem specific.

Our two empirical applications are both informative and encouraging in this respect. They are both problems: (i) which could not have been ana-

lyzed with more traditional tools, and (ii) with sample sizes that are quite small. The small sample sizes do force us to use parsimonious specifications. However the results make it quite clear that the new techniques provide useful information on important parameters; information which could not have been unraveled using more traditional estimation methods. In addition to illustrating the potential of the inequality techniques, the examples are also used to explain details of our inferential procedures.

The first example shows how our setup can be used to analyze investment problems with non-convex or “lumpy” investment alternatives; it analyzes banks’ choices of the number of their ATM locations. It also illustrates the ability of the proposed framework to handle multiple (as well as a single) agent environments; and this particular environment is one where it is clear that there can be many possible “network” equilibria. Finally we use this example to develop the intuition underlying the properties of the estimators.

The second example illustrates how the proposed approach can be used to analyze the nature of contracts emanating from a market with a small number of *both* buyers and sellers. Though markets with a small number of buyers and sellers appear frequently in industrial organization, econometric analysis of their equilibrium outcomes had not been possible prior to this work. Our particular example analyzes the nature of the contracts between health insurance plans and hospitals.

In both examples, the results we obtain are compared to alternative estimators that come to mind for the respective problems. In one example the alternative procedure ignores endogenous regressors. In the other, one of the two alternatives assumes away the non-structural error in the profit measures and the other alternative assumes away the discreteness in the choice set. The empirical results make it clear that accounting for both endogenous regressors and non-structural errors in discrete choice problems can be extremely important. The more detailed substantive implications of the parameter estimates are discussed in Ho (2004) and Ishii (2004).

Related Econometric Literature

A recent and important body of work has considered the general issue of inference for models with partially identified parameters and the more specific problem of estimation subject to inequality restrictions. This section notes our debt to that literature.

Both Chernozhukov, Hong, and Tamer (2003) and Andrews, Berry, and

Jia (2004) consider identification issues, derive properties of set estimators, and consider inferential procedures for models with inequalities. Both these papers are primarily concerned with the econometric issues surrounding a given set of inequalities, rather than with how these inequalities might be obtained from an underlying model¹. As will be noted below our first inferential procedure is closely related to a method in Andrews, Berry, and Jia. Other related econometric literature includes Moon and Schorfheide (2004), who examine an empirical likelihood approach to estimation, and explicitly consider equalities as well as inequalities in their estimation procedures. Earlier work considering partial parameter identification includes Manski (2003), Horowitz and Manski (1998), and Hansen, Heaton, and Luttmer (1995). Imbens and Manski (2003) consider the distinction between inference on the identified parameter set and individual elements of that set, a distinction we come back to below.

2 A Framework for the Analysis

This section describes derives the moment inequalities we take to data. We do this in the context of a Nash equilibrium to a simultaneous move game in pure strategies. However within this frameowrk our assumptions are quite general. In particular we do not restrict choice sets nor require a unique equilibrium, and we allow for both incomplete and assymetric information. After providing our assumptions we illustrate their use with some familiar examples, and then show how they generate the moment inequalities we use as a basis for inference. We conclude with a number of generalizations which show how to allow for; mixed strategies, various forms of non-Nash behavior, and non-simultaneous moves.

¹Andrews, Berry, and Jia (2004) motivate their results with a discrete choice model of interacting agents that exhibits multiple equilibria. In our terminology their example does not allow for a non-structural disturbance and assumes a parametric distribution on the structural disturbance, full information (so all agents know all determinants of profits of each other), and maximizing behavior. They then compute bounds on the probabilities of outcomes which become the inequalities they take to data. Cliberto and Tamer (2004) applies the methods developed in Chernozhukov, Hong, and Tamer (2003) to an entry problem in airline markets.

2.1 Agents' Problem

Suppose players (or “agents”) are indexed by $i = 1, \dots, n$. Let \mathcal{J}_i denote the information set available to agent i before any decisions are made, and $\mathbf{d}_i : \mathcal{J}_i \rightarrow \mathcal{D}_i$ denote a decision to be taken (or a strategy to be played) by agent i . For now we assume these decisions are pure strategies so \mathcal{D}_i is the set of possible values (the support) for \mathbf{d}_i .² Note that we distinguish between \mathbf{d}_i and the actual decision, say d_i , by using boldface for the former.

When $\mathcal{D}_i \subset \mathcal{R}$ it can be either a finite subset (as in “discrete choice” problems), countable (as in ordered choice problems), uncountable but bounded on one or more sides (as in continuous choice with the choice set confined to the positive orthant), or uncountable and unbounded. If \mathbf{d}_i is vector valued then \mathcal{D}_i is a subset of the appropriate product space.³

Let payoffs (or profits) to agent i be given by the function $\pi : \mathcal{D}_i \times \mathcal{D}_{-i} \times \mathbf{Y} \rightarrow \mathcal{R}$, where \mathcal{D}_{-i} denotes $\times_{j \neq i} \mathcal{D}_j$. In particular, returns to i are determined by agent i 's decision, d_i , other agents' decisions, d_{-i} , and an additional set of variables $y_i \in \mathbf{Y}$. Not all components of y_i need to be known to the agent at the time it makes its decisions and not all of its components need to be observed by the econometrician.

Let \mathcal{E} be the expectation operator and \mathbf{y}_i be the random variable whose realizations are given by y_i .⁴ The following assumption characterizes the behavior of agents in the game.

Assumption 1

$$\sup_{d \in \mathcal{D}_i} \mathcal{E}[\pi(d, \mathbf{d}_{-i}, \mathbf{y}_i) | \mathcal{J}_i, \mathbf{d}_i = d] \leq \mathcal{E}[\pi(d_i, \mathbf{d}_{-i}, \mathbf{y}_i) | \mathcal{J}_i, \mathbf{d}_i = d_i]$$

for $i = 1, \dots, n$. ♠

²Some of our examples require us to distinguish agents of different types (e.g. buyers and sellers in buyer-seller networks), but we refrain from introducing a type index until we need it.

³For example \mathcal{D}_i might be a vector of contract offers, with each contract consisting of a fixed fee and a price per unit bought (a two-part tariff). If a contract with one buyer precludes a contract with another, as in “exclusives” which insure a single vendor per market, \mathcal{D}_i becomes a proper subset of the product space of all possible two part tariffs.

⁴Formally this is the expectation corresponding the joint distribution of defined random variables as generated by market or game outcomes. We could have defined the expectation operator corresponding to each agent's perceptions, and then assumed that these perceptions are in fact correct for the play that generated our data. Though this is certainly sufficient for Assumption 1, it is not necessary (see Pakes,2005.)

In single agent problems, this assumption would simply be derived from optimizing behavior. For instance, with $n = 1$ and \mathcal{D}_i a finite set, Assumption 1 is an implication of a standard discrete choice problem. If \mathcal{D}_i is an interval then Assumption 1 generates the standard first order (or Kuhn-Tucker complementarity) conditions for optimal choice of a continuous control. When there are multiple interacting agents, Assumption 1 is a necessary condition for any Bayes-Nash equilibrium. It does not rule out multiple equilibria, and it does not assume anything about the selection mechanism used when there are multiple equilibria. In section 2.4, we discuss the relaxation of Assumption 1 to cover certain kinds of sub-optimal behavior.

Exogeneity and Profit Differences.

We will want to allow for sequential games in which the distribution of the determinants of profits (of our y) in the stages in which profits accrue will depend on the decisions made by all agents in earlier stages (e.g. the profits a bank earns from its ATM investments depend on the equilibrium interest rates in the periods in which those ATM's will be operative which, in turn, depend on the number of ATM's installed by the bank's competitors). This will require some notation and an additional assumption. In particular we let $\mathbf{y} : \mathcal{D}_i \times \mathcal{D}_{-i} \times \mathcal{Z} \rightarrow \mathcal{R}^{\#y}$, so that $\mathbf{y}_i = \mathbf{y}(\mathbf{d}_i, \mathbf{d}_{-i}, \mathbf{z}_i)$ for a random variable \mathbf{z} , with realizations that will be denoted by \mathbf{z} , and make the following assumption.

Assumption 2 *The distribution of $(\mathbf{d}_{-i}, \mathbf{z}_i)$ conditional on \mathcal{J}_i and $\mathbf{d}_i = d$ does not depend on d .*

Conditional independence of agents' decisions (of \mathbf{d}_{-i}) from \mathbf{d}_i is an implication of play in simultaneous move games. Conditional independence of the \mathbf{z}_i is a more substantive restriction. It insures that our approximation for the realization of $\pi(\cdot)$ that would have occurred had our agent made a different choice than the one actually made, i.e. $\pi(d', d_{-i}, y(d, d_{-i}, \mathbf{z}_i))$ for $d' \neq d_i$, has an expectation which conforms to that in Assumption 1.

More precisely if we define

$$\Delta\pi(d, d', d_{-i}, \mathbf{z}_i) = \pi(d, d_{-i}, y(d, d_{-i}, \mathbf{z}_i)) - \pi(d', d_{-i}, y(d, d_{-i}, \mathbf{z}_i)),$$

then assumptions 1 and 2 imply that for any $d' \in \mathcal{D}_i$

$$\mathcal{E}[\Delta\pi(d_i, d', \mathbf{d}_{-i}, \mathbf{z}_i) | \mathcal{J}_i] \geq 0.$$

2.2 Econometrician's Problem

The econometrician may not be able to measure profits exactly but can calculate an approximation to $\pi(\cdot)$, say $r(\cdot; \theta)$, which is known up to the parameter vector θ . $r(\cdot)$ is a function of d_i , d_{-i} , an *observable* vector of the determinants of profits, say z_i , and θ . For now think of z_i as the observable part of \mathbf{z}_i (though we show below that this can be generalized slightly). $\theta \in \Theta$ and its true value will be denoted by θ_0 . We obtain our approximation to the difference in profits that would have been earned had the agent chosen d' instead of d , say $\Delta r(d, d', \cdot)$, by evaluating $r(\cdot)$ at d and d' and taking the difference.

More formally $\Delta r(\cdot) : \mathcal{D}_i^2 \times \mathcal{D}_{-i} \times Z \times \Theta \rightarrow \mathcal{R}$ is a *known* function of; (d, d') , other agents' decisions, or d_{-i} , our observable determinants of profits, $z_i \in \mathbf{Z}$, and a parameter $\theta \in \Theta$. Let \mathbf{z}_i be the random variable whose realizations are given by z_i . Then the relationships between $\Delta\pi(\cdot)$ and $\Delta r(\cdot)$ and z_i and \mathbf{z}_i define the following two unobservables.

Definitions. For $i = 1, \dots, n$, and $(d, d') \in \mathcal{D}_i^2$ define

$$\nu_{2,i,d,d'} = \mathcal{E}[\Delta\pi(d, d', \mathbf{d}_{-i}, \mathbf{z}_i) | \mathcal{J}_i] - \mathcal{E}[\Delta r(d, d', \mathbf{d}_{-i}, \mathbf{z}_i, \theta_0) | \mathcal{J}_i], \text{ and } (1)$$

$$\begin{aligned} \nu_{1,i,d,d'} &= \Delta\pi(d, d', d_{-i}, z_i) - \Delta r(d, d', d_{-i}, z_i, \theta_0) \\ &\quad - \{\mathcal{E}[\Delta\pi(d, d', \mathbf{d}_{-i}, \mathbf{z}_i) | \mathcal{J}_i] - \mathcal{E}[\Delta r(d, d', \mathbf{d}_{-i}, \mathbf{z}_i, \theta_0) | \mathcal{J}_i]\}. \end{aligned} \quad (2)$$

It follows that

$$\Delta\pi(d, d', d_{-i}, z_i) = \Delta r(d, d', d_{-i}, z_i, \theta_0) + \nu_{1,i,d,d'} + \nu_{2,i,d,d'}. \quad (3)$$

$\Delta r(\cdot, \theta)$ provides the econometrician's measure of the profit change that would result from changing from $d_i = d$ to $d_i = d'$. ν_1 and ν_2 are the determinants of the profit difference that are *not observed* by the econometrician. They differ in what the agent (in contrast to the econometrician) knows about them. The agent knows its ν_2 value *before* it makes its decision, i.e. $\nu_2 \in \mathcal{J}$. Since $d_i = \mathbf{d}(\mathcal{J}_i)$, we expect d_i to be a function of $\nu_{2,i,d_i,d'}$. In contrast the agent makes its decision not knowing ν_1 ($\nu_{1,i,d_i,d'}$ can be a function of the realizations of \mathbf{d}_{-i} , \mathbf{z}_i , and z_i ; all variables the agent does not know when it makes its decision). Indeed $\mathcal{E}[\nu_{1,i,d_i,d'} | \mathcal{J}_i] = 0$ by construction, and

as a result $\mathcal{E}[\nu_{1,i,d_i,d'}|d_i] = 0$. Note that the values of ν_2 and of ν_1 differ both across (d, d') couples and across agents (i) .

The importance of accounting for one or both of (ν_1, ν_2) is likely to be different in different applied problems. Differences between ν_1 and zero do not change the agent's expected profits at the time decisions are made. So ν_1 realizations can be caused by either expectational or measurement errors. There are two sources of expectational errors: (i) incomplete information on the environmental variables that will determine the profits that result from the agent's decision, and (ii) asymmetric information, or incomplete information on either the z_{-i} 's or the $\nu_{2,-i}$'s that determine the decisions of the agent's competitors. Similarly there are two sources of measurement errors: (i) classical measurement error in the profit variable (in components of revenues or of costs), and (ii) errors induced by the way $r(\cdot)$ is constructed (we consider measures which contain simulation error presently).

In contrast ν_2 is a "structural" disturbance, i.e. a source of variance in the difference in profits that the agent conditions its decisions on, but that the econometrician does not observe. Variation in ν_2 will be important when $\Delta r(d, d', \cdot)$ does not account for an important source of variance in $\Delta\pi(d, d', \cdot)$ that the agent accounts for when it makes its decision (we are more explicit about how this can happen in discussing our examples).

A number of other points about these definitions are worth noting. First note that we have not had to specify whether the $(z_{-i}, \nu_{2,-i})$ is in agent i 's information set at the time decisions are made. \mathcal{J}_i could contain these values, could contain a signal on their likely values, or could not contain any information on their values at all⁵. Relatedly we need not make a particular assumption on the relationship of the $\{\nu_{2,i}\}$ draws of the different agents.

Second the \mathbf{z}_i that determine $\Delta\pi(\cdot)$ and the \mathbf{z}_i that determine $\Delta r(\cdot)$ (and hence whose outcomes need to be observable), can be different random variables. This allows for the use of simulation estimators. In particular it allows for cases where $\Delta\pi(\cdot)$ is a *nonlinear* function of an *unobserved* random variable whose value may be known to the agent when it makes its decision.

⁵The fact that we need not be explicit about the contents of information sets differentiates our setup and from the setups used in most prior applied work in Industrial Organization. For example Bresnahan and Reiss, 1991, Berry, 1992, assume full information; Seim 2002 and Pakes Ostrovsky and Berry 2003 assume no knowledge of $\nu_{2,-i}$; and Fershtman and Pakes 2004 allow for signals. Of course if we knew (or were willing to assume) more on the properties of the $\nu_{2,i}$ we might well be able to provide more precise estimators of θ (see, for example, Bajari, Hong and Ryan, 2004).

However the distribution of the unobservable that enters the profit function in a nonlinear way, conditional on \mathcal{J}_i , must be *known* to the econometrician. In this case the \mathbf{z}_i contains the value of the unobservable faced by the agent, \mathbf{z}_i contains a simulation draw from the conditional distribution of this unobservable, and $\mathcal{E}[\pi(\cdot)|\mathcal{J}_i] = \mathcal{E}[r(\cdot)|\mathcal{J}_i] + \nu_2$.

On the other hand the combination of assuming that $\nu_{2,-i}$ is unobservable while z_i is observable, and of not making $\Delta r(\cdot)$ a function of $\nu_{2,-i}$, implies that the ν_2 's of the firm's competitors only affects its profits indirectly, through $\nu_{2,-i}$'s effects on (z_i, d_i, d_{-i}) .

Selection.

Our assumptions thus far are not very stringent. In addition to not assuming what each agent knows about its competitors, we *have not* specified either a particular form for the distribution of ν_1 or ν_2 , and we *have* allowed for discrete choice sets and endogenous regressors. We do however require an additional assumption. This because d_i is both a determinant of profits and is partially determined by an unobservable determinant of profits (the $\nu_{2,i}$). This implies that conditional on (z_i, θ) an observation on d_i implies that the draw on $\nu_{2,i}$ was *selected* from a subset of the support of the ν_2 distribution. I.e. if $d_i = d$ and $\mathcal{E}[\Delta r(d, d', \cdot)|\mathcal{J}] \leq 0$, then $\nu_{2,d,d'} > 0$.

The next assumption offers a route for overcoming this selection problem. We will analyze averages of nonnegative linear combinations of our observed proxies for the profit differences (of $\Delta r(d_i, d', \cdot; \theta)$), and consider values of θ which make these averages positive. Assumption 1 insures that the the analogous linear combination of the $\Delta \pi(d, d', \cdot)$ has a positive conditional expectation. Equation (2) then implies that the expectation of the observable $\Delta r(d_i, d', \cdot; \theta)$ will be positive at $\theta = \theta_0$ provided the conditional expectation of $\nu_{1,i,d(i),d'}$ and $\nu_{2,i,d(i),d'}$ are not positive. If the weights are functions of the agents' information sets, the definition of $\nu_{1,i,d(i),d'}$ insures that the relevant linear combinations of the $\nu_{1,i,d(i),d'}$ will have zero expectation. Assumption 3 provides conditions which suffice to insure that the expectation of the same linear combination of the $\nu_{2,i,d(i),d'}$ is not positive.

