Essays on The Industrial Organization of Medical Care

A thesis presented

by

Katherine Emily Ho

to

The Committee of Business Economics

in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the subject of Business Economics

Harvard University
Cambridge, Massachusetts

May, 2005
Abstract

Essays on the Industrial Organization of Medical Care

Katherine Emily Ho

This thesis combines three essays on the industrial organization of medical care. The first two essays consider the restrictions imposed by health insurers on their enrollees' choice of hospitals. The network of hospitals offered by each insurance plan affects consumer welfare, hospital profits and the incentives faced by hospitals to invest in new capacity, technology and quality. The first essay uses a three-step econometric model to predict consumer preferences over health insurers conditional on the hospitals they offer. The results indicate that consumers place a positive and significant weight on their expected utility from the hospital network when choosing plans.

The second essay uses the demand estimates as one input to a model of supply. A simple profit maximization model explains roughly half the observed contracts between insurers and hospitals. A generalization of the model demonstrates an additional effect: hospitals that do not need to contract with all plans to secure demand
(for example, hospitals that are capacity constrained under a limited or selective network) can force insurers to compete for contracts. Some plans may exclude these hospitals in equilibrium. I estimate the expected division of profits between insurers and different types of hospitals using data on insurers’ choices of network. Hospitals that have merged to form systems, and those that are expected to be capacity constrained, capture significantly higher markups than other providers. I show that these high markups imply a negative incentive for hospitals to invest to remove capacity constraints despite a significant benefit to consumers from the investment. The results demonstrate one route by which hospitals may modify their characteristics in order to increase their bargaining power.

The third essay develops a methodology to analyze multiple-agent discrete choice problems where there may be both endogenous regressors and multiple equilibria. This approach is used to estimate plan and hospital profits in the second essay. The identifying assumption is that firms expect their choices to lead to higher returns than those from their alternative options. The approach is a modified method of moments algorithm where the moment conditions hold as inequalities at the true value of the parameter vector.
Contents

Abstract iii

Contents v

Acknowledgements vii

Chapter 1: Introduction 1

1.1 Contracting in the Medical Care Market 2

1.2 A Methodology Using Moment Inequalities 5

Chapter 2: The Welfare Effects of Restricted Hospital Choice 7

2.1 Introduction 7

2.2 Industry Background 11

2.3 Summary of the Estimation Approach 15

2.4 The Data 18

2.5 Details on the Empirical Approach 26

2.6 Results 39

2.7 Welfare Implications of Selective Contracting 48

2.8 Discussion and Conclusion 54

2.9 Appendix A: Details on Hospital and Plan Data 57
Acknowledgements

This dissertation is the result of several years’ work. Many people contributed directly and indirectly its completion; too many to acknowledge directly here. My thanks go in particular to my dissertation committee. Ariel Pakes has been a major influence on the way I think about empirical research. I am still surprised by his willingness to devote so much time and interest to my numerous questions and concerns. Without his insights chapter four would probably never have been completed. Our other co-authors, Jack Porter and Joy Ishii, also contributed significantly to that chapter. I am indebted to David Cutler and Nancy Beaulieu for taking on the task of introducing me to the US medical care sector. David’s valuable advice helped me understand the merits of the previous health economics literature and find a productive approach to my own research. I am also grateful for his willingness to read and comment on my drafts, particularly in the early stages of the process. Thanks to Nancy for helping me struggle with questions of intuition and identification, giving me a much clearer understanding of the models I use than would have emerged without her.

Finally, I am grateful to my husband Victor for his unfailing support and enthusiasm. He now knows far more than he ever wished to about endogeneity, unobservables and instruments. I could not have completed this process without him.
Chapter 1: Introduction

This dissertation contains three essays on the industrial organization of medical care. Chapters 2 and 3 examine the contracts agreed upon by hospitals and health insurers. These contracts determine the choice of hospitals available to enrollees in each insurance plan and therefore have a significant effect on consumer surplus. Chapter 2 studies the demand side of the market. It estimates a model of consumer demand for health insurance plans given the set of hospitals offered by each plan. Chapter 3 considers supply. I provide some descriptive statistics on the formation of insurers’ hospital networks and then use data on the observed contracts to estimate the expected division of profits between plans and different types of hospitals. The final chapter, which is joint work with Ariel Pakes, Jack Porter and Joy Ishii, is methodological. We develop a methodology for analyzing discrete choice problems with multiple agents where there may be both multiple equilibria and endogenous regressors. This is the estimation procedure used in Chapter 3 to estimate plan and hospital profits.

The rest of this introduction provides an overview of each of the research areas, touching on motivations, methodologies and key results.
1.1 Contracting in the Medical Care Market

The effects of managed care health insurers on the price and quality of medical care have been widely researched\(^1\). One aspect of their impact, however, has not been addressed in detail: the restriction imposed by each insurance plan on the network of hospitals from which its enrollees can choose. Chapters 2 and 3 examine the causes and welfare effects of these restrictions. I introduce a new dataset that lists the hospitals in the network of every managed care plan in 43 markets across the US. In that dataset roughly 10% of all hospital-insurer pairs fail or refuse to agree on contracts. 76% of managed care enrollees have a restricted choice of hospitals. I use the data first in the demand estimation of Chapter 2 and then in Chapter 3 to analyze the supply side of the market. I outline these two analyses in turn below.

1.1.1 An Overview of the Demand Estimation

The Model

Chapter 2 estimates a model of consumer demand for hospitals and health insurers and uses it to consider the effect of hospital choice restrictions on consumer welfare. I use a three step econometric model to predict consumer preferences over plans conditional on the hospitals they offer.

In the first step I estimate demand for hospitals using a standard multinomial logit model, treating hospitals as bundles of characteristics and allowing for observed differences across consumers. A particular consumer’s utility from treatment at a particular hospital is assumed to vary with the hospital’s characteristics and with

\(^{1}\text{For example, Miller and Luft (1997) review fifteen studies of the effects of managed care on quality. They find no compelling evidence of a reduction in quality of care, although patients show less satisfaction with managed care than with traditional plans. Cutler, McClellan and Newhouse (2000) consider the causes of the expenditure reductions achieved by managed care plans in the treatment of heart disease. They show that virtually all the difference in spending between indemnity plans and HMOs comes from lower unit prices rather than the quantity of services or a difference in health outcomes.}\)
the consumer’s income, location and diagnosis. The second step takes the estimated parameters and uses them, together with predictions of the probability of each diagnosis for each type of consumer, to predict expected utilities provided by each plan’s hospital network. Finally, the resulting expected utility variable is used as an input to the health plan demand model. Here I use a methodology similar to that suggested by Berry, Levinsohn and Pakes (1995), treating health insurers as bundles of characteristics that include premiums, clinical quality measures, the size of the physician network, and the value of the hospital network to consumers.

Results

The results indicate that consumers place a positive and significant weight on their expected utility from the hospital network when choosing their plans. The coefficient magnitudes imply that a one standard deviation increase in expected utility is equivalent to a reduction in premium of $39 per member per month (a little less than one standard deviation). A welfare analysis demonstrates that a move from the observed hospital networks to an unselective outcome, where all plans contracted with all hospitals, would result in a gain of approximately $1.04 billion per year to consumers and a loss of just $0.8 million to firms across the 43 major US markets considered, assuming fixed premiums.

1.1.2 Understanding the Causes of the Observed Contracts

The Methodology

Chapter 3 considers the supply side of the market. First I use the demand estimates from Chapter 2 to predict the producer surplus (defined as the profits to be divided between the plan and the hospitals in its network) generated by each potential hospital network for each plan in the data. I find that approximately 50% of contractual decisions are explained by this definition of producer surplus. A model of the bargaining
process by which hospital prices are determined rationalizes much of the remaining data. I predict that hospitals that do not need to contract with all insurance plans to secure demand (for example, hospitals that are capacity constrained under a limited or selective network) can force insurers to compete for contracts. Some plans may exclude these hospitals in equilibrium. Hospitals can merge to form “systems”, which may also affect bargaining between hospitals and insurance plans.

The second empirical analysis in this chapter estimates the profits secured by different types of hospitals using insurers’ observed choices of hospital networks and data on the characteristics of both insurers and providers. I assume that plans choose their networks in a simultaneous-moves game conditional on their expectations regarding other plans’ choices and the prices that will be paid to hospitals. The methodology employed here is described in detail in Chapter 4.

**Results and Conclusions**

Hospitals in systems are found to capture markups of approximately 14 per cent of revenues, in contrast to non-capacity constrained, non-system providers whose markups are about 2 per cent. They also impose high penalties on plans that exclude their partners. Providers that are expected to be capacity constrained capture markups of about 15 per cent of revenues. I show that these high markups imply a negative incentive for hospitals to invest to remove capacity constraints, despite a median benefit to consumers of over $330,000 per new bed per year from the change.

The results of this chapter provide evidence that hospitals can increase their markups and their overall profits by reducing their capacity below the level they would choose in a world without bargaining over prices. The chapter therefore demonstrates one avenue by which providers may distort their characteristics in order to increase their bargaining power. The under-investment in capacity that results has a significant negative effect on consumer surplus.
1.2 A Methodology Using Moment Inequalities

Chapter 4, which is joint work with Ariel Pakes, Jack Porter and Joy Ishii, develops a methodology to analyze multiple-agent discrete choice problems where there may be both endogenous regressors and multiple equilibria. This is the methodology used to estimate plan and hospital profits in Chapter 3.

The Methodology

The chapter considers the following problem. The econometrician observes a set of choices made by multiple agents and wishes to analyze the determinants of these choices. The analysis may be complicated by the existence of multiple potential equilibria and by possible problems with endogenous regressors. The approaches used to address these issues in previous papers are feasible only for problems involving small numbers of firms. The method developed here is much simpler to implement.

The identifying assumption used for estimation is the simple necessary condition that each agent expects the choice it makes to lead to higher returns than those from its alternative options. The returns of one agent may be affected by the decisions of the others. We calculate the sample average of the difference in observable returns between the actual choice and the firm’s alternative options, assuming that each firm’s returns can be calculated up to a parameter vector of interest and an additive error. The methodology accepts parameter vectors that make the difference non-negative. The approach is a modified Method of Moments algorithm, where we require the moment conditions at the true value of the parameter vector to hold as inequalities rather than equalities.

The approach addresses the issue of multiple equilibria simply by assuming that the observed outcome constitutes a Nash equilibrium in each market considered. We do not attempt to understand how this equilibrium was chosen from the set of potential outcomes. We will therefore not be able to perform counterfactual experiments.
However, the approach has two major advantages over previous methodologies. First, it is simple to implement, implying that we can analyze problems that have until now been too complex to consider in any detail. Second, we make fewer assumptions than previous methods, for example regarding the information structure of the problem and the distribution of the error terms.

**Empirical Examples**

We discuss two empirical examples. The first is an ordered choice problem: it analyzes how banks choose the size of their ATM networks taking into account the effect of these networks on demand. In the simplest version of the model we estimate an average (across banks) of the marginal cost of buying and installing an ATM. The simplicity of this framework makes it a good example for illustrating how inequality analysis works, and we describe it in some detail. We compare the results to those from two alternative estimators that have been used for ordered problems: ordered probit or logit and first order condition estimators that ignore the discrete nature of the choice.