Assumption 3 constrains the relationship between the ν_2 and the $\mathcal{E}[\Delta r(\cdot)|\mathcal{J}]$ in equation (2). Special cases of this assumption occur when we can find a linear combination of the $\Delta r(\cdot)$ that either does not involve ν_2 , or generates the same ν_2 value no matter the realization of \mathbf{d}_i (for this to occur the $\nu_{2,d,d'}$ values must be constrained in some fashion). In the latter case we employ

instruments to account for possible correlations between ν_2 and the other observable determinants of profits (e.g. (d_{-i}, d_i)). A third case occurs when the linear combination generates a ν_2 with a negative correlation with an $x \in \mathcal{J}$ conditional on $\mathbf{d} = d$. After presenting the assumption we consider some familiar examples which satisfy it.

Assumption 3 *Let h_i be a function which maps x_i into a nonnegative Euclidean orthant. Assume that for an x_i that is both in \mathcal{J}_i and is observed by the econometrician, and a nonnegative weight function $\chi_{d_i, \mathcal{J}_i}^i : \mathcal{D}_i \rightarrow \mathbb{R}^+$ whose value can depend on the realization of \mathbf{d}_i*

$$\mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{d_i, \mathcal{J}_i}^i(d') \nu_{2,i,d_i,d'} h(x_i)\right] \leq 0,^6$$

where $\nu_{2,i,d_i,d'} = \sum_{d \in \mathcal{D}_i} \mathbf{1}\{d_i = d\} \nu_{2,i,d,d'}$.

Two points about Assumption 3 are worth stressing. First x_i can be a proper subset of \mathcal{J}_i . Since both the choice set and the distributions of the unobservables are unrestricted, this allows us to analyze *discrete* choice sets and *endogenous* regressors without making a particular assumption on the distribution of the $\{\nu_2\}$. Second though there is a sense in which the x play the role of an “instrument” in more traditional work, our “instrument” need not generate a moment *equality*; we require only an inequality. I.e. it is sufficient for x to be negatively correlated with the omitted variable the agent knew when it made its decision (our ν_2).

Example 1. $\pi(\cdot)$ is observable up to a parameter vector of interest and an error which is mean zero conditional on the agent’s information set. Formally this is the special case where $\nu_{2,i,d,d'} = 0$ is identically zero for all d, d' , so that Assumption 3 is satisfied with $h = 1$ and any χ^i which weights a $d' \in \mathcal{D}_i$. For example pick any d' and set $\chi^i(d') = 1$ and zero elsewhere. Then

$$\Delta\pi(d_i, d', \mathbf{d}_{-i}, y_i) = \Delta r(d_i, d', \mathbf{d}_{-i}, y_i, \theta_0) + \nu_{1,i,d_i,d'},$$

and our assumptions are satisfied.

⁶An inequality applied to a vector means the inequality holds for every element of the vector.

We note that our functional form and stochastic assumptions are then those of Hansen and Singleton (1982), but our estimator: (i) allows for more general (discrete and/or bounded) choice sets; (ii) allows explicitly for interacting agents (clarifying the conditions that must hold in that case); and (iii), as we discuss in the generalizations, allows for agents whose choices are not always exactly optimal conditional on a prespecified information set. We are able to do this because we assume an ability to compute the profits that would have been earned if the alternative actions had been made up to the parameter of interest and a mean zero disturbance (Hansen and Singleton, 1986, assume an ability to calculate the first derivative of expected returns). Note that these assumptions allow us to analyze discrete choice problems provided the disturbance is unknown to the agent when it makes its decision. This includes models with entry and exit when it is assumed that there are fixed unknown costs of the discrete decisions (as in Bajari, Levin and Benkard, 2004) and an ability to compute continuation values up to expectational and/or measurement error (as in Pakes, Ostrovsky, and Berry, 2004). More generally these assumptions are relevant for any problem for which we can measure profits up to a mean zero error, so they constitute a special case we might often want to test for.

Example 2. (*Fixed Effects.*) Assumption 3 will be satisfied with $h = 1$ by establishing weights such that $\sum_{i=1}^n \sum_{d'} \chi_{d_i, \mathcal{J}_i}^i(d') \nu_{2,i,d_i,d'} = 0$ (the previous example was a special case of this, but one with special relevance for applied work). An example where this might occur is when the ν_2 represent agent specific fixed effects that enter multiple decisions made by the same agent. If for *certain* (vectors of) decisions d , the econometrician knows that there are *certain* alternative decisions d' such that $\nu_{2,i,d_i,d'} = 0$, then we have enough to satisfy Assumption 3. In this case, set $\chi_d^i(d') = 1$ for any decision and alternative (d, d') with $\nu_{2,i,d_i,d'} = 0$ and $\chi^i = 0$ otherwise. The decisions d_i with some nonzero $\chi_{d_i}^i(\cdot)$ must have positive probability of occurring for our inequalities to be informative on θ .

In the fixed effect case the resulting moments are the average of differences, across choices, of the difference in returns between the optimal and an alternative feasible choice; i.e. they are *difference in difference inequalities*. To see this suppose agents make two simultaneous decisions so $d = (d_a, d_b)$ with $d_w \in \{0, 1\}$ for $w = \{a, b\}$. For simplicity, assume also that the profit

function is additive across choices⁷, so that $\Delta\pi(d, d', \cdot) = \sum_w \Delta\pi_w(d_w, d'_w, \cdot)$, and

$$\Delta\pi_w(d_{i,w}, d'_{i,w}, \cdot) = \Delta r_w(d_{i,w}, d'_{i,w}, \cdot) + (d_{i,w} - d'_{i,w})\nu_{2,i} + \nu_{1,i,d_{i,w},d'_{i,w}}^w,$$

for $w = \{a, b\}$. Set $\chi_{d_i}^i(d')$ to one whenever $d = (1, 0)$ and $d' = (0, 1)$ (or vice versa), and zero otherwise. Then if $\mathbf{1}\{\cdot\}$ is notation for the indicator function

$$\begin{aligned} \sum_{d'} \chi_{d_i}^i(d') \Delta\pi(d_i, d', \cdot) &= \sum_{d'} \chi_{d_i}^i(d') [\Delta\pi_a(d_{i,a}, d'_a, \cdot) + \Delta\pi_b(d_{i,b}, d'_b, \cdot)] \\ &= \mathbf{1}\{d_i = (1, 0)\} \Delta r_a(d_{i,a} = 1, d'_a = 0, \cdot) + \Delta r_b(d_{i,b} = 0, d'_b = 1, \cdot) \\ &\quad + \mathbf{1}\{d_i = (0, 1)\} \Delta r_a(d_{i,a} = 0, d'_a = 1, \cdot) + \Delta r_b(d_{i,b} = 1, d'_b = 0, \cdot) \\ &\quad + \sum_{d'} \chi_{d_i}^i(d') \nu_{1,i,d_i,d'}, \end{aligned}$$

and the last term is mean independent of any $x \in \mathcal{J}_i$.

Note that this example covers discrete choice panel data models with fixed effects when the latent dependent variable depends only on the “regression function” and the fixed effect, and the idiosyncratic disturbance is a result of expectational or measurement error (and we do not require assumptions on the distributions of either the fixed effects or the idiosyncratic disturbances, though we do require the usual strict exogeneity assumption). Our second example, Pakes (2005), and Pakes Porter and Wolfram (2005) use generalizations of this idea, i.e. multiple observations containing a disturbance known to the agent but not to the econometrician, to analyze the nature of contracts in buyer/seller networks and the cost functions of electric utilities.

Example 3. (*Ordered choice*). This example assumes weights that yield a mean zero unconditional expectation for the ν_2 's, or

$$\mathcal{E}\left[\sum_{i=1}^n \sum_{d'} \chi_{d_i, \mathcal{J}_i}^i(d') \nu_{2,i,d_i,d'}\right] = 0.$$

Ordered choice, or any discrete choice with an order to the choice set and a determinant of the agent's ordering that is not observed by the econometrician (which becomes ν_2), is covered by this example. Lumpy investment

⁷To analyze the nonadditive case simply assume that for any (d, d') , $\Delta\pi(d, d', \cdot) = \Delta r(d, d', \cdot) + [(d_a - d'_a) + (d_b - d'_b)]\nu_{2,i} + \nu_{1,i,d,d'}$.

decisions (say in the number of stores or machines) are often treated as ordered choice problems, and our first empirical example is a case in point. It has markets consisting of sets of interacting firms each of whom decides how many units of a machine to purchase and install. The parameter of interest (θ) determines the average (across firms) of the cost of installing and operating machines, and the model allows that cost to differ across firms in a way which is known to the firm when they make their decisions but not observed by the econometrician (our ν_2).

With constant marginal costs the difference in profits from installing d versus d' machines includes a cost difference equal to $(d - d')(\theta + \nu_2)$. So if $r(\cdot)$ provides the revenues, the incremental profits from the last machine bought are

$$\Delta\pi(d_i, d_i - 1, \mathbf{d}_{-i}, y_i) = \Delta r(d_i, d_i - 1, \mathbf{d}_{-i}, z_i, \theta_1) - (\theta_0 + \nu_{2,i}) + \nu_{1,i,d,d-1},$$

Since θ_0 is the average marginal cost across firms, then $\mathcal{E}\nu_{2,i} = 0$, and Assumption 3 is satisfied with $h = 1$ and $\chi_{d_i}^i(d') = 1$ only if $d_i = d' + 1$. Consequently

$$\mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{d_i, \mathcal{J}_i(d')}^i \nu_{2,i,d_i,d'} h(x_i)\right] = \sum_{i=1}^n \mathcal{E}[\nu_{2,i,d_i,d_i-1}] = \sum_{i=1}^n \mathcal{E}[-\nu_{2,i}] = 0.$$

Note that here $\nu_{2,i,d,d-1} = \nu_{2,i}$ which does not depend on d . Hence this choice of weight function χ^i eliminates any selection effect. Assumptions 1 and 2 then give the needed moment inequality. We provide a fuller discussion of this case, including a discussion of identification, below⁸.

Example 4. The ν_2 's (or a weighted sum of ν_2 's) are mean independent of a *subset* of the variables that the agents know when they make their decisions, a subset which will become our “instruments,” x ,

$$\mathcal{E}\left[\sum_i \sum_{d'} \chi_{d_i, \mathcal{J}_i}^i(d') \nu_{2,i,d_i,d'} | x_i\right] = 0.$$

Example 3 could be extended to include this case by assuming an $x \in \mathcal{J}_i$ that satisfied $\mathcal{E}[\nu_{2,i} | x] = 0$. Our second empirical example is another, and since it is of some applied interest, we deal with it explicitly here.

⁸The discussion above has implicitly assumed that there are no corners to the choice set (there are feasible choices that are higher and lower than every possible observed choice.). The discussion below considers corners in some detail.

Buyer-seller networks with unobserved transfers. Here we need to introduce a set of types, say $\mathcal{T} = \{\mathbf{b}, \mathbf{s}\}$ for buyers and sellers respectively, and let the choice of inequalities and the form of the primitives (the choice set, profit function,...) differ by type. Type \mathbf{b} 's incremental cost is the cost of purchase, and its incremental expected returns are the expected profit from resale. Type \mathbf{s} 's incremental returns are \mathbf{b} 's purchase cost and its incremental costs are the costs of production. Assume that sellers make take it or leave it offers to buyers. Note that since buyers know the sellers' offers before they determine whether to accept, this is our first example which is not a simultaneous move game. The offers themselves are not public information (they are proprietary), and it is their properties that we want to investigate empirically. We assume that the offers are a parametric function of observables (e.g. an unknown markup per unit purchased) and an error (ν_2).

For now assume there is only one seller and one buyer in each market studied. \mathcal{D}_s is the set of contracts which can be offered. We assume it includes a null contract, say $d_s = \phi$, that is never accepted. A contract which is not accepted does not generate any profit for the seller. $\mathcal{D}_b = \{0, 1\}$ with $d_b = 1$ indicating the contract was accepted. Note that any transfer cost to the buyer is a revenue for the seller, so there is only one value of ν_2 per market and it enters the profits of the buyer and the seller with opposite signs.

Assumption 1 implies that; (i) the expected profits to the seller from the contract it offered are larger than what they would have had the seller offered the null contract, and (ii) if the buyer rejects the offer it is because profits without the contract are higher than profits with the contract. Let $\Delta\pi^s(d_s, \phi, \mathbf{d}_b = 1, \cdot)$ be the increase in seller profits if there is a contract, and $x \in \mathcal{J}_s \cap \mathcal{J}_b$ be an instrument in the sense that $\mathcal{E}[\nu_2|x] = 0$. Then

$$\Delta\pi^s(d_s, \phi, \mathbf{d}_b, \cdot) = I\{\mathbf{d}_b = 1\}[\Delta r^s(d_s, \phi, \mathbf{d}_b = 1, \cdot; \theta) + \nu_2 + \nu_{1,s}(\cdot)],$$

while if the buyer rejects the offer it saves

$$\Delta\pi^b(0, 1, d_s, \cdot) = \Delta r^b(0, 1, d_s, \cdot; \theta) + \nu_2 + \nu_{1,b}(\cdot).$$

Set $\chi_{d(s)}^s(d'_s = \phi) = \chi_{d_b=0}^b(d'_b = 1) = 1$, and the rest of the $\chi^i(\cdot) = 0$. Then

$$0 \leq \mathcal{E}\left[\sum_{i=b}^s \chi_{d(i)}^i(d'_i) \Delta\pi^i(d_i, d_{-i}, \cdot)|x\right] =$$

$$\begin{aligned}
& \mathcal{E}[\Delta r^s(d_s, \phi, d_b = 1, \cdot, \theta) + \nu_2 | x] Pr(d_b = 1 | x) + \mathcal{E}[\Delta r^b(0, 1, d_s, \cdot, \theta) + \nu_2 | x] [1 - Pr(d_b = 1 | x)] \\
&= \mathcal{E}\left[\sum_{i=b}^s \chi_{d(i)}^i(d'_i) \Delta r^i(d_i, d_{-i}, \cdot) | x\right] + E[\nu_2 | x] = \mathcal{E}\left[\sum_{i=b}^s \chi_{d(i)}^i(d'_i) \Delta r^i(d_i, d_{-i}, \cdot) | x\right].
\end{aligned}$$

Our second empirical example is a generalization of this one.

All examples above generated ν_2 averages with zero, in contrast to negative, expectations. However strict inequalities are often needed. For one example add a non-negative cost of switching decisions to Example 2. The discussion of boundaries in our first empirical example provides another.

2.3 Inequality Conditions

Equation (3) implies that

$$\begin{aligned}
& \mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{d_i, \mathcal{J}_i}^i(d') \Delta r(d_i, d', \mathbf{d}_{-i}, \mathbf{z}_i, \theta_0) h(x_i)\right] = \mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{d_i, \mathcal{J}_i}^i(d') \Delta \pi(d_i, d', \mathbf{d}_{-i}, \mathbf{z}_i) h(x_i)\right] \\
& - \mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{d_i, \mathcal{J}_i}^i(d') \nu_{1,i,d_i,d'} h(x_i)\right] - \mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{d_i, \mathcal{J}_i}^i(d') \nu_{2,i,d_i,d'} h(x_i)\right]. \quad (4)
\end{aligned}$$

We consider each of the three terms in equation (4) in turn. Each summand in the first term can be written as

$$\mathcal{E}[\chi_{d_i, \mathcal{J}_i}^i(d') \Delta \pi(d_i, d', \mathbf{d}_{-i}, \mathbf{z}_i) h(x_i)] = \mathcal{E}[\chi_{d_i, \mathcal{J}_i}^i(d') \mathcal{E}_{\mathbf{d}_{-i}, \mathbf{z}_i}[\Delta \pi(d_i, d', \mathbf{d}_{-i}, \mathbf{z}_i) | \mathcal{J}_i] h(x_i)] \geq 0$$

where $\mathcal{E}_{\mathbf{d}_{-i}, \mathbf{z}_i}[\Delta \pi(\cdot) | \mathcal{J}_i]$ is notation for the expectation of $\Delta \pi(\cdot)$ over $(\mathbf{d}_{-i}, \mathbf{z}_i)$ conditional on \mathcal{J}_i , the equality follows from Assumption 2, and the inequality follows from Assumption 1 and the fact that both $\chi_{d_i, \mathcal{J}_i}^i(d')$ and $h(x_i)$ are non-negative and functions of \mathcal{J}_i .

Since the definition of ν_1 in equation (1) insures that

$$\mathcal{E}[\nu_{1,i,d_i,d'} | \mathcal{J}_i, d_i] = 0,$$

and Assumption 3 states that the last term in equation (4) is non-negative, we have

$$\mathcal{E}\left[\sum_{i=1}^n \sum_{d' \in \mathcal{D}_i} \chi_{d_i, \mathcal{J}_i}^i(d') \Delta r(d_i, d', \mathbf{d}_{-i}, \mathbf{z}_i, \theta_0) h(x_i)\right] \geq 0. \quad (5)$$

Equation (5) depends only on observables and θ_0 , so we can form its sample analog and look for values of θ that satisfy it⁹.