The second example is the analysis of the hospital-health plan network problem considered in detail in Chapter 3. We illustrate how this problem fits into the overall framework of the new methodology. We also show how the results differ from those of alternative methodologies, such as the logit estimator, that ignore the issues of endogeneity and multiple equilibria. The empirical results from both examples demonstrate the importance of accounting for these issues.
Chapter 2: The Welfare Effects of Restricted Hospital Choice

2.1 Introduction

Managed care health insurers in the US medical care market restrict their enrollees to visiting hospitals within specific networks. The network offered by each insurer affects consumer welfare, hospital profits and the incentives faced by hospitals to invest in new capacity, new technology and quality. However, there is very little literature on the allocation and impact of plans’ hospital networks, constrained largely by a lack of data on plan contracts. This chapter introduces a new dataset that lists the hospital networks of every managed care plan in 43 markets across the US, making possible an analysis of the phenomenon. The dataset demonstrates that there is significant variation across both plans and markets in the extent to which plans exclude major hospitals from their networks. On average 10% of potential plan-hospital pairs in my data fail to arrange contracts. The proportion varies from 0% in some markets to as many as 50% in others. I define selective markets as those in which at least four of the five major plans fail to reach agreement with at least one major hospital: by this
In this chapter I investigate the effects of selective contracting on consumer welfare. This requires deriving an estimate of consumer demand for health plans conditional on the network of hospitals they offer. The analysis is conducted in three steps. First I estimate a discrete choice model of demand for hospitals, taking into account consumer attributes such as location, diagnosis and income. Identification comes from the variation in individuals’ hospital choice sets across markets. The second step is to use the estimated parameters from this demand system to find each type of consumer’s expected utility from the hospital network offered by each plan in his or her market. Finally, this expected utility variable is included as one input to a discrete choice model for health plans. Here identification comes from the variation in plan choice sets across markets, and from variation in the quality of the hospitals offered by each plan both within and across markets.

I use the demand estimates to predict the total welfare generated by the set of health plans in each market under two scenarios: first, that plans offer their observed hospital networks, and second, that all plans contract with every hospital in the market. This can be interpreted as a comparison of the observed to the first-best outcome, at least in terms of consumer surplus. I find, assuming fixed prices, that the change would lead to a median equivalent variation (the dollar spending needed to compensate consumers for a move back to the observed networks) of $9.89 per person per year. This translates to a total gain of $1.04 billion for the 43 markets in the data. The variation across markets is high: there are four markets in which the equivalent variation is over $100 per person per year and fourteen in which it is under $1 per person per year. Producer surplus effects are small and negative. The change to unrestricted hospital choice would result in a median producer surplus reduction of just $19,000 per market per year, translating to a $0.80 million loss across the markets in the sample. The consumer surplus effects clearly dominate, implying an
increase in total welfare from the change under the assumption of fixed premiums\textsuperscript{2}.

I am not the first to estimate demand for hospitals or for health plans. The hospital choice model in this chapter follows the existing hospital choice literature closely, analyzing encounter-level data and estimating the parameters using maximum likelihood techniques\textsuperscript{3}. In particular, steps 1 and 2 of my analysis (the hospital choice model and creation of the expected utility variable) are closely related to the innovative papers of Capps, Dranove and Satterthwaite (2003) and Town and Vistnes (2001). The plan choice literature is less fully developed, however. Several papers consider the effect of premium changes or quality data on employer and/or consumer choices of health plans\textsuperscript{4}; most of these use reduced form analysis. I advance the literature by explicitly connecting the three key players in this sector of the market - consumers, plans, and hospitals - rather than considering just two of the players and making assumptions or using reduced form analysis to bring in the third, as previous papers have done\textsuperscript{5}. The new dataset mentioned above, collected specifically for this chapter, makes this analysis possible\textsuperscript{6}.

Several caveats to the results should be noted. In particular the demand analysis

\textsuperscript{2}Given the fixed premium assumption, my results essentially estimate the effect of restricted hospital choice on consumer surplus, ignoring the effect of the restrictions on plan bargaining power and therefore on prices. This issue is discussed further in Section 2.8.


\textsuperscript{4}Buchmueller and Feldstein (1997), Beaulieu (2002) and Chernew, Gowrisankaran et al (2004 (a) and (b)) are examples.

\textsuperscript{5}For example, Town and Vistnes (2001) and Capps, Dranove and Satterthwaite (2003) both estimate consumer preferences over hospitals, then regress hospital profits or prices on variables summarizing consumer demand for that hospital. They find that profits increase with value to the consumer, implying that plan willingness-to-pay increases with this measure. In order to investigate not prices but the existence of contracts, I model the impact of consumer demand for the hospital on plan demand explicitly.

\textsuperscript{6}Other important players include primary care physicians, who significantly impact the consumer’s choice of hospital given his plan, and employers, who select the menu of plans from which the consumer chooses. Data limitations imply that their contribution cannot be modelled explicitly.
yields no information on whether the full-choice outcome is in fact an equilibrium for profit-maximizing plans, or if so, on the extent to which premiums would increase in a switch to the new equilibrium. In spite of these issues, however, the results reported here have implications for policy. They provide new evidence regarding the importance of increased hospital choice for consumer welfare, even in markets where consumers already have access to a large subset of the major hospitals in the area. The chapter therefore adds another dimension to the arguments concerning the optimal level of regulation in the medical care market: regulation would be needed to maximize consumer choice of providers if competing plans were unable to reach this equilibrium independently. Further research is clearly needed both to confirm the results of this chapter and to understand the set of potential equilibria more deeply. Chapter 3 models the process by which the equilibrium is determined; the demand estimates provided by this chapter are a key input into that analysis.

Finally, this chapter offers a framework for analyzing other related demand systems. For example the effect of physician network size, or the number of drugs included in a formulary, on demand for health plans could be modeled in an analogous way. Similar demand systems also exist in other industries, where downstream firms such as distributors or retailers contract with networks of upstream firms and may restrict consumer choice in order to reduce costs or avoid competition. The methods used in this chapter may prove useful in these other situations.

This chapter continues as follows. In the next section I describe relevant aspects of the industry and summarize the variation in networks across plans and markets. The demand estimation is outlined in Section 2.3; Section 2.4 describes the data set; and Section 2.5 gives details on the estimation strategy. Demand results are given in Section 2.6. Section 2.7 analyzes the welfare effects of selective contracts and Section 2.8 concludes.
2.2 Industry Background

2.2.1 Firm and Consumer Decisions

Each year, every privately insured consumer in the US chooses a health plan, generally from a menu offered by his employer\(^7\), and pays that plan a monthly premium in return for insurance coverage. The insurer contracts with hospitals and physicians to provide any care needed during the year. When the consumer requires medical care, he may visit any of the providers listed by the health plan, and receive services at zero charge or after making a small out-of-pocket payment. There is some variety in the restrictiveness of different types of managed care plan. If an individual is insured by a Health Maintenance Organization (HMO) he may visit only the hospitals in that plan’s network. Point of Service (POS) plan enrollees can visit out-of-network hospitals but only if referred to them by a Primary Care Physician. Preferred Provider Organizations (PPOs) and indemnity plans are the least restrictive insurers: enrollees do not need a PCP referral to visit an out-of-network hospital, although PPOs may impose financial penalties for doing so, for example in the form of increased copayments or deductibles. The focus of Chapters 2 and 3 is on HMO and POS plans, since their network choices have the strongest effect on both consumers and hospitals. 53\% of the privately-insured population was enrolled in an HMO/POS plan in 2000.

Throughout the analysis I assume that hospital and plan decisions regarding the type and quality of products and services to offer, and also the network of hospitals offered by each plan, are exogenous to the consumer’s choice of insurer and provider. Two additional assumptions simplify the estimation. First, I reduce the two-stage process by which employers choose a menu of plans and then consumers choose a plan from that menu into a single stage representing a "joint" employer/consumer choice.

\(^7\)58\% of the population is insured through an employer, while only 5\% purchase insurance independently. (See the website www.statehealthfacts.kff.org).
There is a small literature that investigates how employers make their decisions, and how these decisions relate to consumer preferences\(^8\), but since my data does not identify employers I am forced to ignore this question. Second, I also assume that health plan choices are made at the individual, not the family, level. I have access only to aggregate plan data; while I could have included the distribution of family size in my estimation, this would have complicated the model substantially.

### 2.2.2 Variation in Size of Hospital Networks

The new dataset introduced in this chapter, which lists the hospital network of every HMO and POS plan in 43 markets across the US, demonstrates significant variation across both markets and plans in the extent to which plans exclude major hospitals from their networks. The data were collected from individual plan websites; missing data were filled in by phone. Figure 1 documents the variation. Markets are categorized on a scale from 1 to 5, where 1 is the least selective, indicating that each of the 5 largest plans (by enrollment) contracts with all 8 largest hospitals (by number of admissions). In markets ranked 5, at least 4 of the largest plans exclude at least one major hospital; the other categories lie between these extremes\(^9\). Markets are fairly evenly spread across the five categories: 16 markets are ranked 1 or 2 (not selective) and 21 are ranked 4 or 5 (very selective). The figure also shows the distribution of plans by the number of major hospitals excluded, and the variation in this distribution across types of market. Plans’ selective behavior varies widely: 217 plans exclude no major hospitals, but 62 plans exclude at least 4 of the 8 major hospitals in their markets. Plans on average exclude more hospitals in selective markets than

---

\(^8\)See, for example, Chernew, Gowrisankaran et al (2004(a)), which uses a dataset listing the health plans available to and offered by employers across markets to examine the effect of quality information on employers’ choices of health plans.

\(^9\)Examples of markets ranked 1 include Baltimore MD and Atlanta GA. Markets ranked 5 include New Orleans LA and Portland OR.
elsewhere, but even within markets there is considerable variation in plan networks. This variation will be a major advantage in terms of identification when I come to estimate the health plan choice equation. For example, as discussed below, it will be important to include market fixed effects to control for cross-market variation in the quality of the outside option. I will be able to do this, while still identifying the coefficient on expected utility from the hospital network, because of the variation in hospital choice sets across plans within each market.

The dataset contains no exclusive contracts (either hospitals reaching agreement exclusively with a single insurer or vice versa), and very few vertically integrated organizations. Many hospitals and health plans attempted vertical integration in the 1990s but this has become increasingly rare in recent years; the literature implies that the breadth of skills needed to run both a hospital and an insurer is too large for the vertically integrated model to be viable except in very specific circumstances\(^\text{10}\). The key exception to this pattern is Kaiser Permanente, a dominant HMO in California and elsewhere that owns a large number of hospitals but contracts with few outside its own organization. I include Kaiser health plans and hospitals in the demand estimation and the welfare calculations: they are important members of the plan and hospital choice sets, particularly in California, and any change in their hospital networks could have significant effects on consumer surplus.

\(^{10}\)See, for example, Burns and Pauly (2002) and Burns and Thorpe (2001).
This figure summarizes the variation in selectivity of plans’ hospital networks both across and within markets. Markets are categorized on a scale from 1 to 5, where 1 is the least selective. Markets are fairly evenly distributed across the categories, as shown in the following table.