2.4 Generalizations

For expositional ease the assumptions used in sections 2.1 and 2.2 were not as general as they could have been. Here we list a number of generalizations and show how they generate moment inequalities that are analogous to those in equation (5).

Generalization 1. (*Non-optimal Decision-making*) It is possible to weaken Assumption 1 considerably. Consider the generalization

$$\sup_{d \in \mathcal{D}_i(d_i)} \mathcal{E}[\pi(d, \mathbf{d}_{-i}, y_i) | \mathcal{J}_i, d_i = d] \leq (1 + \delta) \mathcal{E}[\pi(d_i, \mathbf{d}_{-i}, y_i) | \mathcal{J}_i, d_i = d_i]$$

for $i = 1, \dots, n$. This version of Assumption 1 allows the decision space of the alternative, $\mathcal{D}_i(d_i)$, to be a subset of \mathcal{D}_i , and allows the agent to make decisions which are only within a multiplicative factor $1 + \delta$ of the decisions that maximizes the expected value of the outcome.

When $\mathcal{D}_i(d_i) = \mathcal{D}_i$ and $\delta = 0$, we are back to Assumption 1. However if, for example, $\delta = .5$, then non-optimal choices would be allowed provided they did not, on average, reduce expected profits more than 50% from the expected profits that would be earned from optimal strategies. Complementary reasoning applies to the actions per se when $\delta = 0$ but $\mathcal{D}_i(d_i) \neq \mathcal{D}_i$. For example, if there was a continuous control, and we specified that $\mathcal{D}_i(d_i) = \{d : |d - d_i| \geq \alpha, d_i, d \in \mathcal{D}\}$ for some $\alpha > 0$, then we would be specifying that though small deviations about optimal behavior can occur (deviations that leave the choice within $100\alpha\%$ of the optimal decision), at least on average large deviations do not occur¹⁰. The inequalities carry

⁹In general Assumptions 1, 2, and 3 are sufficient but not necessary for the inequalities in (5), which, in turn, provide the basis for estimation and inference. I.e. we expect that there are alternative conditions that will also suffice.

¹⁰One would typically assume δ and $\mathcal{D}_i(\cdot)$ are set exogenously though we could begin the analysis assuming $\delta = 0$ and $\mathcal{D}_i(d_i) = \mathcal{D}_i$, and then test whether the data is consistent with those assumptions. If it is not, find a relaxation of those assumptions that *is* consistent with the data; for example find a value for δ that satisfies the inequalities (up to sampling error) and the implied estimator of the parameter vector. Note that this procedure maintains our assumptions on functional forms and asks only whether, given those functional

through with this assumption provided we alter the definitions of $\Delta\pi$ and Δr to account for the $1 + \delta$ factor.

Generalization 2. (*Non-simultaneous Move Games.*) To illustrate the problems that can arise in non-simultaneous move games, we relax the assumptions in example 4 to allow for multiple buyers and sellers. Sellers make simultaneous take it or leave it offers to buyers. Buyers respond simultaneously at some later date. The buyers can still accept or reject any given contract without changing any other contract outcome. However the environment that would result if seller s^* changed its contract offer to a buyer who had accepted the original contract, say buyer b^* , from the original $d_{s^*}^{b^*}$ to ϕ (the contract that is never accepted) is no longer necessarily the counterfactual constructed from assuming all the original contracting decisions except the contract between s^* and b^* were retained. This because the optimal response of b^* to the change in s^* 's offer may well include a change in its response to the offers from other sellers.

Let the decision of buyer b be the vector $d^b = (d_s^b, d_{-s}^b) \in [0, 1]^S$ where S is the number of sellers, and $d_s^B = (d_s^{b=1}, \dots, d_s^{b=B}) \in [0, 1]^B$ where B is the number of buyers. The argument above implies that the distribution of \mathbf{d}_{-s}^b conditional on the seller's information set is not independent of the seller's offer, i.e. of $d_b^s \in \mathcal{D}^s$. Assumption 1 implies

$$\begin{aligned} & \mathcal{E}[\pi^s(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, y(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, \mathbf{z})) | \mathcal{J}^s, d^s = (d_b^s, d_{-b}^s) \in \times_b \mathcal{D}^s] \\ & - \mathcal{E}[\pi^s(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, y(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, \mathbf{z})) | \mathcal{J}^s, d^s = (\phi, d_{-b}^s)] \geq 0. \end{aligned}$$

So if we observed or could construct random draws from the distribution of $(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, y(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, \mathbf{z}))$ conditional on $d^s = (\phi, d_{-b}^s)$, we could proceed as we did in the simultaneous move games analyzed above.

The problem is that we do not know how to construct a random draw from the distribution of $(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, y(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, \mathbf{z}))$ conditional on $d^s = (\phi, d_{-b}^s)$. This because without further assumptions we do not know how the buyer would change its responses to other sellers were it faced with a null contract from the seller. One way around this problem is to compute the minimum of $\pi^s(\cdot)$ over all possible choices buyer b could make given the observed realization of (d_{-s}^B, z) and then use the average of the difference between the realized

forms, the relaxation of optimizing behavior needed to rationalize the data is too large to be *a priori* reasonable.

profit and this minimized profit variable as the theoretical inequality we base estimation on. That is since

$$\begin{aligned} & \min_{d \in [0,1]^{s-1}} \pi^s(d_s^b = 0, d, d_{-s}^B, y(d_s^b = 0, d, d_{-s}^B, z)) \\ & \leq \pi^s(d_s^b = 0, d_{-s}^b, d_{-s}^B, y(d_s^b = 0, d, d_{-s}^B, z)) \end{aligned}$$

for every realization of $(\mathbf{d}_{-s}^b, \mathbf{d}_{-s}^B, \mathbf{z})$,

$$\mathcal{E}[\pi^s(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, y(\mathbf{d}_s^B, \mathbf{d}_{-s}^B, \mathbf{z})) | \mathcal{J}^s, d^s = (d_b^s, d_{-b}^s)] \quad (6)$$

$$- \mathcal{E}[\min_{d \in [0,1]^{s-1}} (\pi^s(d_s^b = 0, d, \mathbf{d}_{-s}^B, y(d_s^b = 0, d, \mathbf{d}_{-s}^B, \mathbf{z}))) | \mathcal{J}^s, d^s = (\phi, d_{-b}^s)] \geq 0.$$

To go from the inequalities in (6) to equation (5) we need an analogue of Assumption 3. Moreover empirical implementation requires the additional computational step of finding the required minima. On the other hand this strategy should work more generally in non-simultaneous move games. I.e. in non-simultaneous move games Assumption 2 must be amended to allow the distribution of some components of \mathbf{d}_{-i} conditional on $(\mathcal{J}_i, d_i = d)$ to depend on d . For those components we obtain our alternative profits by finding the required minima.

Generalization 3. There are a number of generalizations that can be incorporated without making any change in the inequalities taken to data.

Individual Effects. Additively separable individual effects that do not interact with the alternative chosen can be added to the returns functions π and r without affecting the inequalities in (5). This because unobserved factors whose effect on the agent's profits regardless do not depend on d are differenced out of (5). Note that this implies that the unobservable ν_2 need only capture the effects of omitted variables that impact on the change in profits in response to a change in d (this assumes $\delta = 0$ in our first generalization).

Mixed Strategies. If agent i plays a mixed strategy then Assumption 1 implies that each pure strategy with positive probability in the mixed strategy must have the same expected return. So our assumptions and notation apply to each decision made, and there is no need for the econometrician to specify whether the underlying strategies are pure or mixed. Of course if we knew mixed strategies were being played, and we could distinguish the mixed strategies associated with a particular information set, then there would be more information available then the information being used in our current inequalities.

Conditioning Sets and Heterogenous Primitives. The notion that \mathcal{J}_i denotes agent i 's information set at the time decisions are made is only used as motivation for Assumption 1. If Assumptions 1, 2, and 3 were known to hold for set of conditioning variables which were not the actual information set, then the required moment conditions could still be formed¹¹. Also we note that the $\pi(\cdot)$, and $r(\cdot)$ functions could be indexed by i as could the instruments (i.e. $x_{i,d,d'}$).

3 Estimation and Inference.

We provide details for the case where there is data on J markets indexed by $j = 1, \dots, J$. A market is a draw on (\mathbf{y}^j, x^j, d^j) where $\mathbf{y}^j \equiv \{\mathbf{y}_i^j\}_{i=1}^{n^j}$, and d^j and x^j are defined similarly. We will assume that the observed markets are independent draws from a population of such vectors with a distribution, say \mathcal{P} , that respects our Assumptions 1 and 2.

The estimation algorithm consists of constructing the sample analogues of the $M \equiv m \times h$ functions in equation (??) and finding the set of θ , say Θ_J , that minimize over $\theta \in \Theta$ (a compact subset of \mathcal{R}^K).

$$\left\| \left(\frac{1}{J} \sum_{j=1}^J \frac{1}{\#Q} \sum_{q \in Q} \sum_{k,d'} \chi_q(k, d_k, d') \Delta r^j(d, d', d_{-k}^j, y_k^j, \theta) \otimes h(x^j) \right) \right\|_- \quad (7)$$

where

$$f(\cdot)_- = \min[f(\cdot), 0],$$

and $\|f(\cdot)_-\|$ is a norm of $f(\cdot)$ (in the empirical examples we use the absolute value). Note that since our restrictions are inequalities, they may well be satisfied by many values of θ . That is, though Θ_J must contain at least a single point, it may well be equal to a larger set of points (more on this distinction below).

Let Θ_0 denote the set of parameter values that satisfy equation (??). In the literature on estimation subject to inequality restrictions, Θ_0 is often called the identified set. Under regularity conditions, one can show set-consistency of Θ_J for Θ_0 ; see Andrews, Berry and Jin (2004) and Chernozhukov, Hong, and Tamer (2003).

¹¹Of course insuring that Assumptions 2 and 3 are satisfied will put conditions on \mathcal{J}_i .

For our general case we only consider the problem of constructing confidence intervals that asymptotically cover (functions of) the true parameter (θ_0) with (at least) a given probability. The procedure that produces the confidence interval also produces a test of the null that there is a value of $\theta \in \Theta$ that satisfies all of our inequality constraints. However as shown in our empirical examples, for many cases we do not expect this test to be very powerful. Consequently we develop a more powerful test in the next section. The section thereafter comes back to a discussion of identification for the special case where the inequalities are linear in the parameters. In the linear case the identified set is convex and this enables us to simplify the discussion of identification considerably. For this case we also provide alternative, easy to construct, confidence intervals which are likely to be more informative than those presented earlier. The examples enable us to provide a deeper discussion of identification and to compare test statistics and confidence intervals in two settings of empirical interest.

3.1 Confidence Regions

In this section, we discuss a method of constructing confidence regions for our general case. The key step to this construction is in finding a (“critical value”) function of θ that exceeds the moments when both are evaluated at θ_0 with a given probability (asymptotically). This idea is first developed in Andrews, Berry and Jin (2004). They use the nonparametric bootstrap to find such a function. We describe how to obtain the confidence regions directly from an estimate of the variance of the data moments.

We will require notation for the sample moments of interest and their population counterparts. To this end let

$$m(y^j, d^j, x^j, \theta) = \frac{1}{\#Q} \sum_{q \in Q} \sum_{k, d'} \chi_q^j(k, d_k, d') \Delta r^j(d, d', d_{-k}^j, y_k^j, \theta) \otimes h(x^j)$$

be the M dimensional vector of moments from which the inequalities are constructed, and

$$m(P_J, \theta) = \frac{1}{J} \sum_{j=1}^J m(y^j, d^j, x^j, \theta).$$

The corresponding population moments are

$$m(\mathcal{P}, \theta) = \mathcal{E}m(\cdot, \theta),$$

and from section 2.2, the assumptions imply that

$$m(\mathcal{P}, \theta_0) \geq 0.$$

Define the variance of the population moments to be

$$\Sigma(\mathcal{P}, \theta) = \text{Var}(m(\cdot, \theta))$$

and denote its sample analogue by

$$\Sigma(P_J, \theta) = \frac{1}{J} \sum_{j=1}^J (m(y^j, d^j, x^j, \theta) - m(P_J, \theta))(m(y^j, d^j, x^j, \theta) - m(P_J, \theta))'.$$

Finally note that with this notation the estimation problem in (7) defines

$$\Theta_J = \arg \min_{\theta} \|m(P_J, \theta)_-\|.$$

We begin with the intuition for the test and confidence region construction. Consider a family of functions, say $\{m^*(P_J, \theta), \theta \in \Theta\}$, and a confidence region defined by

$$\Theta^{CI} = \{\theta : m^*(P_J, \theta) \geq 0, \theta \in \Theta\}. \quad (8)$$

By Assumption 1, $m(\mathcal{P}, \theta_0) \geq 0$, so if

$$m^*(P_J, \theta_0) \geq m(\mathcal{P}, \theta_0), \quad (9)$$

then $\theta_0 \in \Theta^{CI}$. The confidence region is built by finding functions $m^*(\cdot)$ such that the sufficient condition (9) occurs with probability approaching (at least) $1 - \alpha$, and using them to build the Θ^{CI} in equation (8). The test simply asks whether Θ^{CI} is the empty set.

All we need to construct our $m^*(\cdot)$ is: (i) a law of large numbers and a central limit theorem for the sample moment $m(P_J, \theta)$ when that moment is evaluated *at the point* $\theta = \theta_0$, and (ii) a consistent estimator of this variance of the sample estimator at that point.¹² More formally, we assume

¹²Note that we do not require the weak convergence of the empirical process $\{m(P_J, \theta)\}_{\theta \in \Theta}$. In particular, stochastic equicontinuity of the empirical process is not needed. Similarly, consistency of $\Sigma(P_J, \theta)$ is only required at the point θ_0 . Of course, any other consistent estimator of the variance would suffice.

Assumption 4 (a) $\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)] \rightarrow_d \mathcal{N}(0, \Sigma(\mathcal{P}, \theta_0))$,
and
(b) $\Sigma(P_J, \theta_0) \rightarrow_p \Sigma(\mathcal{P}, \theta_0)$. ♠

We now construct $m^*(\cdot)$. For a fixed θ , suppose $Z^*(\theta) \sim \mathcal{N}(0, \Sigma(P_J, \theta))$.¹³ Find a vector $\bar{z}_{J,\alpha}(\theta)$ that satisfies $\Pr(Z^*(\theta) \geq -\bar{z}_{J,\alpha}(\theta) | P_J) = 1 - \alpha$. There are many such vectors and below we suggest standardizing on a particular one that is easy to compute analytically or by simulation. Given a choice of $\bar{z}_{J,\alpha}(\theta)$, set

$$m^*(P_J, \theta) = m(P_J, \theta) + \frac{1}{\sqrt{J}} \bar{z}_{J,\alpha}(\theta). \quad (10)$$

and substitute this expression into the definition of Θ^{CI} in (8). Then Θ^{CI} is an asymptotic $(1 - \alpha)$ level confidence interval for θ_0 .

Theorem 1 *Let*

$$\tilde{\Theta}^{CI} = \{\theta : m(P_J, \theta) + \frac{1}{\sqrt{J}} \bar{z}_{J,\alpha}(\theta) \geq 0, \theta \in \Theta\},$$

and suppose Assumptions 2 and 3, and Condition 1 hold. Then

$$\underline{\lim}_{J \rightarrow \infty} \Pr\{\theta_0 \in \tilde{\Theta}^{CI}\} \geq 1 - \alpha. \quad \spadesuit$$

Proof. Define $m^*(P_J, \theta)$ as in (10),

$$\begin{aligned} \Pr(\theta_0 \in \tilde{\Theta}^{CI}) &= \Pr(m^*(P_J, \theta_0) \geq 0) \\ &\geq \Pr(m^*(P_J, \theta_0) \geq m(P, \theta_0)) = \Pr(\sqrt{J}[m^*(P_J, \theta_0) - m(P, \theta_0)] \geq 0) \\ &= \Pr(\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)] \geq -\bar{z}_{J,\alpha}(\theta_0)) \\ &\rightarrow 1 - \alpha \quad \spadesuit. \end{aligned}$$

Below we discuss easy ways of computing Θ^{CI} for problems that are linear in the parameters.

Corollary 1 *Under Assumptions 2 and 3, and Condition 1*

$$\overline{\lim}_{J \rightarrow \infty} \Pr\{\tilde{\Theta}^{CI} = \phi\} \leq \alpha,$$

where ϕ is the empty set. ♠

¹³Construction of $Z^*(\theta)$ is discussed at the end of section 3.2.

Remarks.

- A natural way of choosing $\bar{z}_{J,\alpha}(\theta)$ is to pick the same cut-off value for each component of the joint normal. This choice of normalization takes away a degree of freedom from the presentation of empirical results, making those results less arbitrary. Doubtless, other methods for constructing $\bar{z}_{J,\alpha}(\theta)$ may be preferable in given situations.¹⁴
- It should not be terribly surprising if Θ_J is a singleton, say θ_J ; i.e. if there is no (or only a single) value of θ that satisfies $m(P_J, \theta) \geq 0$. For example, if $m(\mathcal{P}, \theta_0) = 0$ the probability that $m(P_J, \theta_0)$ is not positive is one minus the probability that every element of an M vector of random variables is above its mean, a probability that can typically be made arbitrarily close to one by choosing M large enough (depending, of course, on the covariances of the moments). On the other hand if there is no value of θ that satisfies $m^*(P_J, \theta) \geq 0$ then that would be surprising, so much so that doubt would be cast on the basic specification.
- Finally, we note again that the confidence interval provided in Theorem 1 is conservative. Just how conservative depends on the properties of both the model and the data; a point discussed in more detail in empirical examples below. There are other ways of obtaining confidence intervals for θ_0 , and some of them are based on less conservative assumptions than the confidence intervals given above. One could, for example, adapt the suggestion in Andrews, Berry, and Jia (2004) to define confidence intervals that condition only on the binding constraints being satisfied. Shortly we consider problems which are linear in the parameter and show that in those problems there are natural, easy to simulate, distributional results available. The simulation procedure will generally produce a sharper confidence interval than the interval above.

¹⁴One alternative is to take the same cut-off after normalizing each component of the vector of moments by its standard error. That is, take the vector of standard deviations (from the component by component square root of the diagonal of $\Sigma(P_J, \theta)$) and denote it by $\sigma(P_J, \theta)$. Then for a scalar $\bar{z}(\theta)$, choose $\bar{z}_{J,\alpha}(\theta)$ to have the form $\sigma(P_J, \theta)\bar{z}(\theta)$.

3.2 Specification Analysis and Testing

There are a number of reasons why specification testing is likely to be particularly important in our context. This section points out three of them and then suggests a test which should be more powerful than the test provided in Corollary 1.

First, as noted above, the actual estimator the researcher uses will depend on the importance of unobservables that are known to the agent when decisions are made but not to the econometrician (ν_2). For every model that does allow for such a disturbance, there is a restricted version which does not and should provide for more efficient estimators. So often it will make sense to start out by testing whether it is necessary to allow for the structural errors.

Second the use of inequalities provides us with an ability to investigate whether any deviation from the null is likely to be due to the behavioral assumption (Assumption 1). Typically specification analyses focuses on the model's functional form or stochastic assumptions (Assumption 2 and Condition 1). The testing of the behavioral assumption limits the alternatives to those that are captured by increases in the δ parameter (which allows choices that cause returns to be less than $\delta\%$ below the optimal returns), or decreasing the number of choices that can be used for comparison (which, for instance, allow choices that are not too "distant" from the optimal choice). Of course this approach conditions on the functional forms and stochastic assumptions. We have not investigated the extent to which it is possible to distinguish between the two types of specification errors.

Finally the use of inequalities allows us to simplify certain aspects of more traditional specification analysis, especially in models with complex choice sets. This simplification occurs because one can now use techniques developed for the specification analysis of models with continuous unbounded outcomes in models with discrete or bounded outcomes. For example, the likely impact of a left out variable in models with discrete outcomes can be analyzed by projecting those variables down onto the included variables and analyzing the sign of the resulting projection coefficients (an analysis that is independent of the particular distributional assumptions made on the disturbances). The fact that the inequality estimators are easy to compute makes this type of specification analysis particularly useful (see the empirical examples below).

A Specification Test

If there is a value of $\theta \in \Theta_J$ for which $m(P_J, \theta) \geq 0$, any reasonable specification test will yield acceptance. However, as noted above, there are frequently good reasons to expect $\min_{\theta} \|m(P_J, \theta)_-\|$ to be different from zero even if the underlying model is a correct. Corollary 1 provides one test of this possibility. We now provide another which, at least in many cases, should be more powerful (see the empirical examples below).

The typical GMM specification test is based on the minimized criterion function value; i.e. it measures the distance between the sample moments and zero. With moment inequalities, a natural specification test of $H_0 : m(P, \theta_0) \geq 0$ vs. $H_1 : m(P, \theta_0) \not\geq 0$ would be based on the extent to which the inequalities are violated, or on $T_J \equiv \min_{\theta} \|(\sqrt{J}m(P_J, \theta))_-\|$.

In general T_J does not have a standardized limit distribution (i.e. it is not asymptotically pivotal), so to use this type of test one needs a method for obtaining appropriate critical values. First, note that under the null

$$\min_{\theta} \|(\sqrt{J}m(P_J, \theta))_-\| \leq \|(\sqrt{J}m(P_J, \theta_0))_-\| \leq \|(\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)])_-\|.$$

So for any ϵ ,

$$\Pr(T_J \geq \epsilon) \leq \Pr(\|(\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)])_-\| \leq \epsilon).$$

If θ_0 were known, the asymptotic distribution of this latter term could be approximated by simulating a normal distribution with mean zero and variance covariance $\Sigma(P_J, \theta_0)$ (the sample variance of the moment at θ_0), and then computing the norm of its negative part.

Since θ_0 is unknown, we consider an $1 - \alpha/2$ level confidence interval for it, denoted $CI_{1-\alpha/2}$. Assume we can construct a family of random variables indexed by θ (a stochastic process in θ), say $\{Z_J(\theta)\}$, with approximately the same distribution at each θ as $\{\sqrt{J}[m(P_J, \theta) - m(P, \theta)]\}$. Let $\bar{z}_{\alpha, J}(\theta)$ be the $1 - \alpha/2$ quantile of $Z_J(\theta)$ and $\bar{z}_{\alpha, J}$ be the supremum of these quantiles over the values of θ in a $1 - \alpha/2$ confidence interval, i.e.

$$\Pr\{\|(Z_J(\theta))_-\| \geq \bar{z}_{\alpha, J}(\theta)\} = \alpha/2, \quad \text{and} \quad \bar{z}_{\alpha, J} \equiv \sup_{\theta \in CI_{1-\alpha/2}} \bar{z}_{\alpha, J}(\theta).$$

Then,

$$\Pr\{T_J \geq \bar{z}_{\alpha, J}\} \leq \Pr\{\theta_0 \notin CI_{1-\alpha/2}\} + \Pr\{T_J \geq \bar{z}_{\alpha, J} \mid \theta_0 \in CI_{1-\alpha/2}\} \leq \alpha,$$

so $\bar{z}_{\alpha,J}$ is an α level confidence interval for T_J . More formally we have the following theorem.

Theorem 2 *Suppose (a) Assumption 4 holds; (b) $CI_{1-\alpha/2,J}$ is such that $\lim_{J \rightarrow \infty} \Pr(\theta_0 \in CI_{1-\alpha/2,J}) \geq 1 - \alpha/2$; and (c) $Z_J^*(\theta)$ is a stochastic process such that at each θ , $Z_J^*(\theta)|P_J \sim \mathcal{N}(0, \Sigma(P_J, \theta))$.*

Now define $\bar{z}_{\alpha,J} = \sup_{\theta \in CI_{1-\alpha/2,J}} \bar{z}_{\alpha,J}(\theta)$, where $\Pr^(\|Z_J^*(\theta)_-\| \geq \bar{z}_{\alpha,J}(\theta)|P_J) \leq \alpha/2$. Then under $H_0 : m(P, \theta_0) \geq 0$,*

$$\overline{\lim}_{J \rightarrow \infty} \Pr(\min_{\theta} \|(\sqrt{J}m(P_J, \theta))_-\| \geq \bar{z}_{\alpha,J}) \leq \alpha. \spadesuit$$

PROOF:

Define $c_{\alpha/2}$ by $\Pr^*(\|Z_J^*(\theta_0)_-\| \geq c_{\alpha/2}|P_J) = \alpha/2$. Now note that

$$\begin{aligned} & \Pr(\inf_{\theta} \|(\sqrt{J}m(P_J, \theta))_-\| \geq \bar{z}_{\alpha,J}) \\ & \leq \Pr(\|(\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)])_-\| \geq \bar{z}_{\alpha,J}) \\ & = \Pr(\|(\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)])_-\| \geq \bar{z}_{\alpha,J} \cap \{\bar{z}_{\alpha,J} \geq c_{\alpha/2}\}) \\ & \quad + \Pr(\|(\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)])_-\| \geq \bar{z}_{\alpha,J} \cap \{\bar{z}_{\alpha,J} < c_{\alpha/2}\}) \\ & \leq \Pr(\|(\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)])_-\| \geq c_{\alpha/2}) + \Pr(\bar{z}_{\alpha,J} < c_{\alpha/2}) \\ & \leq \Pr(\|(\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)])_-\| \geq c_{\alpha/2}) + \Pr(\theta_0 \notin CI_{1-\alpha/2,J}) \end{aligned}$$

The result follows by taking limits. \spadesuit

It still remains to construct $\{Z_J^*(\theta)\}$ and compute $\bar{z}_{\alpha,J}$. Perhaps the computationally simplest method for constructing $\{Z_J^*(\theta)\}$ and finding the associated $\bar{z}_{\alpha,J}$ is as follows. Take repeated draws on $\varepsilon^* \sim N(0, I)$. For each draw set $Z_J^*(\theta) = \Sigma(P_J, \theta)^{1/2}\varepsilon^*$. Now find the largest value of $\bar{z}_{\alpha,J}$ that is less than a fraction $\alpha/2$ of the values of $\sup_{\theta \in CI_{1-\alpha/2}} \|Z_J^*(\theta)_-\|$.¹⁵ As

¹⁵Note that Theorem 2 does not actually require weak convergence of the process $\sqrt{J}[m(P_J, \theta) - m(P, \theta)]$ to a Gaussian process (it only requires asymptotic normality at θ_0). We impose no conditions on the covariances of $\{Z_J^*(\theta)\}$ at different θ 's, ie $Cov(Z_J^*(\theta), Z_J^*(\theta'))$ is unrestricted. Any covariance process for components of $\{Z_J^*(\theta)\}$ will be sufficient as long as it doesn't violate existence of the process and satisfies the variance requirement given above. Consequently a natural alternative to the construction above would be to take $\{Z_J^*(\theta)\}$ as the Gaussian process with mean zero and covariance process given by the sample covariances evaluated at different θ .

we show in the next section this test becomes particularly simple when the underlying moments are linear. There are, however, other ways of computing test statistics for this problem, and we would like a method that obtains a critical value as close as possible to $c_{\alpha/2}$ (as defined in the proof of Theorem 2) with minimal computational burden.¹⁶

3.3 Inference for Linear Moment Models

We consider the special case where $m(P_J, \theta)$ is linear in θ , or¹⁷

$$m(P_J, \theta) = Z_J \theta - W_J, \quad \text{so} \quad \Theta_J = \operatorname{argmin}_{\theta \in \Theta} \|(Z_J \theta - W_J)_-\| \quad (11)$$

for a matrix, Z_J , and a vector, W_J , of sample moments, and θ is the parameter vector which is known to be in $\Theta \subset \mathcal{R}^K$. Analogously if $\mathcal{Z} \equiv \mathcal{E}Z_J$ and $\mathcal{W} \equiv \mathcal{E}W_J$, we note that $\mathcal{Z} \theta_0 \geq \mathcal{W}$, and define

$$m(\mathcal{P}, \theta) = \mathcal{Z} \theta - \mathcal{W}, \quad \text{and} \quad \Theta_0 = \{\theta : \mathcal{Z} \theta \geq \mathcal{W}, \theta \in \Theta\}.$$

Note that we have assumed the sign of \mathcal{W} , and then normalized its coefficient to unity. Since our model only delivers inequalities, we can only hope to estimate its parameters up to “scale” (up to multiplication by a positive constant). The choice of one for the coefficient of \mathcal{W} is a normalization which chooses that scale.

In this setting Θ_0 is the identified set and Θ_J is the corresponding set estimator. Under the assumption that Θ is compact and convex, Θ_0 and Θ_J are compact and convex also. The convexity of these sets simplifies the discussion of identification and consistency considerably. We focus on the problem of finding confidence intervals for components of θ (though, with a bit more notation, one could use analogous reasoning to find joint confidence regions for smooth functions of θ).

¹⁶There is a question of whether one could base a more powerful test on the $\bar{z}_{\alpha, J}(\theta)$. Clearly if one knew θ_0 a test which rejected if $T_J \geq \bar{z}_{2\alpha, J}(\theta_0)$ would be more powerful. In the empirical work below we shall also present $\bar{z}_{2\alpha, J}(\hat{\theta}_J)$ which should approximate the more powerful test statistic.

¹⁷It is straightforward to generalize the results in this section to models which are linear in nonlinear functions of a set of parameters, provided those functions are homogenous of some degree.

Let θ_k denote the k^{th} component of θ , and $\Theta_{k,0} = \{\theta_k : \theta \in \Theta_0\}$. Then the closed convexity of Θ_0 implies that $\Theta_{k,0}$ is a closed interval on \mathfrak{R} . We denote that interval by $\Theta_{k,0} = [\underline{\theta}_{k,0}, \bar{\theta}_{k,0}]$. Sample estimates of the k^{th} component bounds, $\underline{\theta}_{k,0}$ and $\bar{\theta}_{k,0}$, are available from Θ_J (and we provide an easy to use method to compute them below). To derive their properties we need notation for the mappings from a (Z, W) to the corresponding upper and lower bounds:

$$\begin{aligned}\underline{f}_k(Z, W) &= \min\{\theta_k : \theta \in \arg \min_{\tilde{\theta} \in \Theta} \|(Z \tilde{\theta} - W)_-\|\} \\ \bar{f}_k(Z, W) &= \max\{\theta_k : \theta \in \arg \min_{\tilde{\theta} \in \Theta} \|(Z \tilde{\theta} - W)_-\|\}.\end{aligned}$$

Then $\underline{\theta}_{k,0} = \underline{f}_k(\mathcal{Z}, \mathcal{W})$ and $\bar{\theta}_{k,0} = \bar{f}_k(\mathcal{Z}, \mathcal{W})$, and our estimates of these parameters are $\underline{\theta}_{k,J} = \underline{f}_k(Z_J, W_J)$ and $\bar{\theta}_{k,J} = \bar{f}_k(Z_J, W_J)$.

Note that (Z_J, W_J) is a sample average which will obey a law of large numbers and central limit theorem under familiar conditions. Below we provide sufficient conditions for the differentiability of \underline{f}_k and \bar{f}_k . Standard arguments then imply that our estimates of the bounds are consistent and asymptotically normal (see for e.g. Pakes and Pollard, 1989), and we provide the covariance matrix of their limit distribution. We then consider the conditions which might result in \underline{f}_k and/or \bar{f}_k not being differentiable, and briefly consider what might be done in that case.

Given differentiability, an analytic form for the parameters of the asymptotic distribution is available and one could use it to provide consistent estimates for those parameters. We suggest an alternative, simple way to approximate this limit distribution. Begin with simulation draws from a normal distribution centered at (Z_J, W_J) with covariance matrix equal to the sample covariance of these moments. Evaluate the bounds functions \underline{f}_k and \bar{f}_k at the values of the draws. Repeating this procedure, obtain a distribution for the bounds. The next theorem shows that this simulated distribution has the same limiting distribution as the limiting distribution of the estimated coefficients. So variances, confidence intervals, etc., can be taken directly from the simulated distribution.

In particular to find an asymptotic α level confidence interval for $\theta_{k,0}$ we look for numbers, $(d_k^-(\alpha), d_k^+(\alpha))$ such that

$$Pr \{ \theta_{k,0} \notin [\underline{\theta}_{k,J} - d_k^-(\alpha), \bar{\theta}_{k,J} + d_k^+(\alpha)] \} \leq \alpha,$$

where the probabilities are taken from the limit distribution of $(\underline{\theta}_{k,J}, \bar{\theta}_{k,J})$. Since

$$\begin{aligned} & Pr \{ \theta_{k,0} \notin [\underline{\theta}_{k,J} - d_k^-(\alpha), \bar{\theta}_{k,J} + d_k^+(\alpha)] \} \leq \\ & Pr \{ \underline{\theta}_{k,0} \leq \underline{\theta}_{k,J} - d_k^-(\alpha) \} + Pr \{ \bar{\theta}_{k,0} \geq \bar{\theta}_{k,J} + d_k^+(\alpha) \}. \end{aligned}$$

We construct an α level confidence interval by substituting values of $(d_k^-(\alpha), d_k^+(\alpha))$ that make the simulated probability of the latter event α .¹⁸

Lemma 1 in the appendix shows that the next assumption implies that there exists $\rho > 0$ such that for all (Z, W) with $\|(Z, W) - (\mathcal{Z}, \mathcal{W})\| < \rho$, $\underline{f}_k(Z, W)$ is continuously differentiable (and similarly for $\bar{f}_k(Z, W)$). Hence, this assumption provides enough smoothness to guarantee a standard limit theorem.

Assumption 5 (a) *The parameter space Θ is a bounded, convex polyhedron, i.e. it can be expressed as the intersection of a finite number of half spaces and is bounded.* (b) *The linear programs defining the boundary functions \underline{f}_k and \bar{f}_k and the duals to these programs, as defined in the Appendix, have unique nondegenerate solutions at $(\mathcal{Z}, \mathcal{W})$.*

Assumption 4(a) is likely stronger than necessary, but it allows us to take advantage of various findings in the mathematics of linear programming to prove our distributional result. In particular the convex polyhedron condition insures that the bound functions, \underline{f}_k and \bar{f}_k , can be expressed as

$$\begin{aligned} \underline{f}_k &= \operatorname{argmin} \theta_k \text{ s.t. } Z\theta \geq W \text{ and } \theta \in \Theta, \text{ and} & (12) \\ \bar{f}_k &= \operatorname{argmax} \theta_k \text{ s.t. } Z\theta \geq W \text{ and } \theta \in \Theta, \end{aligned}$$

So the bound functions are optimum values of linear programs, and the boundedness of the space insures that these values are finite. Note that many standard computing packages (e.g. matlab) have efficient routines to compute the solution to these programs, so the bounds generated by any given (Z, W) are easy to find. Assumption 4(b) is also likely stronger than necessary. It implies continuous differentiability of the boundary functions

¹⁸Note that this provides a confidence interval for the interval $[\underline{\theta}_{k,0}, \bar{\theta}_{k,0}]$ and hence for $\theta_{k,0}$. In this sense, one might find a shorter interval for just $\theta_{k,0}$. For more on this distinction see Imbens and Manski (2004).

in a neighborhood of $(\mathcal{Z}, \mathcal{W})$, but differentiability will sometimes hold under weaker conditions on the linear programs (see below).