<table>
<thead>
<tr>
<th>Market Category</th>
<th>Definition</th>
<th>Number of markets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The 5 largest plans (by enrollment) contract with all 8 largest hospitals (by admissions)</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>One plan excludes at least one hospital</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>Two plans exclude at least one hospital or three plans exclude one hospital each</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>Three plans exclude at least one hospital; one of them excludes more than one</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>Four or more plans exclude at least one hospital</td>
<td>8</td>
</tr>
</tbody>
</table>

Number of major hospitals excluded by each plan in selective markets (dark bars, categories 4-5 in the table above) compared to unselective markets (pale bars, categories 1-2 in the table)

![Figure 1: Variation in Plan Networks Across and Within Markets](image.png)
2.3 Summary of the Estimation Approach

My main objective in estimating demand is to understand to what extent consumer utility from health plans is affected by the set of hospitals offered by each plan in the market. The analysis requires me to take three steps: these are summarized in Figure 2\textsuperscript{11}.

<table>
<thead>
<tr>
<th>Step of Estimation</th>
<th>Data Inputs</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hospital demand</td>
<td>MEDSTAT: encounter-level data on indemnity/PPO patients’ characteristics and hospital choice in 11 markets, 1997-98 AHA: hospital characteristics</td>
<td>Estimated effect of hospital characteristics on consumer utility given age, gender, diagnosis and location</td>
</tr>
<tr>
<td>2. Expected utility</td>
<td>New dataset listing network of each HMO/POS plan in 43 markets, Q1 2003 Estimated parameters from Step 1</td>
<td>Expected utility of every consumer type in the market from every plan’s network of hospitals</td>
</tr>
</tbody>
</table>

Figure 2: Overview of the Three-Stage Estimation Process

\textsuperscript{11}In principle it would be preferable to estimate the two sets of parameters jointly. Unfortunately this is not possible due to the scale of the problem: convergence would not be achieved within reasonable time limits.
First, I estimate demand for hospitals using a standard multinomial logit model and allowing for observed differences across individuals\textsuperscript{12}. With some probability consumer i becomes ill at time t in market m. His utility from visiting hospital h given diagnosis l is given by:

\begin{equation}
    u_{ihlt} = u(x_{ht}, \eta_h, \nu_{il}/\alpha, \beta)
\end{equation}

where \( x_{ht}, \eta_h \) are vectors of observed and unobserved hospital characteristics respectively, \( \nu_{il} \) are observed characteristics of the consumer such as diagnosis and location, and \( (\alpha, \beta) \) are the coefficients on the specification\textsuperscript{13}. There is no outside option in the hospital choice equation: sick patients are assumed always to go to hospital. Consumers choose hospitals to maximize their utility, so that if consumer i with diagnosis l chooses hospital h then for all other

\begin{equation}
    u_{ihl} = u(x_h, \eta_h, \nu_{il}/\alpha, \beta) \geq u_{ih'l} = u(x_{h'}, \eta_{h'}, \nu_{il}/\alpha, \beta)
\end{equation}

This maximization produces the set \( A_h \) of \( \nu \) that choose hospital h. Thus shares are given by:

\begin{equation}
    s_h(x, \eta/\alpha, \beta) = Pr(\nu \in A_h)
\end{equation}

Details of the estimation process are set out in Section 2.5.1. The next step is to take the estimated parameters \( \alpha, \beta \) and use them, together with predictions of the probabilities of diagnoses for each type of consumer, to predict expected utilities

\textsuperscript{12}This model was first proposed in McFadden (1973).

\textsuperscript{13}Individual i’s choice set is defined by his market m; however, no market subscript is needed on individual-specific variables since individual characteristics i include location. Subscript t is used to define years. Observed hospital characteristics were permitted to vary by year; the estimation was simplified by assuming that unobserved quality and the estimated coefficients were constant across the two years of available data. The time subscript is omitted for the remainder of the paper for ease of exposition.
provided by each plan’s hospital network. The exact methodology is described in Section 2.5.2. Finally the resulting variable, individual i’s expected utility from the set of hospitals offered by plan j in market m, which I denote $EU_{ijm}$, is used as an input to the health plan demand model. In this case I use a methodology similar to that first proposed by Berry, Levinsohn and Pakes (1995) (BLP), and later used by Nevo (2000, 2001(a)), Petrin (2001) and others. The utility of individual i from enrolling in plan j in market m is given by:

$$w_{ijm} = w(z_{jm}, \xi_{jm}, \text{prem}_{jm}, EU_{ijm}, y_i/\theta, \gamma)$$

(4)

where $z_{jm}, \xi_{jm}$ are vectors of observed and unobserved plan characteristics respectively, $\text{prem}_{jm}$ is the vector of plan premiums (prices charged to consumers), $y_i$ is the income of individual i, and $(\theta, \gamma)$ are the coefficients of this specification. The outside option is choosing to remain uninsured. Exactly the same reasoning applies as for the hospital demand model, so that if $B_{jm}$ is the set of consumer types that choose plan j in market m, then plan j’s share in market m is:

$$s_{jm}(z, \xi, \text{prem}/\theta, \gamma) = \Pr[(EU, y) \epsilon B_{jm}]$$

(5)

Details on the empirical approach to estimating this model are given in Section 2.5.3.

Preferences over characteristics in the hospital choice model are identified using variation in the individual’s hospital choice set across markets. The effect of each hospital’s characteristics is uncovered by considering the impact of changes in the choice set on hospital market shares: for example, how does the share of hospital A, which has a high number of nurses per bed, change when hospital B, with fewer nurses but more doctors per bed, is added to the choice set? (The choice set does not vary within markets since, as discussed in the next section, I consider indemnity and
PPO enrollees only.) Identification in the plan choice model comes from variation in the individual’s plan choice set across markets, and also from differences in plans’ hospital networks, both within and across markets. This allows the impact of the key variable, the individual’s expected utility from the plan’s hospital network, on his plan preferences to be estimated.

2.4 The Data

My analysis employs three sets of data. The first includes all the information needed to estimate the consumer utility equation for hospitals. The second identifies the hospitals in the network of every managed care plan in a significant sample of markets; this is used to predict the utility each individual can expect to gain from the hospital network offered by each plan in his or her market. The third dataset contains other plan characteristics and plan market shares; this is the final input needed to estimate the consumer utility equation for health plans.

2.4.1 Hospital Dataset

The core of this dataset is the MEDSTAT MarketScan Research Database for 1997-98. It is constructed from privately insured paid medical claims data provided by approximately 50 employer databases in over 200 markets across the US and gives encounter-level data on all hospital admissions of the relevant enrollees during this two-year period. For each admission, the data includes the patient diagnosis and characteristics, the identity of the hospital and the type of plan. Patient income is not included in the MEDSTAT data; I approximate this using the median income of families in the ZCTA, taken from Census 2000 data. This very detailed data enables me to pin down quite accurately the effect of individual patient characteristics on their choice of hospitals.
My final analysis will investigate consumers’ choice of managed care plans. Thus in an ideal world I would estimate consumers’ hospital choices using data for managed care enrollees and identifying preferences using each individual’s choice of hospital given the choice set specified by his or her plan. However, this is not feasible because the MEDSTAT data does not identify the hospital networks offered by each managed care plan: that is, the hospital choice set of managed care enrollees is unobserved. Instead I examine the choices made by indemnity and PPO enrollees, whose hospital choice sets are unrestricted. Two significant assumptions are required to apply the equation estimated here as an input to the plan choice equation estimated later. The first is that indemnity plan/PPO enrollees have the same preferences over hospitals as HMO/POS enrollees conditional on their diagnosis, income and location. This assumption has been made several times in the existing literature and may not be unreasonable\textsuperscript{14}. The average fee-for-service plan enrollee probably has different preferences over hospitals from the average managed care enrollee before he knows his diagnosis: for example, he may have a stronger desire for choice. However, when informed that he has a specific disease, he might well choose the same hospital as the average managed care enrollee of the same age and living in the same zip code\textsuperscript{15}.

The second assumption regards prices. PPO enrollees may be required to pay

\textsuperscript{14}For example, Town and Vistnes (2001) use data on the hospital selection decisions of Medicare enrollees, assuming that the Medicare population’s valuation of hospitals is a reasonable proxy for that of HMO enrollees. Capps, Dranove and Satterthwaite (2003) make a similar assumption to justify considering patients insured by Medicare, Medicaid, Blue Cross/Blue Shield and indemnity plans.

\textsuperscript{15}I test this assumption by estimating the hospital choice model separately using MEDSTAT data for HMO/POS enrollees in Boston MA, a market in which I observe that the vast majority of plans contract with all hospitals. (I observe this fact for 2003; I assume it also to have been true in 1997/8.) The estimated coefficients are not identical, but are broadly similar, to those estimated using PPO/indemnity enrollee data for Boston MA only. Where HMO/POS enrollees derive positive (negative) utility from a hospital or characteristic, PPO/indemnity enrollees do the same, often with similar coefficient magnitudes. Only 3 out of 36 hospital dummy coefficients and 2 out of 32 interaction coefficients are different in sign across the two models and both significant at \( p=0.1 \). While not overwhelming evidence, I take this to be sufficient to support the assumption, particularly given that no other approach is possible with the available data.
additional copays or deductibles if they choose to go out-of-network. These financial penalties, and the hospitals in the PPO network, are not identified in the dataset; that is, the “price” of the hospital at the point of service is unobserved. I therefore assume that out of pocket prices charged to patients on the margin are zero for both PPO and indemnity patients. This may be reasonable, particularly where prices take the form of increased deductibles, since many of these patients are likely to have spent past their deductible before making their decision\(^\text{16}\).

Further details on the hospital data, and the other datasets used for the demand estimation, are given in Appendix A. Tables 1 and 2 set out summary statistics for the hospital dataset.

**Table 1: Descriptive Statistics for Hospitals, MEDSTAT Dataset**

<table>
<thead>
<tr>
<th>Service</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of beds</td>
<td>286</td>
<td>193</td>
</tr>
<tr>
<td>Teaching status</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>For-profit</td>
<td>0.06</td>
<td>0.25</td>
</tr>
<tr>
<td>Registered nurses per bed</td>
<td>1.24</td>
<td>0.46</td>
</tr>
<tr>
<td>Cardiac services</td>
<td>0.72</td>
<td>0.37</td>
</tr>
<tr>
<td>Imaging services</td>
<td>0.42</td>
<td>0.26</td>
</tr>
<tr>
<td>Cancer services</td>
<td>0.60</td>
<td>0.41</td>
</tr>
<tr>
<td>Birth services</td>
<td>0.82</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Notes: N = 434 hospital-years. Cardiac, imaging, cancer and birth services refer to four summary variables defined in Appendix A. Each hospital is rated on a scale from 0 to 1, where 0 indicates that no procedures in this category are provided by the hospital, and a higher rating indicates that a less common service is offered.