Finally we need to insure that the sample averages, as well as the simulated sample averages, satisfy a central limit theorem. Recall that Z_J and W_J are sample averages. Let $Z_{j,J}$ and $W_{j,J}$ denote the j^{th} observations in the averages.

Assumption 6 For some $\zeta > 0$, $E\|(Z_{j,J}, W_{j,J})\|^{2+\zeta} < \infty$.

To examine the approximation to the limit distribution of $(\underline{\theta}_{k,J}, \bar{\theta}_{k,J})$, we require another bit of notation. Arrange the moments $(\mathcal{Z}, \mathcal{W})$ into the vector $\mathcal{S} = \text{vec}(\mathcal{Z}, \mathcal{W})$, and analogously let $S_J = \text{vec}(Z_J, W_J)$. Also, set $V_S = \text{var}(S_{j,J})$.

Theorem 3 Given Condition 1, and Assumptions 2, 5, and 6,

(a)

$$\sqrt{J} \begin{pmatrix} \underline{f}_k(S_J) - \underline{f}_k(\mathcal{S}) \\ \bar{f}_k(S_J) - \bar{f}_k(\mathcal{S}) \end{pmatrix} \xrightarrow{d} N(0, V)$$

where $V = \Gamma V_S \Gamma'$ and $\Gamma = (\nabla \underline{f}_k(\mathcal{S})', \nabla \bar{f}_k(\mathcal{S})')'$ the stacked partial derivatives of $\underline{f}_k(\mathcal{S})$ and $\bar{f}_k(\mathcal{S})$; and

(b) Let

$$S_J^* = S_J + \frac{V_S^{1/2} \epsilon^*}{\sqrt{J}} + o_{as}(1/\sqrt{J}) \quad (13)$$

where ϵ^* is a mean zero normal random variable with an identity covariance matrix (independent of the sample). Then, for almost every sample sequence (with empirical distribution P_J)

$$\sqrt{J} \left(\underline{f}_k(S_J^*) - \underline{f}_k(S_J), \bar{f}_k(S_J^*) - \bar{f}_k(S_J) \right) \Big| P_J \xrightarrow{d} N(0, V).$$

Proof of Theorem.

Let T_η denote the η neighborhood of \mathcal{S} given by the conclusion of Lemma 1. Also, let $f_k(\cdot) = (\underline{f}_k(\cdot), \bar{f}_k(\cdot))'$. All partial derivatives of f_k exist (and are continuous) on T_η . Let $\Gamma(S) = \begin{pmatrix} \nabla \underline{f}_k(S) \\ \nabla \bar{f}_k(S) \end{pmatrix}$ (so $\Gamma = \Gamma(\mathcal{S})$).

We prove conclusion (b), since conclusion (a) is standard and follows by analogous reasoning. A superscript ω will be used to denote a particular sample sequence. Let the $o_{as}(1)$ term in (13) be denoted τ_J . Let $A = \{\omega : \lim_{J \rightarrow \infty} S_J^\omega = \mathcal{S}, \lim_{J \rightarrow \infty} \tau_J^\omega = 0\}$. By the SLLN and the assumption that $\tau_J = o_{as}(1)$, $P(A) = 1$. Take $\omega \in A$, then there exists \bar{J} such that $S_J^\omega \in T_\eta$ for all $J \geq \bar{J}$. Let P^* denote the probability for $S_J^{*\omega}$ along the sample sequence given by ω . Then, $|S_J^{*\omega} - S_J^\omega| \xrightarrow{as^*} 0$ and $S_J^\omega \rightarrow \mathcal{S}$, so $S_J^{*\omega} \xrightarrow{as^*} \mathcal{S}$.

For $J \geq \bar{J}$,

$$\begin{aligned}
& \sqrt{J}(f_k(S_J^{*\omega}) - f_k(S_J^\omega)) \\
&= \sqrt{J}(f_k(S_J^{*\omega}) - f_k(S_J^\omega))\mathbf{1}\{S_J^\omega \in T_\eta\} \\
&= \sqrt{J}(f_k(S_J^{*\omega}) - f_k(S_J^\omega))\mathbf{1}\{S_J^\omega \in T_\eta\}\mathbf{1}\{S_J^{*\omega} \in T_\eta\} + o_{as^*}(1) \\
&= \sqrt{J}[\Gamma(S_J^\omega)(S_J^{*\omega} - S_J^\omega) + o(S_J^{*\omega} - S_J^\omega)]\mathbf{1}\{S_J^\omega, S_J^{*\omega} \in T_\eta\} + o_{as^*}(1) \\
&= \sqrt{J}\Gamma(\mathcal{S})(S_J^{*\omega} - S_J^\omega)\mathbf{1}\{S_J^\omega, S_J^{*\omega} \in T_\eta\} + o_{as^*}(1) \\
&= \sqrt{J}\Gamma(\mathcal{S})(S_J^{*\omega} - S_J^\omega) + o_{P^*}(1) + o_{as^*}(1) \\
&\xrightarrow{d} N(0, \Gamma V_S \Gamma')
\end{aligned}$$

where the first equality follows by the choice of ω and J ; the second from the fact that $S_J^{*\omega} \xrightarrow{as^*} \mathcal{S}$ ¹⁹; the third from the differentiability proven in Lemma 1(b); the fourth from an argument similar to the first, the continuity of $\Gamma(\cdot)$ at \mathcal{S} and the fact that $\sqrt{J}(S_J^{*\omega} - S_J^\omega) = O_{P^*}(1)$. The last step follows directly from the definition of $S_J^{*\omega}$. ♠

Remarks.

- The theorem states that if we find random variables $\{S_J^*\}$ that satisfy its conditions and substitute them into the linear program in (12), then the distribution of the solutions to that program will be identical to the limit distribution of $(\underline{\theta}_{k,J}, \bar{\theta}_{k,J})$. If \hat{V}_S is the sample covariance of S , then $\hat{V}_S \xrightarrow{as} V_S$. So we can set S_J^* to be $S_J + \frac{\hat{V}_S^{1/2} \epsilon}{\sqrt{J}}$. I.e., there

¹⁹Let $A^* = \{\omega^* : \lim_{J \rightarrow \infty} S_J^{*\omega^*}(\epsilon^{\omega^*}) = \mathcal{S}\}$. Then $P(A^*) = 1$. By the definition of A^* , for a given $\omega^* \in A^*$ there exists \bar{J}^* such that $|S_J^{*\omega^*}(\epsilon^{\omega^*}) - \mathcal{S}| < \eta$ for all $J \geq \bar{J}^*$. Hence, $\mathbf{1}\{S_J^{*\omega^*}(\epsilon^{\omega^*}) \notin T_\eta\} = 0$ for all $J \geq \bar{J}^*$ and $\sqrt{J}(f_k(S_J^{*\omega^*}(\epsilon^{\omega^*})) - f_k(S_J^\omega))\mathbf{1}\{S_J^\omega \in T_\eta\}\mathbf{1}\{S_J^{*\omega^*}(\epsilon^{\omega^*}) \notin T_\eta\} = 0$ for all $J \geq \bar{J}^*$. Then, $P(\{\omega^* : \lim_{J \rightarrow \infty} \sqrt{J}(f_k(S_J^{*\omega^*}(\epsilon^{\omega^*})) - f_k(S_J^\omega))\mathbf{1}\{S_J^\omega \in T_\eta\}\mathbf{1}\{S_J^{*\omega^*}(\epsilon^{\omega^*}) \notin T_\eta\} = 0\}) \geq P(A^*) = 1$, i.e. $\sqrt{J}(f_k(S_J^{*\omega^*}(\epsilon^{\omega^*})) - f_k(S_J^\omega))\mathbf{1}\{S_J^\omega \in T_\eta\}\mathbf{1}\{S_J^{*\omega^*}(\epsilon^{\omega^*}) \notin T_\eta\} = o_{as^*}(1)$.

is a natural way to construct the distribution of moments substituted into the linear program; take draws from a normal random variable centered at the sample moments with variance equal to the estimate of the variance of those moments.

- Theorem 3 can also be applied to two step estimators. Assume that $\tilde{\beta}$ is estimated in a preliminary stage and that $S_J = S_J(\tilde{\beta})$ where $S_J(\tilde{\beta}) = S_J(\beta_0) + \Upsilon_{\beta_0}(\tilde{\beta} - \beta_0) + o_p(1/\sqrt{J})$. Then, the expression for V_S is the asymptotic variance of

$$\sqrt{J}(S_J(\tilde{\beta}) - \mathcal{S}) = (1, \Upsilon_{\beta_0}) \begin{pmatrix} \sqrt{J}(S_J(\beta_0) - \mathcal{S}) \\ \sqrt{J}(\tilde{\beta} - \beta_0) \end{pmatrix} + o_P(1)$$

The asymptotic variance of the right-hand side expression above is usually obtained by writing $\tilde{\beta}$ in terms of its influence function, which can also be used to obtain an estimator of the desired variance-covariance.

- The theorem states that in the limit the simulated distribution will be the same as a normal distribution with (consistently) estimated mean and variance. However in finite samples the two distributions will, in general, be different. For example, the normal with estimated mean and variance can have positive probability of yielding values of the lower bound estimate larger than the upper bound estimate. The simulated distribution will not have this problem. More generally the simulated distribution uses a normal approximation for the distribution of the sample averages, and then finds the implied distribution for the nonlinear transformation of that normal that solves our linear programming problem. The normal with estimated mean and variance obtains its approximation by linearizing the solution to the linear programming problem. We think it likely that the normal approximation to means of data moments is more accurate than the normal approximation to the solution of an extremum problem. Also the simulation estimator may be appropriate for some problems that don't satisfy Assumption 5 (this is a subject we are currently exploring). We know, however, that there are cases, though, in a sense to be discussed below unlikely cases, where the simulation estimator will not converge to the true asymptotic distribution.

If Assumption 4(a) is satisfied but 4(b) is not, then the solutions to the linear programs in (12) still define \underline{f}_k and \overline{f}_k , but these bound functions

may not be continuously differentiable in a neighborhood of $(\mathcal{Z}, \mathcal{W})$. For Assumption 4(b) to be inappropriate these solutions must be either degenerate (when there are more than K inequalities going through the solution) or non-unique (when the solution to the linear program occurs at more than one point on the same inequality). One could view these cases as “knife-edge” cases and ignore them²⁰. However even in these cases directional derivatives may still exist at $(\mathcal{Z}, \mathcal{W})$, in which case though the limit distribution may not be normal, one may still be able to derive its analytic form and approximate it directly (we do not pursue this in this version of the paper).

Finally, we should point out that the specification test given in section 3.2 can also be used in the case with linear moments. Simulations from the stochastic process $Z_j^*(\theta)$ given in Theorem 2 take on a particularly simple form. Specifically, let $U_j^* = \frac{\hat{V}_S^{1/2} \epsilon^*}{\sqrt{J}}$, which is just S_j^* centered at zero. Note that U_j^* implicitly gives simulation draws on (Z_j, W_j) , ie $U_j^* = \text{vec}(Z_j^*, W_j^*)$, which in turn gives draws on the desired process $Z_j^*(\theta) = Z_j^* \theta - W_j^*$. From these simulation draws it is straightforward to obtain $\bar{z}_{\alpha, J}$ in Theorem 2.

4 Empirical Examples.

We now introduce our two empirical examples. One is an ordered choice problem while the other is a bargaining problem. In each case we begin by outlining the substantive problem. Next we describe our method of moments inequality estimators, discuss their properties and compare them to familiar alternatives. We conclude with a brief discussion of the empirical results

4.1 Ordered Choice.

This section is based on Ishii (2004). She analyzes how ATM networks affect market outcomes in the banking industry. The part of her study considered here is the choice of the number of ATMs. More generally the example shows how the techniques proposed in this paper can be used to empirically analyze multiple agent “lumpy” investment problems, or investment problems subject to adjustment costs which are not convex for some other reason²¹. We treat

²⁰In particular if it does occur for one choice of $(\mathcal{D}(d_i))$, it is unlikely to occur for another.

²¹Actually Ishii’s problem has two sources of non-convexities. One stems from the discrete nature of the number of ATM choice, the other from the fact that network effects

these problems as multiple agent ordered choice problems.

Ishii uses a two period model with simultaneous moves in each period. In the first period each bank chooses a number of ATMs to maximize its expected profits given its perceptions on the number of ATMs likely to be chosen by its competitors. In the second period interest rates are set conditional on the ATM networks in existence. Note that there are likely to be many possible Nash equilibria to this game, and one of the motivations for the study is to compare the observed network to other possible equilibria.

Ishii (2004) estimates a demand system for banking services and an interest rate setting equation. Both are estimated conditional on the number of ATMs of the bank and its competitors, i.e. on (d_i, d_{-i}) . The demand system has consumers choosing among a finite set of banks with consumer and bank specific unobservables (as in Berry, Levinsohn, and Pakes 1995). The indirect utility of the consumer depends on the distance between the consumer's home and the nearest bank branches, the consumer's income, interest rates on deposits, bank level of service proxies, the size of the ATM network, and the distribution of ATM surcharges (surcharges are fees that ATM users pay to an ATM owner when that owner is not the user's bank). Interest rates are set in a simultaneous move Nash game. This setup provides Ishii (2004) with the parameters needed to compute the banks' earnings conditional on its and its competitors ATM networks²².

To complete her analysis of ATM networks Ishii requires estimates of the cost of setting up and running ATMs. These costs are central to the public debate on alternative "market designs" for the ATM network (of particular interest is the analysis of systems that do not allow surcharges). This paper provides initial estimates of those costs, while Ishii (2004) provides robustness tests and considers the implications of the results.

4.1.1 The ATM Choice Model: Theory and Econometric Issues

To obtain the cost estimates we model the choice of the size of a network, that is the choice of $d_i \in \mathcal{D} \subset \mathcal{Z}^+$, the non-negative integers. In the simplest

can generate increasing returns to increasing numbers of ATMs

²²These earnings are calculated as the earnings from the credit instruments funded by the deposits minus the costs of the deposits (including interest costs) plus the fees associated with ATM transactions. The ATM fee revenue is generated when non-customers use a bank's ATMs and revenue is both generated and paid out when customers use a rival's ATMs.

version of the model that we begin with, we only attempt to estimate an average (across banks) of the marginal cost of buying and installing an ATM. Let

$$\pi(y_i, d, d_{-i}, \theta) = r(y_i, d, d_{-i}) - (\nu_{2,i} + \theta)d + \nu_{1,d,i}, \quad (14)$$

where $r(y_i, d, d_{-i})$ is the profits that would be earned in the second stage if the firm chose d and its competitors chose d_{-i} in the first stage, θ is the average (across banks) of the marginal cost of purchasing and installing ATM's, and the $\nu_{2,i}$ is the unobserved bank specific deviation from that marginal cost. ν_1 and ν_2 are defined as in section 2.2.

$r(\cdot)$ in equation (14) is obtained from the first stage of Ishii's analysis. Note that to find the returns that would be earned were $d \neq d_i$ (the firm's actual choice), we have to solve out for the equilibrium interest rates that would prevail were the alternative network chosen.

Clearly a necessary condition for an optimal choice of d_i is that expected profits from the observed d_i is greater than the expected profits from either $d_i - 1$ or $d_i + 1$. We use these two differences as our $\Delta\pi(\cdot)$.²³ So $m = 2$ and $\mathcal{E}\Delta\pi(\cdot)$ is the vector consisting of

$$\mathcal{E}[r(y_i, d_i, d_{-i}) - r(y_i, d_i - 1, d_{-i})|\mathcal{J}_i] - \theta - \nu_{2,i},$$

and

$$\mathcal{E}[r(y_i, d_i, d_{-i}) - r(y_i, d_i + 1, d_{-i})|\mathcal{J}_i] + \theta + \nu_{2,i}.$$

The simplicity of the model makes this a particularly good example for illustrating how inequality analysis works. Recall that we form moment conditions by interacting $\Delta\mathbf{r}(\cdot)$ with $h(x)$ (since ν_2 is mean independent of x , it averages out). Consider first using only the moment conditions generated by $h(x_i) \equiv 1$, i.e. by $\Delta\mathbf{r}(\cdot) \otimes 1$. Then the moment condition from the profitability difference that arises as a result of decreasing the value of d_i , or the change "to the left", is

$$m_L(P_J, \theta) = \frac{1}{J} \sum_j \frac{1}{n_j} \sum_i [r(y_i^j, d_i^j, d_{-i}^j) - r(y_i^j, d_i^j - 1, d_{-i}^j) - \theta] \quad (15)$$

²³These conditions will also be sufficient if the expectation of $\pi(\cdot)$ is (the discrete analogue of) concave in d_i for all values of d_{-i} . This condition which can not be checked without more detail on the model, but the realizations of profits evaluated at the estimated value of θ were concave in d_i for almost all banks.

$$= \frac{1}{J} \sum_j [\Delta \bar{r}_L^j(\cdot) - \theta] = \Delta \bar{r}_L - \theta,$$

where

$$\Delta \bar{r}_L^j(\cdot) \equiv \frac{1}{n_j} \sum_i [r(y_i^j, d_i^j, d_{-i}^j) - r(y_i^j, d_i^j - 1, d_{-i}^j)], \text{ and } \Delta \bar{r}_L \equiv \frac{1}{J} \sum_j \Delta \bar{r}_L^j(\cdot).$$

Analogously the moment condition from the profit change that would result from increasing the value of d_i^j , or the change to the right, is

$$m_R(P_J, \theta) \equiv \Delta \bar{r}_R + \theta. \quad (16)$$

The set of θ that minimize of the objective function in equation (7) are the values of θ that make both equations (15) and (16) positive. Since $(\Delta \bar{r}_L, \Delta \bar{r}_R)$ are the changes in revenue resulting from first and increase and then a decrease in the number of ATM's, we expect $\Delta \bar{r}_L$ to be positive while $\Delta \bar{r}_R$ should be negative. If $-\Delta \bar{r}_R < \Delta \bar{r}_L$ then

$$\Theta_J = \{\theta : -\Delta \bar{r}_R \leq \theta \leq \Delta \bar{r}_L\}$$

while if $-\Delta \bar{r}_R \geq \Delta \bar{r}_L$ there is a single θ which minimizes (7) and it is given by²⁴

$$\Theta_J = \{\theta_J = .5[-\Delta \bar{r}_R + \Delta \bar{r}_L]\}.$$

Increasing The Number of Instruments.

If we increase the number of instruments each new instrument produces a pair of additional inequalities (one for the change from the left and one for the change from the right). Indeed if h indexes instruments

$$\Theta_J = [\max_h \{-\Delta \bar{r}_{h,R}\}_h, \min_h \{\Delta \bar{r}_{h,L}\}_h],$$

where $\Delta \bar{r}_{h,R}$ provides the upper bound from the h^{th} instrument and so on. So Θ_J becomes shorter (weakly) as the number of instruments increases. Now there will be constraints that do not bind and our estimate of the lower

²⁴In the simple case where $h(x) \equiv 1$, if $r(\cdot)$ is concave in d_i then, at least in expectation, $\Delta \bar{r}_L \geq -\Delta \bar{r}_R$, so we do not expect the set of minimizers to be a singleton. Once we add instruments, however, concavity no longer insures that the inequalities will be satisfied by a set of θ values.

bound is the greatest lower bound while our estimate of the upper bound becomes the least upper bound.

The greatest lower bound is the maximum of a finite number of moments each of which distribute (approximately normally) about a separate $\theta_h < \theta_0$. By using this max as our estimator we should expect a positive bias in the binding constraint in finite samples (if h binds, $\hat{\theta}_h$ will tend to be larger than θ_h). This bias should increase with the number of inequalities. So when there are a large number of inequalities and some give lower bounds close to θ_0 we should not be surprised if the estimated lower bound is greater than θ_0 . Analogously, since the estimate of the upper bound is a minimum, it should not be surprising if the upper bound estimate is less than θ_0 . Of course, if the lower bound is greater than θ_0 *and* the upper bound is less than θ_0 , then the estimate Θ_J is just a point (even if the true Θ_0 is an interval). This accentuates the need for a test with good small sample properties²⁵.

Tests for the Presence of ν_2 .

Assume the x used as instruments are contained in the appropriate information sets and are uncorrelated with any unobserved cost differences known to the agents. Then the only difference between models with and without ν_2 is that in the model without ν_2 we can use the actual decision, or d , as an instrument, and in the model with a ν_2 use of d as an instrument would violate Assumption 3. Accordingly one way of determining the importance of ν_2 is to compare the test statistics from two estimation runs; one which uses d as an instrument and one which does not.

Increasing the Number of Parameters.

Change the specification so that the cost of setting up and operating an ATM equals $\theta_0 + \theta_1 x$ where x can be either bank or market specific. Again beginning with the case that $h(x) \equiv 1$ we have our two moment restrictions as

$$m_L(P_J, \theta) = \Delta \bar{r}_L - \theta_0 - \theta_1 \bar{x} \geq 0,$$

where $\bar{x} = J^{-1} \sum_j n_j^{-1} \sum_i x_i^j$, and

$$m_R(P_J, \theta) = \Delta \bar{r}_R + \theta_0 + \theta_1 \bar{x} \geq 0.$$

²⁵Similar issues arise in obtaining the upper and lower bounds to the distribution of values in independent private value auctions; see Haile and Tamer (2003). It suggests a role for a small sample correction, but that is a topic beyond the scope of this paper.

If we plot these two inequalities on a graph, their boundaries will be given by two parallel lines. If $\Delta\bar{r}_L > -\Delta\bar{r}_R$, then Θ_J , the estimate of Θ_0 , will be the area between the two parallel lines. If we add a covariance between the two differences and another instrument, say the number of branches, then provided Θ_J is not a singleton, it will be the intersection of the area between two sets of parallel lines with different slopes, or a parallelogram. If further moments are added, we obtain the intersection of the areas between a larger number of parallel lines. With three parameters we would look for the intersection between planes, and so on.

Boundaries.

If the choice set has a boundary that is chosen by some agents, then there may be moments which we can not construct for those agents (we do not have d' on one side of the boundary). Then the sample mean of the structural error converges to the expected value of the structural error conditional on not being at the boundary, and we have to check that the sign of that conditional expectation is negative as is required by Assumption 3.

In our example there are markets in which a number of banks chose not to purchase ATM's, and so do not have a change from the left. If we simply drop the $d_i = 0$ banks and form the average of $\Delta r_{L,d_i,i}$ among the firms with $d_i \geq 1$, then our inequality becomes

$$\mathcal{E}[\Delta r_{L,i}(\cdot)|\mathcal{J}_i, d_i \geq 1] \geq \theta + \mathcal{E}[\nu_{2,i}|\nu_{2,i} \leq -\Delta r_i(0, 1; \cdot) - \theta] \leq \theta.$$

So the fact that there is a boundary or corner on the left can indeed cause a violation of Assumption 3.

Note that if the important source of structural error were in sales and not costs, or if the boundary was from above rather than below, then the conditional expectation of ν_2 for those observations that were not bounded would have had the opposite sign. In these cases the boundaries do not violate our Assumption 3 and all one has to do to deal with the boundary is drop those observations which are constrained by it.

In cases where there is a boundary problem we should be able to get an indication of its likely magnitude by using a function of an instrument to select a subsample of firms for which there are likely to be no firms at the boundary, and redoing the estimation procedure. A large difference in the estimates indicates a need to modify the estimator to account for the boundary

problem. In these cases one can obtain a consistent estimate of the bound by substituting a random variable which is known to have the appropriate inequality relationship to the $\nu_{2,i}$ for the missing observations, and averaging across the full sample (in the ATM example, we could substitute a number which is larger than any reasonable ATM cost for the missing incremental revenues from the first ATM for the banks without ATM's).

Alternative Estimators: Ordered Choice Models.

In our notation the ordered choice model sets $\nu_1 \equiv 0$ in equation (14), assumes a particular distribution for ν_2 , and forms the likelihood of the observed d . It is one of two models traditionally used for such problems, and it does *not* allow for either expectational or measurement errors, or for firm specific fixed effects (none of which affect our inequality estimator).

Regardless of the distribution chosen, the ordered likelihood of any θ in our data is *minus infinity*. I.e. the ordered model can not be estimated. This occurs because if our “difference from the left” is less than our “difference from the right” for one or more observations there will be no value of $\theta + \nu_2$ that rationalizes the observed choices (if it was profitable to purchase the d^{th} ATM, the model says that it must have been profitable to purchase the next ATM). Note that as long as there is some uncertainty when decisions are made we should expect some agent's difference from the left to be less than its difference from the right even if all agents are behaving optimally.

One could modify the simple ordered model to allow for measurement error, and then form a likelihood that could be maximized. This, however, essentially “assumes away” a likely source of the problem (expectational error). Moreover to deal explicitly with expectational error we would have to assume a great deal more about the game and actually compute Nash equilibria conditional on values for the parameter vector (which would be nontrivial).

Alternative Estimators: First Order Conditions.

Hansen and Singleton's (1982) first order condition estimator can be applied to ordered choice problems if one is willing to ignore the discrete nature of our control. The stochastic assumptions used in Hansen and Singleton are the opposite of those required by the ordered choice model. The first order condition estimator assumes that there is no structural error ($\nu_2 \equiv 0$), and

attributes all differences in outcomes not explained by observables to ν_1 .

Given these assumptions and some mild regularity conditions, if the agents are maximizing with respect to continuous controls the first order condition for agents with a $d > 0$ must have an expectation of zero conditional on their information sets. As a result, provided $x \in \mathcal{J}$, a consistent estimator of θ can be found by minimizing

$$\left\| \frac{1}{J} \sum_j \frac{1}{n^j} \sum_i \{d_i^j > 0\} \left(\left(\frac{\partial r(y_i^j, d, d_{-i}^j)}{\partial d} \right) \Big|_{d=d_i^j} - \theta \right) \times h(x_i^j) \right\|.$$

There are two differences between these moment conditions and those that define the inequality estimator. First the inequality estimator uses a discrete analogue of the derivative; i.e. the derivative is replaced with inequalities from two discrete changes (one from the left and one from the right).²⁶ Whether or not this causes a substantial difference in the estimates is likely to depend on the ‘‘lumpiness’’ of the investment good.

Second, as originally formulated the first order condition estimator; (i) could use d as an instrument, and (ii) does not need to worry about boundaries. However we could reformulate the first order condition model to allow for an additive ν_2 error and chose instruments and treat boundaries precisely as we do for the inequality estimator. Then Hansen and Singleton’s (1982) estimator would retain its desirable properties.

4.1.2 Empirical Results

The dataset consists of a cross-section of all banks and thrifts in Massachusetts metropolitan statistical areas in 2002. A market is defined as a primary metropolitan statistical area, and the sample is small: it contains a total of 291 banks in 10 markets.²⁷ Our moment inequalities are derived

²⁶This assumes an inequality model with maximizing behavior and that the inequality estimator only uses the inequalities generated from the two adjacent possible choices. The first order condition model is not sufficiently flexible to estimate subject to the weaker behavioral assumptions we considered in our Assumption 1, and does not enable the researcher to add other inequalities to improve the efficiency of the estimator.

²⁷The data set is described in Ishii(2004), and is carefully put together from a variety of sources including the Summary of Deposits, the Call and Thrift Financial Reports, the 2000 Census, the Massachusetts Division of Banks, and various industry publications. The number of banks varies quite a bit across markets (from 8 in the smallest market to 148 in Boston), as does the number of distinct ATM locations per bank (which averages 10.1 and

as described above. The instruments used when we refer to the full set of instruments (and for the first order condition estimator) include a constant term, the market population, the number of banks in the market, and the number of branches of the bank (its mean is 6 and standard deviation is 15).

Table 1 contains the inequality estimators.²⁸ The first row provides the results when only a constant term is used as an instrument (the $h(x) = 1$ case). Then the estimator is an interval, $\Theta_J = [32,066, 32,492]$, but the interval is quite short. The confidence interval places the true θ_0 between \$23,300 to \$42,000 dollars with 95% probability.

Not surprisingly then when the rest of our instruments are added, the interval collapses to a point \$32,492, with a simulated confidence interval which shortens to \$29,431 to \$38,444. Given that the estimates in row 2 are point estimates we want to test whether the data is consistent with the inequalities holding, that is we want to use the test for $H_0 : m(P, \theta_0) \geq 0$ provided in Theorem 2.

The simulated distribution of the test statistic from two thousand simulation draws is described in figure 1 (which partitions the draws on the test statistic into twenty-five bins) and figure 2 (into fifty bins). With 25 bins it is hard to tell the difference between that distribution and a half normal; recall that the test takes its values from the negative parts of mean zero moments. In the 50 bin figure we see the differences between the simulated and half normal distributions generated by the fact that different moments will bind in different simulation draws.

The bottom rows provide the ratio of the value of our objective function to the simulated critical value of the test statistic when $\alpha = .05$. When the actual decision was not included in the instrument set, the ratio was .96. This accepts the null but is close to the critical value of one. Moreover the fact that the test has to allow for an interval of possible θ values decreases its power. So we looked at the ratio of T evaluated only at $\hat{\theta}$ to what the

has a standard deviation of 40.1). Since the number of banks per market varies so widely, we weighted our market averages with the square root of the number of banks in each market before averaging across markets (this generates a small improvement in confidence intervals).

²⁸All estimators for both empirical problems analyzed in this paper were obtained using the “fmincon” algorithm in Matlab. In the linear case, “fmincon” finds the *argmin* of $F(\theta)$ subject to the linear constraints $A\theta \leq B$. By setting $F(\theta) = \theta_k$ and then $F(\theta) = -\theta_k$ for the different components of θ we obtain the vertices of Θ_J . For details on the search method used in fmincon see <http://design1.mae.uf.edu/enkin/egm6365/AlgorithmConstrained.pdf>.

critical value of the α level test would be were we to assume $\hat{\theta} = \theta_0$. This ratio was .97, also less than one. Next we added the actual number of ATMs chosen to the instrument set. The test ratio then jumped to 1.36, indicating rejection at any reasonable significance level. Thus the data indicate that we should allow for unobserved cost components, but provide no reason to worry about the specification once we do.

Table 1: Inequality Method, ATM Costs*

	θ_J	95% CI for θ	
		LB	UB
1. $h(x) \equiv 1$	[32,006, 32,492]	23,301	41,197
2. $h(x) = \text{full } d > 0$	32,492	29,431	38,444
3. $h(x) = \text{full}, d \geq 0$	32,522	29,571	38,478
<i>Different Choices of $D(d_i)$ ($h(x) = \text{full}$)</i>			
4. $\{d : d - d_i = 2\}$	36,188	31,560	38,947
5. $\{d : d - d_i = 1, 2\}$	36,188	31,983	36,869
<i>Extending the Model ($h(x) = \text{full}$)</i>			
6. θ_b (in branch ATM)	36,649	32,283	38,871
7. θ_r (remote ATM)	38,348	26,179	47,292
<i>Test Statistics</i>		$d \notin IV$	$d \in IV$
T(observed)/T(critical at 5%)		.96	1.36

Table 2: First Order Conditions, ATM Costs*.

	Coeff.	Std. Error
θ_{01} (constant)	38,491	7,754
θ_0 (in-branch constant)	50,132	11,102
θ_1 (remote constant)	55.065	12,010

* There are 291 banks in 10 markets. The FOC estimator requires derivatives with respect to interest rate movements induced by the increment in the number of ATMs. We used two-sided numerical derivatives of the first order conditions for a Nash equilibria for interest rates.

Five percent of the observations have $d = 0$. In the first two columns we keep the inequality from the right for these observations, and simply drop

those observations in constructing the inequalities from the left. The banks that did not have ATM's were the smallest banks, and our estimates indicate that the returns to the first ATM is increasing in bank size. If we substitute the average of the returns from the first ATM's for the banks that did have ATM's for the unobserved returns for the banks that did not we get the estimates in the "full, $d \geq 0$ " row. As expected this increases the estimates, but by under 1%, indicating that boundary problems are likely to have only a minor impact on the empirical results.

Next we return to the model in Assumption 1 and assume that $D(d_i) = \{d : |d - d_i| = 2\}$ and $\delta = 0$. This allows agents to make ATM choices that are one ATM more or less than the optimal, but not more than one. The results using the weaker restriction on choices yield an estimate which is a little higher than the original estimate, but still well within the original confidence interval. When we consider alternatives that are one or two ATMS from the observed number, $D(d_i) = \{d : |d - d_i| = j \text{ for } j = 1, 2\}$, the estimate remains at \$36,188, but the length of the simulated confidence interval is now only \$31,983 to \$36,869.

Finally we consider if there is a difference in cost for "in-branch" and "remote" ATM locations. To do so the model is extended to allow for a choice of in-branch ATM's, say d_b , and remote ATM's, say d_r . The amended model has $\pi_i(\cdot, d) = r_i(\cdot, d) - d_b\theta_b - d_r\theta_r - \nu_i(d_b + d_r)$. We get point estimates for each cost, with θ_r about 5% higher than θ_b . However the confidence interval for θ_r covers that for θ_b , which, in turn, is similar to the confidence intervals we obtained when we did not differentiate branch locations²⁹.

The fact that the results when we set $D(d_i) = \{d : |d - d_i| = 2\}$ are similar to those when we set $D(d_i) = \{d : |d - d_i| = 1\}$, indicates that there is no reason to doubt that firms, at least on average, act optimally. This bodes well for the first order condition estimator that we now turn to, as that estimator can not be modified to allow for non-optimal choices. Table 2 shows the first order condition cost estimate to be \$38,491 with a standard error of \$7,754. \$38,491 is larger than the upper bound of the CI for the inequality estimator, but twice its standard error covers the inequality estimator's entire CI. When

²⁹We initially expected a cost advantage to in-branch locations. However on going back to the data we found that 16 banks own remote ATM's sites while having branches that lack an ATM; a fact which indicates either lower costs or greater benefits to remote ATM's for at least some banks. Also banks may find it optimal to install more, and/or more expensive, machines in their branches thus offsetting other branch cost advantages. Unfortunately we do not have the data nor the model needed to investigate these possibilities further.

we allow for a separate θ_r and θ_b , the estimates are larger than those obtained from the inequality estimator, but they are also much less precise. It seems that when we substitute first order conditions for inequalities we generate a less precise estimator, and possibly one with a positive bias.