\(^\text{16}\)I also have to assume that employees in the MEDSTAT dataset are representative of all privately-insured consumers in the market, conditional on age and sex.
### Table 2: Patient Descriptive Statistics, MEDSTAT Dataset

<table>
<thead>
<tr>
<th>Category</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis: Neurological</td>
<td>0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>Cardiac</td>
<td>0.11</td>
<td>0.32</td>
</tr>
<tr>
<td>Labor</td>
<td>0.17</td>
<td>0.38</td>
</tr>
<tr>
<td>Baby</td>
<td>0.07</td>
<td>0.26</td>
</tr>
<tr>
<td>Digestive</td>
<td>0.09</td>
<td>0.28</td>
</tr>
<tr>
<td>Cancer</td>
<td>0.08</td>
<td>0.27</td>
</tr>
<tr>
<td>Distance to chosen hospital (miles)</td>
<td>10.61</td>
<td>9.95</td>
</tr>
<tr>
<td>Distance to all hospitals (miles)</td>
<td>20.98</td>
<td>12.93</td>
</tr>
<tr>
<td>Female</td>
<td>0.63</td>
<td>0.48</td>
</tr>
<tr>
<td>Age: 0-17</td>
<td>0.13</td>
<td>0.34</td>
</tr>
<tr>
<td>18-34</td>
<td>0.22</td>
<td>0.42</td>
</tr>
<tr>
<td>35-44</td>
<td>0.13</td>
<td>0.33</td>
</tr>
<tr>
<td>45-54</td>
<td>0.19</td>
<td>0.39</td>
</tr>
<tr>
<td>55-64</td>
<td>0.29</td>
<td>0.45</td>
</tr>
<tr>
<td>Over 64</td>
<td>0.03</td>
<td>0.17</td>
</tr>
<tr>
<td>Industry: Manufacturing (durable)</td>
<td>0.31</td>
<td>0.46</td>
</tr>
<tr>
<td>Manufacturing (nondurable)</td>
<td>0.06</td>
<td>0.24</td>
</tr>
<tr>
<td>Transport, Communications, Utilities</td>
<td>0.02</td>
<td>0.14</td>
</tr>
<tr>
<td>Finance, Insurance, Real Estate</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>Services</td>
<td>0.46</td>
<td>0.50</td>
</tr>
<tr>
<td>State and Local Government</td>
<td>0.03</td>
<td>0.17</td>
</tr>
<tr>
<td>Working status: Full time</td>
<td>0.74</td>
<td>0.44</td>
</tr>
<tr>
<td>Part time</td>
<td>0.002</td>
<td>0.05</td>
</tr>
<tr>
<td>Early retiree</td>
<td>0.19</td>
<td>0.39</td>
</tr>
<tr>
<td>Retiree</td>
<td>0.04</td>
<td>0.19</td>
</tr>
<tr>
<td>PPO enrollee</td>
<td>0.51</td>
<td>0.50</td>
</tr>
<tr>
<td>Emergency admission</td>
<td>0.05</td>
<td>0.21</td>
</tr>
</tbody>
</table>

N = 28,666 encounters
I consider the 11 largest markets in the MEDSTAT data (those with over 1000 observations per market-year): the dataset therefore includes 217 hospitals, 434 hospital-years and 28,666 encounters in total. The hospitals have 286 beds and 1.24 registered nurses per bed on average; 20% are teaching hospitals. 51% of patients are enrolled in a PPO rather than an indemnity plan; 63% are women. The most common diagnoses are labor (17% of encounters) and cardiac admissions (11%). The employers providing data are not identified; however, the most common industry sectors are services (including hotels, personal services and health care; these comprise 46% of observations) and durable good manufacturing (31% of observations). Other industry sectors represented include non-durable good manufacturing, transport, communications and utilities, and state and local government.

2.4.2 Plan Networks

The link between the hospitals and the HMO/POS plans in each market is provided by a uniquely constructed dataset that defines, for every HMO/POS plan in the 43 markets covered by the plan characteristics data, the network of hospitals offered to enrollees in March/April 2003. The dataset includes 516 HMO/POS plans and 665 hospitals in total (on average 12 plans and 15 hospitals in each market). The data are summarized in Section 2.2.2 above.

2.4.3 Plan Characteristics

The final dataset contains aggregate data on health plan characteristics, including plan enrollment, for the 516 HMO/POS plans covered by the plan networks dataset. Each potential plan choice is defined as an insurer - product (HMO/POS) – market. Thus Aetna HMO is a separate choice from Aetna POS plan in Boston. Data are brought together from several sources. Atlantic Information Services data provide information on enrollment and characteristics (premiums earned, number of enrollees,
tax status, plan age, and other information) for the 3rd and 4th Quarters of 2002 respectively. The Weiss Ratings’ Guide to HMOs and Health Insurers includes additional plan characteristics for over 500 HMOs in Fall 2002. Finally, data on plan performance are taken from the Health Employer Data and Information Set (HEDIS) and the Consumer Assessment of Health Plans (CAHPS) 2000 data, both of which are published by the National Committee for Quality Assurance (NCQA). These data measure clinical performance and patient satisfaction in 1999, and are taken from the first year in which NCQA required plans to hire third-party firms to audit the data.17

Table 3 contains variable definitions and summary statistics for the plan dataset. The average market share of the HMO/POS plans in the dataset is 3%.18 Premiums average $141 per member per month. 35% of insurers are POS plans; 76% have been in existence for over 10 years. HEDIS scores vary widely, from an average rating of 0.15 (for the percent of children receiving all required doses of MMR, Hepatitis B and VZV vaccines before their 13th birthday) to an average of 0.73 (the proportion of women aged 52-69 who had received a mammogram within the previous two years). The two most frequently-occurring plans are Aetna and CIGNA, with 15% and 10% of observations respectively.

---

17 Individuals choosing health plans in 2002 would in fact have been informed by NCQA 2001 data, to which I do not have access. I therefore choose the HEDIS and CAHPS 2000 variables that are most highly correlated with their 1999 counterparts, assuming that their correlation with the 2001 data will also be high. All variables used have correlations with the 1999 data of over 0.65.

18 Shares are measured as percent of the nonelderly population in the market. See Appendix A for details of the methodology used to generate this variable and robustness tests of the resulting data.
Table 3: Descriptive Statistics for HMO/POS Plans

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>N</th>
<th>Mean</th>
<th>Std Devn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market Share</td>
<td>Plan share of non-elderly market</td>
<td>516</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Premium pmpm ($)</td>
<td>premiums earned per member per month</td>
<td>478</td>
<td>140.75</td>
<td>44.27</td>
</tr>
<tr>
<td>Physicians per thousand popln</td>
<td>number of physician contracts per 1000 popln in markets covered by plan</td>
<td>418</td>
<td>1.56</td>
<td>1.51</td>
</tr>
<tr>
<td>Breast cancer screening</td>
<td>% of women aged 52-69 who received a mammogram within last 2 years</td>
<td>352</td>
<td>0.73</td>
<td>0.05</td>
</tr>
<tr>
<td>Cervical cancer screening</td>
<td>% of adult women who received a pap smear within last 3 years</td>
<td>352</td>
<td>0.72</td>
<td>0.07</td>
</tr>
<tr>
<td>Check-ups after delivery</td>
<td>% of new mothers receiving a check-up within 8 weeks of delivery</td>
<td>351</td>
<td>0.72</td>
<td>0.11</td>
</tr>
<tr>
<td>Diabetic eye exam</td>
<td>% of adult diabetics receiving eye exam within last year</td>
<td>350</td>
<td>0.45</td>
<td>0.11</td>
</tr>
<tr>
<td>Adolescent immunization 1</td>
<td>% of children receiving all required doses of MMR and Hep B vaccines before 13th birthday</td>
<td>346</td>
<td>0.31</td>
<td>0.16</td>
</tr>
<tr>
<td>Adolescent immunization 2</td>
<td>% of children receiving all required doses of MMR, Hep B and VZV vaccines before 13th birthday</td>
<td>313</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>Advice on smoking</td>
<td>% of adult smokers advised by physician to quit</td>
<td>213</td>
<td>0.63</td>
<td>0.07</td>
</tr>
<tr>
<td>Mental illness checkup</td>
<td>% of members seen as outpatient within 30 days of discharge after hospitalizn for mental illness</td>
<td>307</td>
<td>0.68</td>
<td>0.15</td>
</tr>
<tr>
<td>Care quickly</td>
<td>Composite measure of member satisfaction re: getting care as soon as wanted</td>
<td>304</td>
<td>0.75</td>
<td>0.05</td>
</tr>
<tr>
<td>Care needed</td>
<td>Composite measure of member satisfaction re: getting authorizations for needed/desired care</td>
<td>304</td>
<td>0.72</td>
<td>0.06</td>
</tr>
<tr>
<td>Age 0-2</td>
<td>Dummy for plans aged 0 - 2 years</td>
<td>516</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>Age 3-5</td>
<td>Dummy for plans aged 3 - 5 years</td>
<td>516</td>
<td>0.06</td>
<td>0.23</td>
</tr>
<tr>
<td>Age 6-9</td>
<td>Dummy for plans aged 6 - 9 years</td>
<td>516</td>
<td>0.17</td>
<td>0.37</td>
</tr>
<tr>
<td>Aetna</td>
<td>Plan fixed effect</td>
<td>516</td>
<td>0.15</td>
<td>0.36</td>
</tr>
<tr>
<td>CIGNA</td>
<td>Plan fixed effect</td>
<td>516</td>
<td>0.10</td>
<td>0.31</td>
</tr>
<tr>
<td>Kaiser</td>
<td>Plan fixed effect</td>
<td>516</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>BCBS</td>
<td>Dummy for ownership by BCBS</td>
<td>516</td>
<td>0.16</td>
<td>0.36</td>
</tr>
<tr>
<td>POS plan</td>
<td>Dummy for POS plan</td>
<td>516</td>
<td>0.35</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Descriptive statistics for the 665 hospitals with which these plans can choose to contract are given in Table 4\textsuperscript{19}. The hospitals have 339 beds and 1.26 registered nurses per bed on average; 20% are teaching hospitals.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of beds (set up and staffed)</td>
<td>338.66</td>
<td>217.19</td>
</tr>
<tr>
<td>Teaching status</td>
<td>0.195</td>
<td>0.397</td>
</tr>
<tr>
<td>For-profit</td>
<td>0.202</td>
<td>0.401</td>
</tr>
<tr>
<td>Registered nurses per bed</td>
<td>1.263</td>
<td>0.498</td>
</tr>
<tr>
<td>Cardiac services</td>
<td>0.812</td>
<td>0.310</td>
</tr>
<tr>
<td>Imaging services</td>
<td>0.539</td>
<td>0.287</td>
</tr>
<tr>
<td>Cancer services</td>
<td>0.647</td>
<td>0.402</td>
</tr>
<tr>
<td>Birth services</td>
<td>0.857</td>
<td>0.348</td>
</tr>
</tbody>
</table>

Notes: N = 665 hospitals. Cardiac, imaging, cancer and birth services refer to four summary variables defined in Appendix A. Each hospital is rated on a scale from 0 to 1, where 0 indicates that no procedures in this category are provided by the hospital, and a higher rating indicates that a less common service is offered.

Several assumptions are needed to link the datasets. In order to apply the estimated parameters from the 11 markets in the hospital dataset to the 43 markets used for the plan choice model I assume that hospital preferences given demographics and diagnosis are fixed across markets and over time. In addition, because patient diagnosis and severity of illness influence hospital choice but are not observed in the

\textsuperscript{19}These are the 665 hospitals located in the 43 markets covered by the plan dataset. They are identified in the plan network dataset described in Sections 2.2.2 and 2.4.2. I use their characteristics in the calculation of the expected utility variable. They form a separate sample from the providers in the MEDSTAT dataset discussed in Section 2.4.1 and summarized in Table 1.
health plan data, I assume invariance of both factors across markets and over time. I then predict diagnosis probabilities in different markets based on age and sex and assume that the predicted choices from an equation that excludes severity can be interpreted as average choices across the severity distribution\textsuperscript{20}.

### 2.5 Details on the Empirical Approach

#### 2.5.1 Demand for Hospitals

The specification of the utility function giving rise to PPO/indemnity enrollees’ demand for hospitals is given by:

\[ u_{ihl} = \delta_h + x_h\nu_{il}\beta + \varepsilon_{ihl} \]

where \( \varepsilon_{ihl} \) captures unobserved idiosyncratic tastes which are assumed to be iid distributed according to a Type 1 extreme value distribution. The hospital-specific variable \( \delta_h \) is given by:

\[ \delta_h = \eta_h + x_h\alpha \] (6)

where \( \eta, x \) are as defined in Section 2.2.3. This formulation implies that the share equation can be written as:

\[ s_h = \sum_{i,l} N_{il} \frac{\exp(\delta_h + x_h\nu_{il}\beta)}{\sum_{p\in H} \exp(\delta_p + x_p\nu_{il}\beta)} \] (7)

\textsuperscript{20}Linking the various sources of plan data also requires an assumption: since the network dataset was compiled after the AIS enrollment data, I assume that plans’ hospital networks were stable between 2001, when individuals choosing plans for Q4 2002 made their decisions, and Q1-2 2003 when the network data were observed. It is reassuring to note that I gathered 25% of the plan network data twice: first in Q4 2002 and second in Q2 2003. There was very little change in networks between these two dates.
where $N_{il}$ is the number of individuals in consumer-type $i$ who are hospitalized with diagnosis $l$, $N$ is the number of individuals admitted to hospital in the market, and $H$ is the set of hospitals in the market.

To ensure consistency of the estimate of $\beta$, I include hospital fixed effects in my estimation of equation (7): these give me the predicted $\hat{\delta}'$s\textsuperscript{21}. Since all the variables in equation (7) are observed, the estimation can be performed using MLE. A number of interaction terms are also included. Previous studies have shown that distance travelled to hospital has a significant negative effect on utility: the distance between hospital and patient residence zipcodes, distance squared, and distance interacted with market dummies and a dummy indicating an emergency admission are included to account for this. The other interactions are between patient characteristics (the seven diagnosis categories listed in Appendix A, income, a dummy for emergency admissions and a PPO dummy) and hospital characteristics (teaching status; tax status; the number of nurses per bed and four variables that summarize the services offered by each hospital)\textsuperscript{22}. Interactions that should have no effect (for example, a cancer diagnosis interacted with provision of birth services) are restricted to be zero.

Since the next step is to use the estimated coefficients in markets and years outside my sample, I need to predict the value of $\hat{\delta}$ using variables included in the plan dataset. Assuming that $\mu_h$ is the sampling error in $\delta_h$, $\hat{\delta}_h = \delta_h + \mu_h$, I estimate the following equation:

$$\hat{\delta}_h = x_h \alpha + \eta_h + \mu_h$$

(8)

where $\eta_h$ is unobserved hospital quality as defined above. I estimate equation (8)

\textsuperscript{21}The alternative would be to include uninteracted hospital characteristics (the $x$’s) in the specification: that is, to estimate $\alpha$ directly. However, this would imply ignoring unobserved quality $\eta$; the estimates of $\alpha$ and $\beta$ would therefore be inconsistent.

\textsuperscript{22}The hospital service variables are described in detail in Appendix A. Market and time fixed effects cannot be included since there is no outside option in the hospital choice equation.
by regressing the estimated hospital dummy coefficients on 31 hospital characteristic variables taken from the AHA 1997 and 1998 data. Market fixed effects are also included\textsuperscript{23}. To account for the heteroscedasticity introduced by $\mu$, I adjust the standard errors of the OLS regression using White’s (1980) heteroscedastic consistent standard error estimator\textsuperscript{24}.

The estimated coefficients (the $\alpha$’s in equation (8)) are used to predict the $\delta$’s in equation (7) when the analysis is extended out-of-sample. The specification for predicted utility used in the subsequent analysis is therefore given by the following:

$$\hat{u}_{ihl} = \eta_h + x_h\hat{\alpha} + x_h\nu_{il}\hat{\beta} + \varepsilon_{ihl}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimated in the two-stage process described above, $\varepsilon$ is assumed to be distributed iid according to the Type 1 extreme value distribution, and $\eta$ is unobserved hospital quality.

\textsuperscript{23}The variables included are: number of beds; distance from City Hall; distance from City Hall squared; registered nurses per bed and nurses per bed squared; doctors per bed; dummy variables for JCAHO accreditation, cancer program approved by ACS, residency training program, medical school, member of Council of Teaching Hospitals of the Association of American Medical Colleges, Independent Practice Association, Foundation, Indemnity Fee for Service Plan, osteopathic hospital, and operating subsidiary corporations; Control/ownership dummies (control by county, Church and For-profit partnership); and a list of service dummies (for neonatal intensive care, angioplasty, cardiac catheterization laboratory, computed-tomography scanner, positron emission tomography, single photon emission computerized tomography, oncology services, obstetric services, emergency department, breast cancer screening, burn care, and alcohol/drug abuse inpatient care). Market fixed effects are needed to normalize the baseline hospitals - that is, the randomly-chosen hospitals, one in each market, whose dummies are excluded from the choice model - across markets. The fixed effects are set to zero when the analysis is extended out-of-sample.

\textsuperscript{24}White’s estimator does not adjust for serial correlation. However, the off-diagonal elements of $E(\mu\mu')$, the variance-covariance matrix for $\hat{\delta}$, are of order one hundredth the magnitude of the diagonal elements. I assume that $E(\eta\eta') = \sigma^2 I$, so by assumption there is no serial correlation in $\eta$. It therefore seems reasonable to assume no serial correlation in $\eta + \mu$. 

28
2.5.2 Expected Utility from each Plan’s Hospital Network

2.5.2.1 Computing the Expected Utility Variable

The analysis now moves from the hospital dataset to the plan data, taking the estimated coefficients from equations (7) and (8) and using them to predict the utility that each type of individual expects to receive from the hospital network of each HMO/POS plan in his market\(^{25}\). Individual i’s expected utility from the hospitals offered by plan j in market m is given by:

\[
EU_{ijm} = \sum_{l} p_{il} \ E_{\hat{u}_{ihl}}(\max_{h \in H_{jm}} (\hat{u}_{ihl}))
\]

where \(p_{il}\) is the probability that individual i will be hospitalized with diagnosis l and \(H_{jm}\) is the set of hospitals offered by HMO/POS plan j in market m\(^{26}\). Ben-Akiva (1973) shows that, under the assumption of Type 1 extreme value errors, this formula reduces to:

\[
EU_{ijm} = \sum_{l} p_{il} \log \left( \sum_{h \in H_{jm}} \exp(\eta_h + x_h \hat{\alpha} + \nu_{il} \hat{\beta}) \right)
\]  

Six diagnoses (l) are included here: all the categories described in Appendix B except newborn babies\(^{27}\).

\(^{25}\)This is the point in the analysis when I begin to use the dataset listing the hospital network of each HMO/POS plan. Up until now I have considered only indemnity/PPO enrollees, who I assume have access to all hospitals in the market; I now switch to HMO/POS enrollees whose choice is limited.

\(^{26}\)The expectation over values of \(\varepsilon\) implies an assumption that each individual’s \(\varepsilon\) is unknown when he chooses his plan. The alternative assumption, that he knows his \(\varepsilon\) when making the choice, leads to a different expression for expected utility: \(EU_{ijm} = \sum_{l} p_{il} \max(\hat{u}_{ihl})\). I estimate this quantity as a robustness check by taking draws of \(\varepsilon\) from the Type 1 extreme value distribution, and assigning a draw to each simulated individual in each ZCTA-age-sex cell, for every possible diagnosis. Using this new expected utility variable in the health plan choice model had little effect on the final results.

\(^{27}\)Newborn infants are included in the hospital choice model but not the expected utility calculation. Inclusion in the hospital model implies that the baby is considered by the hospital to be a separate unit from its mother. In contrast, I assume that the health plan does not consider the baby as a separate entity, whose preferences should be considered separately from those of its mother, until after the birth episode. (The mother considers her preferences regarding the hospital in which to give birth when she chooses her health plan; she does not choose a plan for her baby until after
2.5.2.2 Methodologies to Account for Unobserved Hospital Quality

The rest of the methodology would be straightforward if \( \eta \) was known; unfortunately I do not observe this variable directly. I do, however, have information on the distribution of \( \eta + \mu \) (the residuals from estimation of equation (8)). This enables me to use a methodology first introduced by Hanushek (1974) and developed by Lewis (2000) to estimate the variance of \( \eta \). The idea is to use OLS regression to estimate the variance of \( \eta + \mu \), and then use the estimate of \( E(\mu \mu') \) obtained in the first stage choice model to back out a variance estimate for \( \eta \). Details are given in Appendix B.

I find that the variance of \( \eta \) is negligible compared to that of \( \mu \): only a very small proportion of the "lack of fit" in estimating equation (8) is due to unobserved hospital quality. For this reason I assume \( \eta = 0 \) in the main analysis.

I conduct one robustness test for this assumption. It turns out that, if we take draws \( \eta^* \) from the distribution of \( \eta \), and assign one draw as each hospital’s unobserved quality (measured with error), then the error in the resulting expected utility variable \( EU_{ijm}^{\eta^*} \) (the input to the plan demand equations) is:

\[
\text{Error} = EU_{ijm}^{\eta} - EU_{ijm}^{\eta^*} = \sum_l p_l \log \left( \sum_{h \in H_{jm}} C_{hl} \exp(\eta_h) \right) - \sum_l p_l \log \left( \sum_{h \in H_{jm}} C_{hl} \exp(\eta^*_h) \right)
\]

where \( C_{hl} = \exp(x_h \hat{\alpha} + x_h \nu_{il} \hat{\beta}) \). This implies that the expected value of the error is zero, since \( \eta \) and \( \eta^* \) are iid draws from the same distribution. In a linear context, such as the logit demand framework described below, we can therefore adjust for the measurement error using an instrument correlated with \( EU_{ijm}^{\eta^*} \) but not with the error in measuring this variable. I use \( EU_{ijm}^{\eta^*} \), the predicted expected utility from the plan’s hospital network found using the methodology above but ignoring \( \eta \), as an

its birth.) (Less than 0.1% of children aged 0-18 are under than one week old, assuming a constant birth rate, so the assumption that newborn infants do not count towards market shares is reasonable even though they are not explicitly excluded from the share denominator.)
instrument for $EU_{ijm}^\eta$. I assume that $\mu = 0$ when creating $EU_{ijm}^\eta$; the test therefore provides an upper bound for the impact of $\eta$ on the model's predictions. The results are encouraging: the test's logit model predictions are very similar to those for the main model, implying that the assumption that $\eta = 0$ does not lead to significant bias. While the robustness test strictly applies only to the logit methodology (the nonlinearity in the full plan choice model prevents me from conducting the same test there), the combined results of this section are sufficient to justify setting $\eta = 0$ for the main analysis.