Perhaps the most notable characteristic of the inequality estimators is how stable they were. Regardless of the set of instruments or the choice of alternatives to the observed choice, the average cost estimate is between \$32,000 and \$36,200 dollars.³⁰ When we considered richer models the confidence intervals do widen but this was to be expected given the size of the sample, and the point estimates from the richer model are similar to those from the parsimonious model. Ishii(2004) use these numbers to examine; the effect of surcharges on concentration in banking, the welfare impacts of alternative market designs conditional on the given ATM network, and the optimality of the size of that network.

4.2 Discrete Choice and Buyer/Seller Networks.

The section is based on Ho(2004a and 2004b). She analyzes the interactions between privately insured consumers, Health Maintenance Organizations (HMOs), and hospitals in the U.S. health care market. Ho considers a three stage game. In the last stage consumers choose an HMO given the networks of hospitals the HMOs have contracted with and the premiums set by the HMOs. In the second stage HMOs set their premiums in a Nash equilibrium which conditions on both consumer preferences and the hospital networks of the HMOs. The first stage sets contracts which determine which hospitals each HMO has access to and the transfers the hospitals receive for the services supplied to the HMOs they contract with.

This paper provides the analysis of the first stage, that is of the HMO-hospital contracting game. To do so we develop a framework capable of empirically analyzing the nature of contracts that arise in a market in which there are a small number of both buyers and sellers all of whom have some “market power”. Similar issues arise in the analysis of many markets where vertical relationships are important.

³⁰These estimates of costs are for costs over a six month period. Ishii (2004) notes that these are quite a bit higher than previous cost estimates and explains the difference by pointing out that the prior estimates leave out significant components of costs and are for smaller markets.

There are a number of ways to model buyer-seller interactions in these markets and then use Assumption 1 to provide inequalities which can be matched to data. We assume that sellers (or hospitals) make take it or leave it offers to buyers (or HMO's) of contracts that consist of a fixed fee and per patient markups. We then analyze the "reduced form" relationship between the contract parameters and buyer, seller, and market characteristics.

We make no attempt to uncover the structural model which determines the precise nature of contracts. Moreover there are likely to be many configurations of hospital networks that satisfy the resulting inequalities, and we do not investigate how this multiplicity of possible equilibria gets resolved (and, therefore, will not be able to perform counterfactual experiments). The hope is that the characterization of contracts obtained here will help determine the relevance of alternative more detailed models, and, to the extent policy and environmental changes do not effect the reduced form relationship per se, provide some idea of how changes in the environment are likely to impact on HMO/hospital transfers.

4.2.1 The Model

We begin with a brief overview of how the consumer demand for HMO's was obtained (for more detail see Ho, 2004). A consumer's utility from a hospital network conditional on the consumer having a given diagnosis is estimated from observed consumer hospital choices and a discrete choice demand system. The consumer's expected utility from a given network is then constructed as the sum of (demographic group specific) probabilities of various medical conditions times the utility the consumer gets from the network should it have a medical condition. The individual chooses its HMO as a function of this expected utility, the premiums the network charges, and other observed and unobserved plan characteristics. The function determining the utility from different HMO's is estimated from market level data on consumer's HMO choices (as in Berry, Levinsohn, and Pakes 1995).

Premiums are assumed to constitute a Nash equilibria conditional on the contracts that have been established between the HMOs and hospitals, the demand system, and the illness and hospital selection probabilities described above. Given premiums we construct each HMO's profits as premiums from the consumers who chose that HMO minus the costs of hospital care for those consumers, all conditional on the hospital networks of each HMO. We assume that these profits are uniquely determined (though the actual network

of hospitals each HMO contracts with need not be).

We adopt a parsimonious notation which does not do justice to the details of the model, but simplifies the exposition. Hospitals are indexed by h , HMO's by \mathbf{m} , the hospital network of HMO \mathbf{m} by $H_{\mathbf{m}}$ (this is just a vector of indicator functions which tell us whether there are contracts between the HMO and the various hospitals), and, analogously, the HMO network of hospital h by M_h . We let $q_{\mathbf{m}}(H_{\mathbf{m}}, H_{-\mathbf{m}})$ be the number of members of HMO \mathbf{m} , and $q_{\mathbf{m},h}(H_{\mathbf{m}}, H_{-\mathbf{m}})$ will be the number of patients HMO \mathbf{m} sends to hospital h . Then if $pr_{\mathbf{m}}$ is the premiums set by HMO \mathbf{m} and $T_{\mathbf{m},h}$ are the transfers it sends to hospital h , the HMO's profits are

$$\pi_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{-\mathbf{m}}) = pr_{\mathbf{m}}q_{\mathbf{m}}(H_{\mathbf{m}}, H_{-\mathbf{m}}) - \sum_{h \in H_{\mathbf{m}}} T_{\mathbf{m},h}, \quad (17)$$

while if c_h is the per patient cost of hospital h , its profits are

$$\pi_h^H(M_h, M_{-h}) = \sum_{\mathbf{m} \in M_h} T_{\mathbf{m},h} - c_h \sum_{\mathbf{m} \in M_h} q_{\mathbf{m},h}(M_h, M_{-h}). \quad (18)$$

We assume that the transfers implicit in the contracts are determined by a fixed fee, $fc(\cdot)$, a per patient markup $mk(\cdot)$, and, possibly, a disturbance known to both agents when they make their decisions (our ν_2). The actual contracts are largely proprietary, and when accessible are too complicated to summarize in a small number of variables. The purpose of the empirical exercise is to determine what the transfers implicit in them must have depended upon for the equilibrium we observe in the data to have satisfied Assumption 1. As a result we write $fc(\cdot) = fc(x, \beta)$ and $mk(\cdot) = mk(x, \beta)$ where x is a vector of HMO, hospital, and market characteristics, and β is a parameter vector to be estimated³¹.

Substituting for the transfers we write HMO profits as

$$\pi_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{-\mathbf{m}}, \beta) = r_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{-\mathbf{m}}, \beta) - \sum_{h \in H_{\mathbf{m}}} \nu_{2,\mathbf{m},h} + \nu_{1,H_{\mathbf{m}},H_{-\mathbf{m}}}, \quad (19)$$

where

$$r_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{-\mathbf{m}}, \beta) \equiv pr_{\mathbf{m}}q_{\mathbf{m}} - \sum_{h \in H_{\mathbf{m}}} fc(x_{\mathbf{m},h}, \beta) - \sum_{h \in H_{\mathbf{m}}} q_{\mathbf{m},h}(H_{\mathbf{m}}, H_{-\mathbf{m}})mk(x_{\mathbf{m},h}, \beta),$$

³¹Two previous papers analyze the marginal value of the hospital to a network conditional on the networks in existence, but do not attempt to analyze the payments from the hospitals to the HMO; see Capps, Dranove and Satterthwaite (2003), and Town and Vistnes (2001).

and hospital profits as

$$\pi_h^H(M_h, M_{-h}, \beta) = r_h^H(M_h, M_{-h}, \beta) + \sum_{i \in M_h} \nu_{2, M_h, M_{-h}} + \nu_{1, M_h, M_{-h}}, \quad (20)$$

where

$$r_h^H(M_h, M_{-h}, \beta) \equiv \sum_{i \in M_h} fc(x_{m,h}, \beta) - \sum_{i \in M_h} q_{m,h}(M_h, M_{-h})mk(x_{m,h}, \beta),$$

and the ν_2 and the ν_1 are as defined in section 2.2.

Note that because ν_2 is a disturbance in the transfers between HMO's and hospitals, each $\nu_{2,m,h}$ that appears in (20) for a given hospital appears with an opposite sign in (19) for the respective HMO. Expectational or measurement errors in membership, patient flows, and/or costs generate ν_1 ³².

Moment Inequalities When $\nu_2 \equiv 0$.

Recall that the model has hospitals making simultaneous take it or leave it offers to HMOs. Then Assumption 1 implies that HMO's accept or reject contract offers according as the offers increase or decrease their expected profits. With $\nu_2 = 0$ this implies that we can use the difference between our estimate of the HMO's profits from the observed HMO network and our estimate of what those profits would have been from the alternative network obtained by reversing the plan's contract choice with each of the hospitals in the market as the $\Delta\pi^M(\cdot)$ in our objective function (equation (7)). I.e. if the plan did contract with the hospital we compare to a situation in which it did not contract with the hospital, and vice versa³³.

The game is sequential, so without further assumptions the implications of Assumption 3 on the inequalities we can derive for the hospitals (i.e. the $\Delta\pi^H(\cdot)$ that we can use in our objective function) use the second generalization in section 2.4. I.e. the change in hospital profits between a contract

³²What we do assume away is a structural error that affects one participant in a contract but not the other; be it in the surplus or in the cost of contracting. However we could allow for a structural error with these properties if it had a negative expectation conditional on the instruments, and therefore satisfied our Assumption 3.

³³More precisely we use the change in profits from reversing the decision of each HMO with each of the six largest hospitals separately, and then formed one more HMO inequality by summing over the $\Delta\pi^M(\cdot)$ of the remaining hospitals. So $\Delta\pi^M(\cdot)$ had seven elements. The six largest hospitals by capacity cover an average of 57% of the admissions to hospitals.

that was accepted and the minimum profits the hospital could earn were the HMO to not accept the hospital's contract (the minimum being over the HMO's possible decisions with other hospitals) should be positive. On the other hand were we to assume the existence of a contract that would induce an HMO which accepted a hospital's offer to reverse its decision with that hospital without changing its decisions with other hospitals, then Assumption 1 would imply a stronger inequality; that the difference in hospital's profits from a network that includes an HMO which accepted its offer and one that does not is expected to be positive. This is an assumption, then, that we might want to test for.

Alternative Estimators: A Logit Model

We compare the results to those from a logit model. The logit model assumes that the plan chooses the network that maximizes its profits knowing the value of the unobserved disturbances in the profit equation. Profits for the different networks are calculated as in equation (19) and disturbances are assumed to distribute i.i.d. extreme value. Estimation is by maximum likelihood.

The assumptions on the disturbance term in the logit model are problematic. First they imply that there is no expectational or measurement error in the profit measure (in our terminology the assumption is $\nu_1 \equiv 0$). This only leaves the structural disturbance, but that disturbance can not be both independent of the observed determinants of profits and a determinant of the firm's decision (since those decisions determine profits). Accordingly we expect maximum likelihood to lead to inconsistent estimates.³⁴

4.2.2 The Data and Empirical Results

The primary data set contains every HMO in 43 major US markets, and considers the network of hospitals these HMOs offered to enrollees in March/April 2003. It has 451 plans and 665 hospitals, and contains a number of plan, market, and hospital characteristics put together from different data sources (for

³⁴Note also that the errors for the different possible networks consist of the sum of the errors for the hospitals in those networks. Since the same hospitals are in different networks it is unlikely that the composite error in the network choice equation is either independent across choices for a given HMO, or independent across different HMO's. As a result we should expect the disturbance in the profit equation to be correlated with the choices of the firm's competitor's as well as with its own choice, and the choices of the firm's competitors are also determinants of the firm's profits.

more detail, see Ho 2004b). As in the ATM example market size varies quite a bit, and we found that we obtain somewhat shorter confidence intervals when the market averages are weighted by the square root of the number of plans in the market before averaging across markets to form the moments used in estimation.

This is not a large data set and we have no theoretical restrictions on the form of the equations determining the transfers, so again we have to be somewhat parsimonious. We experimented with determinants of $fc(\cdot)$ and $mk(\cdot)$ suggested by standard IO and/or bargaining models. Those that mattered for the markup were; a constant term, a measure of the extent particular hospitals are likely to be capacity-constrained (obtained by calculating the number of patients treated at each hospital under the thought experiment that every plan contracts with every hospital in the market), and a measure of hospital costs per admission. Those that mattered for the fixed cost were; a dummy variable for whether the hospital negotiated in a system, and another for whether the network of the HMO excluded a same-system hospital³⁵.

All these variables except hospital costs are also used as instruments (hospital costs is not used as an instrument due to concerns over measurement error in this variable). The additional instruments used to construct $h(x)$ in equation (7) include other market and plan characteristics known to the agents at the time of the contracting decision³⁶. Finally the results presented below do not use the patient flows as instruments, as we wanted to allow for error in our model’s estimates of these flows. On the other hand when we ran

³⁵The hospitals that are “capacity constrained” are hospitals for which the predicted number of patients exceeds the number of beds \times 365 / average length of stay in the hospital. We note that when we estimated models allowing all the variables above to affect both the marginal and fixed costs, the individual coefficients ended up being insignificant, but there was little difference in the implications of the estimates and that Ho (2004b) considers robustness of the results to the inclusion of a number of other variables.

³⁶They include indicator variables for high proportion of population aged 55-64, a high proportion of the population with high cholesterol, a high number of beds per population, high proportion of hospitals in the market being integrated into systems of hospitals (see below), whether the plan is local, whether the plan has good breast cancer screening services, whether the plan has poor mental health services, and some of these characteristics interacted with the standard deviation of the distance between hospitals in the market (travel distance is a major determinant of hospital choice, and hence of surplus from a given system). Here low proportion means less than the mean percentile, except for beds per population and breast cancer screening rates where quartiles of the distribution were used.

the model adding these flows to the instruments there was very little change in either estimates or standard errors.

Estimates (without structural errors)

We began with assumptions that reduce the computational burden of obtaining the estimates. They assumed the existence of a contract that would induce an HMO which accepted a hospital's offer to reverse its decision with that hospital without changing its decisions with other hospitals (so we did not need to compute the minimums that insure our inequalities for the non-simultaneous move game; see generalization two above). Also though we allowed the plan membership and the patient flows to adjust for the alternative hospital networks formed by changing the contract status of one hospital (HMO), we did not allow the premiums to adjust (since this requires computing a new equilibrium for premiums and the premium adjustment for a change in one component of the network is likely to be small). Under these assumptions it took about ten minutes for a run of two hundred simulation draws on pentium three processor with a 1.33 GHZ hardrive and 512 MB of RAM, so it was easy to try numerous specifications. We then examined the robustness of the results to our simplifications.

Table 3 provides the base results and Table 4 presents a selection of (the many) robustness tests we ran. The estimate of Θ_0 from every specification was a singleton, i.e. there was no parameter vector that satisfied all the inequality constraints. All specifications have over eighty inequality constraints so this should not be surprising. Perhaps more telling is the fact that the value of the objective function was always less than .3 of the critical value of the simulated test statistic from our Theorem 2. So there is no indication of specification error, though the variance in the moment conditions in this example was quite large, so the test may not be too powerful.

The point estimates in Table 3 all have the expected sign. Four of the five coefficients are significantly different from zero at the 5% level, and the other, the constant term in the markup, was significant at the 10% level; but the confidence intervals are reasonably large. Figures 1 and 2 provide the simulated distributions for four of our coefficients. They seem right skewed with most of their mass between zero and a third of the upper bound to the confidence intervals.

Perhaps surprisingly the information on contractual relationships when combined with plan, hospital, and market characteristics seems to be enough

to identify variables that are correlates of the outcomes of the negotiations between hospitals and HMOs. I.e. hospitals in systems seem to take a larger fraction of the surplus and penalize an HMO that does not contract with all its members. Moreover markups are higher for capacity constrained hospitals and for hospitals with lower costs. Interestingly our point estimates imply an equilibrium configuration where lower cost hospitals get to keep only about a half of their cost savings.

Though, at least with the current sample size, our estimates of magnitudes are not very precise, the point estimates are consistent with the (little bit) of information from other sources at our disposal. The American Hospital Association (2001) reports that the total average cost of an admission is roughly \$11,000 (this includes interest and depreciation costs). The markups over these costs that we estimate varies by cost and type of hospital. For hospitals that are neither capacity constrained nor in a system, the point estimates imply very low average markups of about 2%. This is lower than the Kaiser Family Foundation Report (2004) estimates for community hospitals of 4.2%. However we estimate that capacity constrained hospitals receive an extra \$1440 per patient which translates into an average markup of approximately 13% of their costs. Hospitals that are not capacity constrained but are in systems capture \$179,000 in incremental profits per month per plan, which given their average patient load, translates into a markup of about 14% of costs. Also we are estimating a third of that as a penalty for excluding a hospital from a system, but this happens only rarely. Again these figures do have reasonably large confidence intervals associated with them, but they are, to our knowledge, the first estimates of the way the bargaining surplus is split between hospitals and HMOs that is available.

The most striking difference between our estimates and the logit estimates is that the latter indicate that we should be quite certain that hospitals in systems receive *lower* markups (the t-value for this coefficient is about seven). A natural explanation for this finding is the logit model's inability to account for endogeneity. Hospitals that are in systems are demanded by a disproportionate number of plans (the system is formed with demand patterns in mind and there is a penalty imposed when an in-system hospital is excluded). The logit model rationalizes these patterns by estimating that an HMO can contract with in-system hospitals at a lower cost. The logit model also estimates markups for capacity constraints and for dropping a hospital from the same system which are implausibly large in absolute value.

Table 3: Determinants of Hospital/HMO Contracts.
(See the notes to Table 4)

Characteristics of Hospitals	θ	Simulated		Logits	
		95% CI		θ	SE
Fixed Component (Units = \$/million, per month).					
SysHos	.14	.60	.07	-.57	.085
DropHos	.04	.23	.01	.63	.06
Per Patient Component (Units = \$/thousand, per patient).					
const	1.5	11.7	-3.9	.33	2.59
capcon	1.8	10.8	0.32	4.96	1.33
c_h	-.49	-.33	-1.5	-.53	.17

Table 4: HMO/Hospital Contracts: Robustness Analysis.