2.5.2.3 Creating the Expected Utility Variable: Other Issues

The coefficients on market-distance interactions, emergency-distance interactions, and interactions with PPO dummies and newborn infant dummies in the hospital utility equation are assumed to be zero in the expected utility calculation. I assume that patients in POS plans, like those in HMOs, are restricted to hospitals within their insurer's network (that is, I ignore the possibility that their PCP might refer them to an out-of-network hospital). I have no choice about making this assumption: my plan data is at the aggregate level so I have no way to observe which hospitals enrollees in particular POS plans actually visit. The expected utility of consumers choosing indemnity plans or no insurance are also needed. Consumers with indemnity plans can go to any hospital in the market and their expected utility is predicted accordingly. I assume that uninsured consumers expect to gain zero utility from hospitals.

---

28 The expected utility variable used in the logit framework, which as describe below is effectively an average over individuals within the most populated ZCTA in the market, also has the property that $E(\text{error}/x, v) = 0$, implying that the instrumental variables methodology is valid.

29 I assume that consumers do not consider the hospital they would choose in an emergency when choosing a health plan. My assumption that preferences over hospital characteristics, given diagnosis and demographics, are fixed across markets implies no need to include market-distance interactions: I assume the coefficient on distance is that estimated for Boston, my "base case" market.

30 This is not quite right: uninsured consumers can be admitted to public hospitals as emergency cases. The implications of this are discussed later in the paper.
2.5.3 Demand for Health Plans Given the Hospital Network Offered

The final step is to take the predicted expected utility from the hospital network of each plan, for each type of individual, and include it as an input to the plan demand equation. I begin with a simple logit model, ignoring the distribution of individual characteristics within markets, before moving on to a fuller model where variation in individual attributes is taken into account.

2.5.3.1 The Benchmark Model: Logit Formulation

The logit framework assumes common coefficients for all individuals. It has the advantage of generating an equation that is easy to estimate, making it a useful benchmark model. However, estimating such a model requires a plan-level, rather than plan-individual-level variable representing expected utility from the hospital network. I therefore define a representative agent in each market as an individual living in the most populated ZCTA and having the weighted average probability of diagnoses and of hospital admission of people resident in that ZCTA. I define the plan-level expected utility as that of the representative agent in the relevant market, that is:

\[
EU_{rep}^{jm} = \sum_{l} p_{l}^{rep} \log \left( \sum_{h \in H_{jm}} \exp(\eta_{h} + x_{h} \hat{\alpha} + x_{h} \nu_{l}^{rep} \hat{\beta}) \right)
\]

where \( p_{l}^{rep} \) is the weighted average probability of diagnoses of individuals in the most populated ZCTA, and \( \nu_{l}^{rep} \) is the vector of other characteristics of an individual in that ZCTA (income and location, the only individual characteristics other than diagnosis that are not set to zero, are both defined at the ZCTA level). I control for the unobserved \( \eta \) first by assuming \( \eta = 0 \) and then by using the instrumental variables methodology already described. The utility of consumer \( i \) from choosing plan \( j \) in market \( m \) is therefore given by:
\[ w_{ijm} = \xi_{jm} + z_{jm}\theta + \gamma_1 EUrep_{jm} + \omega_{ijm} \]

where premium is included in the observed plan characteristics \( z \), and I assume that the \( \omega_{ijm} \) is distributed iid according to a Type 1 extreme value distribution\(^{31}\). Normalizing the utility of all consumers from the outside good (good 0) to be zero, I obtain the standard equation for estimation:

\[
\log(s_{jm}) - \log(s_{0m}) = \xi_{jm} + z_{jm}\theta + \gamma_1 EUrep_{jm}
\tag{11}
\]

where \( s_{0m} \) is the share of the outside good in market \( m \).

The logit model can be estimated using a simple two stage least squares methodology (instrumenting for premiums since these are likely to be correlated with the unobserved quality variable \( \xi \)). The basic plan characteristics included in \( z \) are: premium per member per month, number of physicians per 1000 members of the population and age of the plan (coded into four dummies: less than 3 years, 3-5 years, 6-9 years and over 10 years of age). The HEDIS measures used are the breast cancer and cervical cancer screening rates; the rate of check-ups after live deliveries; the proportion of diabetic patients with annual eye exams; the proportion of adolescents receiving final immunizations before their 13th birthday; the proportion of smokers advised by their physician to quit; and the proportion of patients seen on an outpatient basis within 30 days of discharge from a mental illness admission. The CAHPS measures are "getting needed care" and "getting care quickly". Each is an aggregation of responses to several CAHPS questions. The CAHPS measures of plan performance are highly

\(^{31}\)Both the logit and the full demand specifications assume that consumer utility from the hospital network (which is a function of the probabilities of different possible diagnoses) is additively separable from other aspects of plan quality in the plan demand equation. In terms of a utility equation where expected health is the key input, this implies that the individual’s expectation of the effect of a hospital on his health depends on his diagnosis, but his expectation of the effect of the plan is independent of diagnosis. This makes sense given that many plan functions are preventive whereas most if not all hospital interventions are specific to a diagnosis.
correlated with one another; so are the HEDIS measures. Missing variable dummies are included in all specifications. Insurer fixed effects are included for insurers that are active in at least 10 of the major markets defined by AIS\textsuperscript{32}, and market dummies are also included in some specifications.

The model is completed by defining the outside good. The simplest definition to implement would be a composite of non-managed care private coverage and uninsured (I exclude Medicare by considering only the non-elderly population, and exclude Medicaid by assumption; see Appendix A for details). However, indemnity coverage and no coverage are at opposite ends of the spectrum in terms of price and many aspects of quality (especially hospital access) so this outside good would be non-homogeneous. Instead I define the outside good as "choosing to be uninsured" and create a separate choice in each market, defined as "choosing indemnity or PPO insurance", and assumed to be homogeneous within each market\textsuperscript{33}. None of the data sources provides information on non-managed care coverage, so assumptions must be made to complete the dataset. Indemnity plans are assumed to be over ten years old; to have premiums equal to the highest managed care premium in the relevant market; and to offer a physician network size equal to the largest offered by a managed care plan in the market. Indemnity plan performance ratings (both HEDIS and CAHPS) are assumed to equal the average of managed care plans in the market\textsuperscript{34}. Average quality for uninsured consumers is not identified in the plan choice model (unless I make more assumptions or normalize one of the "inside" goods) so I normalize it to

\textsuperscript{32} These are Aetna, CIGNA, Coventry, Health Net, Kaiser Permanente, One Health, PacifiCare, United, UNICARE and Blue Cross Blue Shield plans.

\textsuperscript{33} Assuming homogeneity of this option, and making the assumptions needed to define its characteristics, are problematic. The quality of this option will affect consumer choices of HMO and POS plans. Data limitations prevent me from modelling this option more completely; instead I include market fixed effects in the demand specification to account for cross-market differences in both this and the outside option.

\textsuperscript{34} An alternative would be to assume that indemnity plan ratings = 0 for all HEDIS and CAHPS measures. The analyses were repeated under this assumption with little change in results.
zero.

The instruments used for the premium variable, in addition to the usual set of plan characteristics (the $z$'s), are the average hourly hospital wage and the average weekly nurse wage across the markets in which each health plan is observed to be active. The assumption required for these to be valid instruments is that health plan costs are correlated with premiums but not with unobserved health plan quality. Hospital wage data comes from the Centers for Medicaid and Medicare Services (CMS) 1999; nurse wage data is from the Bureau of Labor Statistics 1999.

The last issue is the need to adjust the estimated standard errors for the three-step estimation process being used here\footnote{The first step is the estimation of the hospital choice model including hospital fixed effects; the second is the regression of the fixed effect coefficients on hospital characteristics; the third is estimation of the health plan choice model including the expected utility variable.}. I need to take into account the fact that the expected utility variable is constructed from estimated parameters. This has no impact on the consistency of the results but, since the estimator is not adaptive, will affect the standard error estimates.

To incorporate the effect of these estimated parameters into the estimates of the variance-covariance matrix, I take advantage of the GMM structure of the estimation procedure. Incorporating estimated parameters into a GMM estimator is fairly straightforward. An extension of Pakes (1997) shows that if there exist first stage parameter estimates $\beta$ such that

$$\sqrt{N_1}(\beta_n - \beta_0) = L_1 \sqrt{N_1}^{-1} \sum_{h,i} f(x_h, \nu_i) + o_p(1)$$

and second stage parameter estimates $\alpha$ such that

$$\sqrt{N_2}(\alpha_n - \alpha_0) = L_2 \sqrt{N_2}^{-1} \sum_h m(x_h) + o_p(1)$$

then, under standard regularity conditions, the third stage parameter estimates...
\( \theta_n \) are distributed asymptotically normal as

\[
\sqrt{n}(\theta_n - \theta_0) \rightarrow_d N[0, (\Gamma'^{-1} \Gamma)^{-1}\Gamma' A^{-1} V A^{-1} \Gamma (\Gamma'^{-1} \Gamma)^{-1}]
\]

where \( \Gamma \) is the derivative of the moment condition with respect to the parameters, 
\( A \) is the weight matrix and \( V \) is given by

\[
V = a_{var} \left( \frac{1}{\sqrt{n}} \sum_{j,m} \{ g_{jm}(z, x, \nu, \alpha_0, \beta_0, \theta_0) + \sqrt{n} M_1 (\beta_n - \beta_0) + \sqrt{n} M_2 (\alpha_n - \alpha_0) \} \right) \\
= E(gg' + \frac{n}{N_1} M_1 v_{C1} M_1' + \frac{n}{N_2} M_2 v_{C2} M_2') + 6 \text{ covariance terms}
\]

where \( g_{jm} \) are the third-stage moments, \( M_1 \) is a matrix of derivatives of \( g(.) \) with respect to the elements of \( \beta_0 \), \( M_2 \) is a matrix of derivatives of \( g(.) \) with respect to the elements of \( \alpha_0 \), and \( v_{C1} \) and \( v_{C2} \) are the variance-covariance matrices from the first two stages. The covariance terms can be reduced to an expression depending solely on the expectations of products of \( M_1, M_2, L_1, L_2 \), and the individual moment conditions. Further details of the methodology are given in Appendix C.

2.5.3.2 The Full Demand Specification

Since the focus of the analysis is on the weight consumers attach to the expected utility variable in the plan demand equation, the obvious extension to the logit model is to allow a richer specification of this expected utility term. The full demand specification includes the full variable \( EU_{ijm} \) rather than the summary measure used in the logit formulation, and therefore accounts for the impact of the distribution of individual locations, income and demographics within each market on plan market shares. This has the additional advantage of avoiding the logit model’s well-known unattractive implication: the imposition of the independence of irrelevant alternatives (IIA) substitution pattern for an individual patient’s choice of plan. The IIA assumption implies that cross-price effects are a function solely of plan shares and are indepen-
dent of plans’ relative positions in the characteristic space. This is clearly inaccurate
to the extent that consumers substitute more readily between plans that are ‘closer’
in terms of characteristics (for example a consumer switching from a high-premium,
high-choice plan is more likely to choose another high-premium plan with similar
qualities than a low-premium competitor that offers restricted choice of providers).

The utility of consumer i choosing plan j in market m in the full demand speciﬁcation is given by:

\[ w_{ijm} = \xi_{jm} + z_{jm} \theta + \gamma_1 EU_{ijm} + \gamma_2 \frac{\text{prem}_{jm}}{y_i} + \omega_{ijm} \]  

(12)

where \( \text{prem}_{jm} \) is plan j’s premium in market m, \( y_i \) is the median family income of
individual-type i (deﬁned by ZCTA), and the other variables are as speciﬁed above.
This is similar to the model introduced in BLP (1995) in which random coefﬁcients,
which are functions of demographic variables taken from market-level census data,
are interacted with product characteristics. The difference is that, in this equation,
both consumer-speciﬁc terms (\( EU_{ijm} \) and \( y_i \)) are observed, either at the ZCTA level
(in the case of income) or at the ZCTA-agegroup-sex level (for the expected utility
variable). Therefore simulation methods are not needed to evaluate the estimation
algorithm\(^{36}\). The share equation reduces to:

\[ s_{jm} = \sum_{i} \left( \frac{n_i}{n_m} \right) s_{ijm}(\theta, \gamma) \]  

(13)

where \( n_i \) is the number of individuals in consumer-type i, \( n_m \) is the number in the
market, and \( s_{ijm}(\theta, \gamma) \), the share of type - i individuals choosing plan j in market m,
is deﬁned by:

\(^{36}\)The ZCTA-level income measure, and consumer cell-level expected utility measure, are them-
selves assumed to be integrals over the types of people within the cell. The methodology used here
therefore effectively integrates out over individuals, as in BLP.
s_{ijm}(\theta, \gamma) = \frac{\exp(\xi_{jm} + z_{jm} \theta + \gamma_1 E_{ijm} + \gamma_2 \frac{\text{prem}_{jm}}{y_i})}{1 + \sum_{k \in P_m} \exp(\xi_{km} + z_{km} \theta + \gamma_1 E_{ikm} + \gamma_2 \frac{\text{prem}_{km}}{y_i})} \quad (14)

The presence of the unobserved quality measure \(\xi\) implies that MLE cannot be used to estimate the model (as it was for the plan demand equation)\(^{37}\). Instead, the contraction mapping introduced in BLP (1995) is used to transform equation (13) into a linear equation for \(\xi\) and the coefficients estimated using a GMM methodology, again as in BLP (1995)\(^{38}\). The variables included in \(z\) are the same as those used in the logit specification. I instrument for premium for the same reasons as in the logit framework; I add two instruments to the hospital and nurse wage variables used before. These are the average expected utility in the most populated ZCTA in the market (the \(E_{U_{rep}}_{jm}\) variable defined for the logit model) and the average income across the other markets in which the plan is observed to be active. It is clear that \(E_{U_{rep}}_{jm}\) should be correlated with premiums and the setup of the model already implies that it is uncorrelated with unobserved quality \(\xi\) as required for a valid instrument. Some additional assumptions are needed to include the average income across the rest of the plan’s markets as an instrument. First, the average income in each market must affect aspects of the plan’s unobserved quality, such as promotional activity, in that market but not in others, and this activity must affect plan costs. Second, plans must set premiums in market \(j\) taking account of costs in other markets. Put simply, the assumption is that plans determine variables such as promotional activity locally but set premiums (at least partly) centrally\(^{39}\).

\(^{37}\)If plan fixed effects were included they would subsume the unobservable \(\xi\) and MLE could be used. However, the dataset is too small to successfully estimate all 516 plan dummy coefficients in addition to the other parameters.

\(^{38}\)I use heteroscedasticity-robust standard errors which also allow for correlated errors across the three stages of estimation. The weight matrix used in this analysis is the two stage least squares weight.

\(^{39}\)The assumptions needed are somewhat troubling. However, the instrument does seem to be valid: the estimated coefficients change very little, and the standard errors go down, when the instrument is added. The Difference-in-Sargan or C-statistic for this instrument (Hayashi p218) was
Standard error adjustments are again needed to account for the variance introduced in the first two steps of estimation; the methodology used is exactly analogous to that described for the logit model.

2.6 Results

2.6.1 Hospital Demand: Results

Tables 5 and 6 set out the results of the hospital demand specification. Table 5 shows the results of the first stage: the estimation of the hospital choice model using MLE and including hospital fixed effects. Only the interaction term coefficients are displayed in the table.

The results are very much in line with the previous hospital choice literature, and are intuitive. If a hospital moves an additional mile away from a patient’s home, this reduces the probability that the patient will choose it by 21%. Emergency admissions are even more sensitive to distance than non-emergencies. The non-interacted effect of teaching hospitals on patient utility is subsumed in the fixed effects; however, the interaction coefficients show that patients with the most complex conditions (neurological diagnoses) attach the highest positive weight to these types of hospital. Cancer patients have very strong preferences for hospitals with a large number of nurses per bed. Women in labor are the only types of patient attaching a positive weight to for-profit hospitals. Finally, and reassuringly, cardiac diagnoses place a strong positive weight on hospitals with good cardiac services; cancer patients on hospitals with good cancer services (although this coefficient is not significant at p =0.1), and women in labor and newborn babies have a strong preference for hospitals with good labor services.

1.37, implying that the null that income is a valid orthogonal instrument cannot be rejected even in a test of size 0.24.
Table 5: Hospital Demand Results, MLE Estimation

<table>
<thead>
<tr>
<th>Interaction Terms</th>
<th>Variable</th>
<th>Estimated coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distance (miles)</td>
<td>-0.215** (0.004)</td>
</tr>
<tr>
<td></td>
<td>Distance squared</td>
<td>0.001** (0.000)</td>
</tr>
<tr>
<td></td>
<td>Emergency * distance</td>
<td>-0.008** (0.004)</td>
</tr>
<tr>
<td>Interactions: Teaching</td>
<td>Cardiac</td>
<td>0.090 (0.060)</td>
</tr>
<tr>
<td></td>
<td>Cancer</td>
<td>0.192** (0.069)</td>
</tr>
<tr>
<td></td>
<td>Neurological</td>
<td>0.546** (0.175)</td>
</tr>
<tr>
<td></td>
<td>Digestive</td>
<td>-0.145** (0.062)</td>
</tr>
<tr>
<td></td>
<td>Labor</td>
<td>0.157** (0.048)</td>
</tr>
<tr>
<td></td>
<td>Newborn baby</td>
<td>0.038 (0.075)</td>
</tr>
<tr>
<td></td>
<td>Income ($000)</td>
<td>0.007** (0.001)</td>
</tr>
<tr>
<td></td>
<td>PPO enrollee</td>
<td>-0.067 (0.050)</td>
</tr>
<tr>
<td>Interactions: Nurses per bed</td>
<td>Cardiac</td>
<td>-0.096 (0.070)</td>
</tr>
<tr>
<td></td>
<td>Cancer</td>
<td>0.445** (0.079)</td>
</tr>
<tr>
<td></td>
<td>Neurological</td>
<td>0.130 (0.200)</td>
</tr>
<tr>
<td></td>
<td>Digestive</td>
<td>-0.028 (0.076)</td>
</tr>
<tr>
<td></td>
<td>Labor</td>
<td>-0.002 (0.063)</td>
</tr>
<tr>
<td></td>
<td>Newborn baby</td>
<td>0.071 (0.087)</td>
</tr>
<tr>
<td></td>
<td>Income ($000)</td>
<td>0.005** (0.001)</td>
</tr>
<tr>
<td></td>
<td>PPO enrollee</td>
<td>-0.099* (0.056)</td>
</tr>
<tr>
<td>Interactions: For-Profit</td>
<td>Cardiac</td>
<td>-0.164 (0.181)</td>
</tr>
<tr>
<td></td>
<td>Cancer</td>
<td>-0.197 (0.202)</td>
</tr>
<tr>
<td></td>
<td>Neurological</td>
<td>0.229 (0.379)</td>
</tr>
<tr>
<td></td>
<td>Digestive</td>
<td>0.195 (0.150)</td>
</tr>
<tr>
<td></td>
<td>Labor</td>
<td>0.300** (0.107)</td>
</tr>
<tr>
<td></td>
<td>Newborn baby</td>
<td>0.194* (0.122)</td>
</tr>
<tr>
<td></td>
<td>Income ($000)</td>
<td>-0.001 (0.003)</td>
</tr>
<tr>
<td></td>
<td>PPO enrollee</td>
<td>-0.036 (0.090)</td>
</tr>
<tr>
<td>Interactions: Cardiac Services</td>
<td>Cardiac</td>
<td>1.222** (0.134)</td>
</tr>
<tr>
<td></td>
<td>Income ($000)</td>
<td>0.001 (0.001)</td>
</tr>
<tr>
<td></td>
<td>PPO enrollee</td>
<td>0.080 (0.088)</td>
</tr>
<tr>
<td>Interactions: Imaging Services</td>
<td>Cardiac</td>
<td>-0.188** (0.094)</td>
</tr>
<tr>
<td></td>
<td>Cancer</td>
<td>-0.052 (0.107)</td>
</tr>
<tr>
<td></td>
<td>Neurological</td>
<td>-0.084 (0.287)</td>
</tr>
<tr>
<td></td>
<td>Digestive</td>
<td>-0.182* (0.104)</td>
</tr>
<tr>
<td></td>
<td>Labor</td>
<td>-0.071 (0.084)</td>
</tr>
<tr>
<td></td>
<td>Newborn baby</td>
<td>0.398** (0.129)</td>
</tr>
<tr>
<td></td>
<td>Income ($000)</td>
<td>0.004** (0.001)</td>
</tr>
<tr>
<td></td>
<td>PPO enrollee</td>
<td>-0.061 (0.072)</td>
</tr>
<tr>
<td>Interactions: Cancer Services</td>
<td>Cancer</td>
<td>0.073 (0.082)</td>
</tr>
<tr>
<td></td>
<td>Income ($000)</td>
<td>-0.005** (0.001)</td>
</tr>
<tr>
<td></td>
<td>PPO enrollee</td>
<td>0.087 (0.056)</td>
</tr>
<tr>
<td>Interactions: Labor Services</td>
<td>Labor</td>
<td>3.544** (0.391)</td>
</tr>
<tr>
<td></td>
<td>Newborn baby</td>
<td>3.116** (0.487)</td>
</tr>
<tr>
<td></td>
<td>Income ($000)</td>
<td>-0.003** (0.002)</td>
</tr>
<tr>
<td></td>
<td>PPO enrollee</td>
<td>0.045 (0.077)</td>
</tr>
<tr>
<td></td>
<td>Hospital fixed effects</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Pseudo-Rsquared</td>
<td>0.43</td>
</tr>
</tbody>
</table>

N = 28,666 encounters. SEs in parentheses; **significant at p=0.05; *significant at p=0.1
The interaction terms between PPO dummies and hospital characteristics are almost all insignificant. This provides support for the assumption that indemnity and PPO enrollees have identical preferences over hospitals conditional on demographics and diagnosis. Most of the interactions between patient income and hospital characteristics are either positive or insignificant. This probably implies that higher-income individuals have a stronger preference for high-tech hospitals and more personalized care, rather than that they are willing to pay more for their services, although the latter may also be relevant for PPO enrollees choosing out-of-network hospitals. The fit of the overall model is reasonable given that only a cross-section of data is used: the pseudo-$R^2$ is 0.43\textsuperscript{40}.