Variable	Star Hos.			Prem. Adj.			Enrollees		
	θ	Sim CI		θ	Sim CI		θ	Sim CI	
Enrole	-	-	-	-	-	-	.03	.11	-.02
Fixed Components (Units=\$/million, per month).									
SysHos	.16	.6	.11	.16	.65	.08	.32	.42	.09
DropHos	.045	.19	.02	.055	.24	.01	.12	.15	-.03
Per Patient Components (Units=\$/thousand, per patient).									
const	.13	2.3	-.8	9.0	16.2	-.57	.62	5.7	.08
capcon	1.5	10.2	-1.4	3.0	3.2	-2.3	2.1	5.1	-2.3
c_h	-.48	-1.4	-.33	-1.13	-1.8	-.65	-.58	-.41	-1.1
USnews	1.9	7.7	-9	-	-	-	-	-	-

Notes: Enrole refers to the total number of enrollees of the plan, “SysHos” to whether the hospital is in a system, “DropHos” to a system hospital in a network that does not include a same system hospital, “capcon” to the capacity constraint, and c_h to the cost per hospital admission.

Table 4 presents results from a selection of the robustness checks. When we allow premiums to adjust the magnitudes of the various components of the markup

(but not of the fixed cost) did change; the point estimates of the constant and the effect of capacity constraints are larger, and the cost term is more negative. However the sign and importance of the various variables, and the implications on overall markups, are not very different. The second set of robustness results accounts for the non-hospital related costs associated with plan enrollees. These costs clearly exist but the number of enrollees do not change very much when we change networks by one hospital (if they did not change at all they would act as the fixed effect in generalization three and could be omitted). Consequently it gets a positive, but imprecisely estimated coefficient, and generates an increase in confidence intervals for the other fixed cost components. Table 4's results are indicative of what we obtained when we tried other generalizations. Though precise magnitudes of coefficients did vary, the qualitative nature of the results did not.

4.2.3 Allowing For Structural Errors.

It is useful here to introduce some additional notation. Let H_m be the observed hospital network of HMO m , $H_m \setminus h$ be all hospitals in that network but hospital h , and $H_m \cup h$ be the network obtained when we add hospital h to H_m . Now define

$$\begin{aligned}\Delta\pi_m^M(+h, \beta) &= \pi_m^M(H_m, H_{-m}, \beta) - \pi_m^M(H_m \setminus h, H_{-m}, \beta) \\ \Delta\pi_m^M(-h, \beta) &= \pi_m^M(H_m, H_{-m}, \beta) - \pi_m^M(H_m \cup h, H_{-m}, \beta) \\ \Delta\pi_h^H(+m, \beta) &= \pi_h^H(M_h, M_{-h}, \beta) - \pi_h^H(M_h \setminus i, M_{-h}, \beta).\end{aligned}$$

and $\Delta r_m^M(-h, \beta)$, $\Delta r_m^M(+h, \beta)$, and $\Delta r_h^H(+m, \beta)$ analogously. Finally let $\chi(m, h)$ be the indicator function which takes the value of one if HMO m and hospital h contract, and zero elsewhere.

Assumption 1 implies that hospitals expect to increase their profit from signed contracts and that HMOs only reject offers when their expected profits are higher without them. Thus the expectation of

$$U^\pi(m, h; \beta) \equiv \chi(m, h)\Delta\pi_h^H(+m, \beta) + (1 - \chi(m, h))\Delta\pi_m^M(-h, \beta),$$

conditional on any $x \in \mathcal{J}_m \cap \mathcal{J}_h$ is positive.

The profit functions for the HMOs and the hospitals (equations (19,20)) imply

$$U^\pi(\mathbf{m}, h; \beta) =$$

$$\begin{aligned} & \chi(\mathbf{m}, h)[\Delta r_h^H(+\mathbf{m}, \beta) + \nu_{1, M_h, M_h \setminus \mathbf{m}} + \nu_{2, \mathbf{m}, h}] + (1 - \chi(\mathbf{m}, h))[\Delta r_{\mathbf{m}}^M(-h, \beta) + \nu_{1, H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h} + \nu_{2, \mathbf{m}, h}] \\ & = U^r(\mathbf{m}, h; \beta) + [\chi(\mathbf{m}, h)\nu_{1, M_h, M_h \setminus \mathbf{m}} + (1 - \chi(\mathbf{m}, h))\nu_{1, H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h}] + \nu_{2, \mathbf{m}, h}. \end{aligned}$$

So if $x \in \mathcal{J}_{\mathbf{m}} \cap \mathcal{J}_h$, and is an instrument in the sense that $\mathcal{E}[\nu_{2, \mathbf{m}, h}|x] = 0$, then $\mathcal{E}[U^r(\mathbf{m}, h, ; \beta)|x] \geq 0$, and $U^r(\cdot)$ can be one component of the $\Delta \mathbf{r}(\cdot)$ used to construct the moment inequalities in our objective function (equation (7)).

A second component can be obtained as the sum of HMO and hospital profits if they contract and zero otherwise. Assumption 1 insures that each component of the sum has positive expectation and since the sum does not depend on the transfers between these two agents it does not depend on $\nu_{2, \mathbf{m}, h}$ (though it does depend on transfer between both of them and other agents, and hence on β). I.e. if

$$S^r(\mathbf{m}, h; \beta) = \chi(\mathbf{m}, h)[\Delta r_h^H(+\mathbf{m}, \beta) + \Delta r_{\mathbf{m}}^M(+h, \beta)]$$

then it can be used for the second component of $\Delta \mathbf{r}(\cdot)$ in (7).

“Effects” Models.

Here we constrain the above model and assume $\nu_{2, \mathbf{m}, h} = \nu_{2, \mathbf{m}}$, that is that there are only hospital effects.³⁷ This assumption provides two more components for $\Delta \mathbf{r}(\cdot)$ in (7). First for every $\tilde{h} \notin H_{\mathbf{m}}$ and $h \in H_{\mathbf{m}}$ we have

$$\Delta \pi_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \cup \tilde{h}, \cdot) + \Delta \pi_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h, \cdot) =$$

$$\Delta r_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \cup \tilde{h}, \cdot) + \Delta r_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h, \cdot) + \nu_{1, \mathbf{m}, H_{\mathbf{m}}, H_{\mathbf{m}} \cup \tilde{h}} + \nu_{1, \mathbf{m}, H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h}.$$

Assumption 1 insures that each component of this sum is positive, and the sum itself does not contain $\nu_{2, \mathbf{m}}$, so

$$\mathcal{E} \left[\Delta \pi_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \cup \tilde{h}, \cdot) + \Delta \pi_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h, \cdot) | \mathcal{J}_{\mathbf{m}} \right] = \mathcal{E} \left[\Delta r_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \cup \tilde{h}, \cdot) + \Delta r_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h, \cdot) | \mathcal{J}_{\mathbf{m}} \right]$$

³⁷A more complete analysis of effects models in buyer seller networks, one which allows for both buyer and seller effects and considers the case with and without instruments, is given in Pakes, 2005.

is non-negative. If $\#$ denotes the cardinality of a set, this gives us $\sum_{\mathbf{m}} \#H_{\mathbf{m}}(\#H - \#H_{\mathbf{m}})$ “difference in difference” inequalities.

Also note that now

$$\begin{aligned} U^\pi(\mathbf{m}, h; \beta) &= \\ &= U^r(\mathbf{m}, h; \beta) + [\chi(\mathbf{m}, h)\nu_{1, M_h, M_h \setminus \mathbf{m}} + (1 - \chi(\mathbf{m}, h))\nu_{1, H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h}] + \nu_{2, \mathbf{m}}. \end{aligned}$$

So if we define

$$\bar{U}^r(\mathbf{m}) = \frac{1}{\#H} \sum_h U^r(\mathbf{m}, h),$$

then

$$\mathcal{E}[\bar{U}^r(\mathbf{m})|\mathcal{J}] \geq -\nu_{2, \mathbf{m}}.$$

Consequently for $h \in H_{\mathbf{m}}$

$$\begin{aligned} 0 &\leq \mathcal{E}[\Delta\pi_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h, \cdot)|\mathcal{J}] = \mathcal{E}[\Delta r_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h, \cdot)|\mathcal{J}] - \nu_{2, \mathbf{m}} \\ &\leq \mathcal{E}[\Delta r_{\mathbf{m}}^M(H_{\mathbf{m}}, H_{\mathbf{m}} \setminus h, \cdot)|\mathcal{J}] + \mathcal{E}[\bar{U}^r(\mathbf{m})|\mathcal{J}], \end{aligned}$$

providing us with another $\sum_h \#H_{\mathbf{m}}$ inequalities.

5 Conclusion

Appendix

Let $\underline{Q}(Z, W)$ denote the following linear program:

$$\min \theta_k \quad \text{s.t.} \quad Z\theta \geq W \quad \text{and} \quad \theta \in \Theta$$

and $\bar{Q}(Z, W)$ denote the following linear program:

$$\max \theta_k \quad \text{s.t.} \quad Z\theta \geq W \quad \text{and} \quad \theta \in \Theta.$$

The value of the program $\underline{Q}(Z, W)$ is $\underline{f}_k(Z, W)$ as defined previously, and similarly for $\bar{f}_k(Z, W)$. Under Assumption ??, $\underline{Q}(Z, W)$ has a unique optimal basis, which we will denote $\beta \subset \{1, \dots, h\}$. Also let $\underline{\theta}$ denote the solution to $\underline{Q}(Z, W)$.

Lemma 1 *Under Assumption 5, there exists $\eta > 0$ such that for all (Z, W) with $\|(Z, W) - (\mathcal{Z}, \mathcal{W})\| < \eta$,*

(a) β is the unique optimal basis for $\underline{Q}(Z, W)$ (and similarly for $\overline{Q}(Z, W)$); and

(b) $\underline{f}_k(Z, W)$ is continuously differentiable (and similarly for $\overline{f}_k(Z, W)$).

Proof:

If the conclusion does not hold, then there exists a sequence $(Z_n, W_n) \rightarrow (\mathcal{Z}, \mathcal{W})$ such that β is not the unique optimal basis for $\underline{Q}(Z_n, W_n)$. Without loss of generality, we can assume (Z_n, W_n) is close enough to $(\mathcal{Z}, \mathcal{W})$ to ensure that $\underline{Q}(Z_n, W_n)$ is feasible (i.e. that there is a value of θ which solves this problem). The constraints $\theta \in \Theta$ and Assumption 5 imply that if $\underline{Q}(Z_n, W_n)$ is feasible then a bounded solution exists. Since β is not the unique optimal basis for $\underline{Q}(Z_n, W_n)$, there exists a different basis α_n that is optimal. Also let θ_n denote the solution corresponding to the basis α_n . Note that α_n is a sequence that can only take on a finite number of possible values. The sequence α_n must then have at least one limit point in this finite set of values. For any such limit point, say α , there exists a subsequence n' such that $\alpha_{n'} = \alpha$. Now note that $\theta_{n'}$ is a sequence in the compact space Θ and so must have a convergent subsequence, $\theta_{n''} \rightarrow \theta^*$.

The unique solution to $\underline{Q}(\mathcal{Z}, \mathcal{W})$ is bounded, then by Goldfarb and Todd (1989) Theorem 4.2(b), the dual program to $\underline{Q}(Z_n, W_n)$ has a bounded optimal solution. Since the solution to the dual is also unique, it follows that that solution is bounded. Under these conditions, Martin (1975) Theorem 1.1, shows that the value of the program, \underline{f}_k , is continuous at $(\mathcal{Z}, \mathcal{W})$. Hence, $\theta_k^* = \underline{\theta}_k (= \underline{\theta}_{k,0})$.

For a matrix Z , let Z^α denote the matrix consisting of the rows of Z in α . Similarly $Z^{-\alpha}$ denote the matrix consisting of all the rows of Z not in α . By the definitions of α and β ,

$$\mathcal{Z}^\beta \underline{\theta} = \mathcal{W}^\beta \text{ and } \mathcal{Z}^{-\beta} \underline{\theta} > \mathcal{W}^{-\beta} \quad (21)$$

and

$$Z_{n''}^\alpha \theta_{n''} = W_{n''}^\alpha \text{ and } Z_{n''}^{-\alpha} \theta_{n''} \geq W_{n''}^{-\alpha}.$$

Taking limits,

$$\mathcal{Z}^\alpha \theta^* = \mathcal{W}^\alpha \text{ and } \mathcal{Z}^{-\alpha} \theta^* \geq \mathcal{W}^{-\alpha}.$$

Thus, θ^* is a feasible solution to $\underline{Q}(\mathcal{Z}, \mathcal{W})$. Since $\theta_k^* = \underline{\theta}_k$, θ^* is also optimal. By uniqueness, we must have $\theta^* = \underline{\theta}$. Since $\alpha \neq \beta$, there exists an element l such that $l \in \alpha$ and $l \notin \beta$. But then $\mathcal{Z}^l \theta^* = \mathcal{Z}^l \underline{\theta} = \mathcal{W}^l$, which contradicts the strict inequality in (21). The conclusion (a) follows for $\underline{Q}(Z, W)$ and the argument for $\overline{Q}(Z, W)$ is symmetric.

The conclusion (b) follows almost immediately from (a). Given the unique optimal basis β , the unique optimal solution to $\underline{Q}(Z, W)$ for (Z, W) in the η -neighborhood of $(\mathcal{Z}, \mathcal{W})$ is given by $\theta = (Z^\beta)^{-1} W^\beta$. Then, $\underline{f}_k(Z, W) = e_k' (Z^\beta)^{-1} W^\beta$, where e_k is the vector with a one in the k^{th} component and zero elsewhere. Since Z^β is nonsingular in this neighborhood, \underline{f}_k is clearly continuously differentiable. Similarly, the result holds for \overline{f}_k . ♠

References.

- American Hospital Association Annual Survey Database: Fiscal Year 2001.
- Andrews, D., Berry, S., and P. Jia (2004), "Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location," manuscript, Yale University.
- Bajari, P., Benkard, L., and J. Levin (2004), "Estimating Dynamic Models of Imperfect Competition", manuscript, Stanford University.
- Bajari, P., Hong, H., and P. Ryan (2004), "Identification and Estimation of Discrete Games of Complete Information", manuscript, Duke University.
- Berry, S. (1992); "Estimation of a Model of Entry in the Airline Industry", *Econometrica*, vol. 60, no. 4, pp. 889-917.
- Steve Berry, Jim Levinsohn, and Ariel Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, vol. 63, no. 4, pp. 841-890.
- Bickel, P. and D. Freedman (1981), "Some Asymptotic Theory for the Bootstrap," *Annals of Statistics*, 9, 1196-1217.
- Bresnahan, Timothy and Peter Reiss (1991): "Entry and Competition in Concentrated Markets", *Journal of Political Economy*, vol. 99, no. 5. pp. 977-1009.

- Chernozhukov, V., Hong, H., and E. Tamer (2003), "Parameter Set Inference in a Class of Econometric Models," manuscript.
- Cliberto, and E. Tamer (2004), "Market Structure and Multiple Equilibria in the Airline Markets," manuscript.
- Dove Consulting, 2002 ATM Deployer Study, Executive Summary, February 2002.
- Fershtman, C. and A. Pakes, 2004, "Dynamic Games with Assymmetric Information; A Computational Framework", *mimeo*, Harvard University.
- Haile, P. and E. Tamer, 2003, "Inference with an Incomplete Model of English Auctions", *The Journal of Political Economy*, 111, 1-51.
- Hansen, Lars, 1982, "Large Sample Properties of Method of Moments Estimators", *Econometrica*, 50, 1029-1054.
- Hansen, Lars Peter, and Kenneth J. Singleton, 1982, "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models", *Econometrica*, vol. 50, no. 5, pp. 1269-86.
- Ho, K., (2004a), "The Welfare Effects of Restricted Hospital Choice in the US Medical Care Market", *mimeo*, Harvard University.
- Ho, K., (2004b), "Insurer-Provider Networks in the Medical Care Market", *mimeo*, Harvard University.
- Horowitz, J. and C. Manski (1998), "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281-302.
- Imbens, G. and C. Manski (2003), "Confidence Intervals for Partially Identified Parameters," manuscript.
- Ishii, Joy, 2004, "Interconnection Pricing, Compatibility, and Investment in Network Industries: An Empirical Study of ATM Surcharging in the Retail Banking Industry", *mimeo*, Harvard University.
- Kaiser Family Foundation report, "Trends and Indicators in the Changing Health Care Marketplace", 2004.
- Manski, C. (2003), *Partial Identification of Probability Distributions*, Springer: New York.

Pakes, A. and D. Pollard, 1989; "Simulation and the Asymptotics of Optimization Estimators" *Econometrica* vol 57, pp. 1027-57.

Pakes, A., M. Ostrovsky, and S. Berry, 2003; "Simple Estimators for the Parameters of Discrete Dynamic Games (with Entry-Exit Examples)" *National Bureau of Economic Research WP* 10506.

Pakes, A., 2005; "Theory and Econometrics in Empirical I.O.", The Fischer-Schultz lecture, World Congress of the Econometric Society, *mimeo* Harvard University.

Seim, K.(2002);"Geographic Differentiation and Firms' Entry Decisions: The Video Retail Industry", *mimeo*, GSB, Stanford.