The results of the regression of the predicted hospital dummy coefficients on hospital characteristics are shown in Table 6. Many of the hospital characteristics included (particularly the hospital service variables, each of which was set to 1 if the hospital offered the relevant service, and 0 if not) are correlated with each other, making the individual results difficult to interpret. However, it is clear from the results that consumers place a positive value on the number of nurses per bed, the number of doctors per bed, and overall on hospital accreditation. Hospitals at a higher distance from the city center are more popular with patients, probably because distance from the patient’s home is negatively correlated, on average, with distance from the city. The fit of this model is similar to the first stage: the $R^2$ is 0.44\textsuperscript{41}.

\textsuperscript{40}The pseudo-$R^2$ is defined as $1 - (L_1/L_0)$, where $L_0$ is the constant-only log likelihood and $L_1$ is the full model log-likelihood. The model chi-squared(261) = 74,810, with p-value=0.0000.

\textsuperscript{41}As already noted, almost 100% of the variance here is due to the sampling error in $\delta$ ($\mu$), rather than to unobserved hospital quality $\eta$. 
Table 6: Regression of Hospital Dummy Coefficients on Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neonatal Intensive Care</td>
<td>-1.79 (1.46)</td>
</tr>
<tr>
<td>Angioplasty</td>
<td>-1.51 (1.61)</td>
</tr>
<tr>
<td>Cardiac Catheterization Laboratory</td>
<td>5.90** (1.82)</td>
</tr>
<tr>
<td>Computed-tomography scanner</td>
<td>6.53* (3.88)</td>
</tr>
<tr>
<td>Positron emission tomography</td>
<td>4.55** (1.57)</td>
</tr>
<tr>
<td>Single photon emission computerized tomography</td>
<td>-3.23** (1.09)</td>
</tr>
<tr>
<td>Oncology services</td>
<td>2.90 (2.08)</td>
</tr>
<tr>
<td>Obstetric services</td>
<td>-1.93 (1.69)</td>
</tr>
<tr>
<td>Emergency Department</td>
<td>-4.29* (2.20)</td>
</tr>
<tr>
<td>Breast cancer screening/mammograms</td>
<td>-4.44* (2.49)</td>
</tr>
<tr>
<td>Burn care</td>
<td>2.10 (1.88)</td>
</tr>
<tr>
<td>Alcohol/drug abuse inpatient care</td>
<td>0.51 (1.24)</td>
</tr>
<tr>
<td>Number of beds</td>
<td>0.01** (0.004)</td>
</tr>
<tr>
<td>Distance from City Hall</td>
<td>0.63** (0.19)</td>
</tr>
<tr>
<td>Distance from City Hall squared</td>
<td>-0.02** (0.01)</td>
</tr>
<tr>
<td>Registered nurses per bed</td>
<td>28.18** (4.97)</td>
</tr>
<tr>
<td>Nurses per bed squared</td>
<td>-9.74** (1.76)</td>
</tr>
<tr>
<td>Doctors per bed</td>
<td>3.78** (1.79)</td>
</tr>
<tr>
<td>JCAHO accreditation</td>
<td>6.83* (3.45)</td>
</tr>
<tr>
<td>Cancer Program approved by ACS</td>
<td>4.32** (1.72)</td>
</tr>
<tr>
<td>Residency Training Program</td>
<td>-4.46** (1.46)</td>
</tr>
<tr>
<td>Medical School</td>
<td>4.72** (1.36)</td>
</tr>
<tr>
<td>Member of Council of Teaching Hospitals of the Association of American Medical Colleges</td>
<td>-0.29 (1.86)</td>
</tr>
<tr>
<td>Independent Practice Association - hospital</td>
<td>5.27** (1.15)</td>
</tr>
<tr>
<td>Foundation</td>
<td>-6.79** (2.04)</td>
</tr>
<tr>
<td>Indemnity Fee for Service Plan - hospital</td>
<td>2.56 (2.27)</td>
</tr>
<tr>
<td>Primarily osteopathic hospital</td>
<td>1.56 (3.72)</td>
</tr>
<tr>
<td>Operates subsidiary corporations</td>
<td>2.24** (1.05)</td>
</tr>
<tr>
<td>Controlled/owned by county</td>
<td>-9.51** (3.80)</td>
</tr>
<tr>
<td>Controlled/owned by Church</td>
<td>-4.49** (1.41)</td>
</tr>
<tr>
<td>Controlled/owned by For-profit partnership</td>
<td>19.69** (5.06)</td>
</tr>
<tr>
<td>Constant</td>
<td>-32.10** (5.57)</td>
</tr>
<tr>
<td>Market fixed effects</td>
<td>Yes</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.44</td>
</tr>
</tbody>
</table>

N = 434 hospital-years. Robust standard errors are reported in parentheses; ** significant at p=0.05; * significant at p=0.1
2.6.2 Plan Demand: Results

Table 7 sets out the results of the logit demand specification under the assumption that \( \eta = 0 \). Results are reported for three specifications. Specification 1, the baseline model, has no fixed effects. Specification 2 includes large insurer fixed effects which measure consumer preferences for specific national plans (such as the "Aetna effect"). Specification 3 includes both large insurer and market fixed effects; the latter control for unobserved differences in the outside option (being uninsured) across markets.

Including large insurer fixed effects improves the fit of the model; adding market fixed effects improves it still further (the \( R^2 \) increases from 0.36 to 0.59 to 0.67 when both types of fixed effect are included). The coefficient on premium is negative but not significant in any of the specifications; its magnitude suggests an elasticity of roughly -1.3, which is broadly consistent with the existing literature\(^{42}\).

The key variable, \( EU_{rep_{jm}} \), has a positive coefficient in all three specifications, implying that consumers care about hospital networks: a plan’s market share would be predicted to decrease if it excluded hospitals. The magnitude of the coefficient increases when market fixed effects are added\(^{43}\); it is significant at \( p=0.05 \) in Specification 3. The most straightforward way to interpret the result is to compare the marginal effect of \( EU_{rep_{jm}} \) to that of plan premiums. The results imply that a 1 standard deviation increase in expected utility from the hospital network is equivalent to a $39 reduction in premium per member per month (a change in premium of a little less than one standard deviation).


\(^{43}\) The reason for the increase is probably that consumers are more likely to choose the outside option (being uninsured) in markets where hospitals on average offer higher utility. This effect will be absorbed into and bias down the coefficient on \( EU_{rep_{jm}} \) unless market fixed effects are included. The fixed effects account for differences in the outside option across markets, leaving the expected utility coefficient to measure the effect of differences in \( EU_{rep_{jm}} \) across plans within each market. For this reason, Specification 3 gives the most relevant results to the question being considered and will be used for the remainder of the paper.
Table 7: Plan Demand Results, Logit Specification

<table>
<thead>
<tr>
<th></th>
<th>No Fixed Effects</th>
<th>Large Plan Fixed Effects</th>
<th>Plan and Market FEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Premium ($00 pmpm)</td>
<td>-1.26 (3.15)</td>
<td>-1.09 (1.67)</td>
<td>-0.92 (1.10)</td>
</tr>
<tr>
<td>EUrep</td>
<td>0.14 (0.14)</td>
<td>0.22* (0.11)</td>
<td>0.55** (0.14)</td>
</tr>
<tr>
<td>Physicians/1000 popln</td>
<td>0.30** (0.13)</td>
<td>0.23** (0.08)</td>
<td>0.21** (0.07)</td>
</tr>
<tr>
<td>Breast cancer screening</td>
<td>4.77 (4.66)</td>
<td>-1.71 (3.21)</td>
<td>-0.36 (2.48)</td>
</tr>
<tr>
<td>Cervical cancer screening</td>
<td>4.66** (1.83)</td>
<td>4.19** (1.69)</td>
<td>4.46** (1.75)</td>
</tr>
<tr>
<td>Check-ups after delivery</td>
<td>-0.53 (1.64)</td>
<td>0.26 (1.07)</td>
<td>0.14 (1.03)</td>
</tr>
<tr>
<td>Diabetic eye exams</td>
<td>0.39 (1.68)</td>
<td>-0.83 (1.19)</td>
<td>-1.20 (1.08)</td>
</tr>
<tr>
<td>Adolescent immunizn 1</td>
<td>-0.77 (1.29)</td>
<td>-2.19* (1.08)</td>
<td>-4.11** (1.16)</td>
</tr>
<tr>
<td>Adolescent immunizn 2</td>
<td>-1.74 (1.83)</td>
<td>2.19* (1.47)</td>
<td>3.16** (1.40)</td>
</tr>
<tr>
<td>Advice on smoking</td>
<td>-7.07** (2.76)</td>
<td>2.75* (1.90)</td>
<td>6.20** (1.80)</td>
</tr>
<tr>
<td>Mental illness check-ups</td>
<td>-0.34 (2.46)</td>
<td>2.02 (1.79)</td>
<td>2.67** (1.25)</td>
</tr>
<tr>
<td>Care quickly</td>
<td>6.64 (6.10)</td>
<td>4.55 (4.47)</td>
<td>0.75 (3.93)</td>
</tr>
<tr>
<td>Care needed</td>
<td>3.77 (6.58)</td>
<td>-1.85 (4.30)</td>
<td>0.81 (3.60)</td>
</tr>
<tr>
<td>Plan age: 0 - 2 years</td>
<td>-1.30 (0.98)</td>
<td>0.52 (1.17)</td>
<td>1.33 (0.94)</td>
</tr>
<tr>
<td>Plan age: 3 - 5 years</td>
<td>-2.31* (1.43)</td>
<td>-0.97* (0.53)</td>
<td>-0.63 (0.42)</td>
</tr>
<tr>
<td>Plan age: 6 - 9 years</td>
<td>-1.63* (0.83)</td>
<td>-0.26 (0.24)</td>
<td>-0.25 (0.22)</td>
</tr>
<tr>
<td>POS plan</td>
<td>-1.35** (0.22)</td>
<td>-1.10** (0.13)</td>
<td>-1.11** (0.13)</td>
</tr>
<tr>
<td>Constant</td>
<td>-9.38 (7.47)</td>
<td>-6.75* (3.78)</td>
<td>-10.94** (2.89)</td>
</tr>
<tr>
<td>Large plan fixed effects</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Market fixed effects</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.362</td>
<td>0.592</td>
<td>0.671</td>
</tr>
</tbody>
</table>

N=559. Standard errors (adjusted for the three-stage estimation process) are reported in parentheses. ** significant at p=0.05; * significant at p=0.1.
The coefficient on the number of physicians per 1000 population is also positive and significant, implying that consumers prefer plans with large physician networks. The coefficients on HEDIS and CAHPS variables are difficult to interpret because the variables are fairly highly correlated with each other. However, most of the coefficients that are significant are positive as expected.

Consumers seem to significantly prefer HMO to POS plans. At first sight this is surprising, particularly since POS plans offer more flexibility to use hospitals outside the network. The reason is probably twofold. First, premiums set by POS plans are in general higher than their HMO equivalent; unfortunately the available data in general reports the same premium amount for both HMO and POS plans in a given organization. Second, there is a selection problem: plans would prefer to attract HMO enrollees, who have more restricted benefits, rather than more expensive POS enrollees, and probably take unobserved steps (e.g. through promotions) to do so. This issue is beyond the scope of this chapter; I assume no selection. Both these complicating factors will lead to a negative bias on the POS dummy coefficient\textsuperscript{44}.

Several robustness checks are used to test the stability of the results to changes in assumptions and in the underlying data. First, the methodology is repeated using just the plans whose enrollment is reported by AIS at the PMSA level\textsuperscript{45}. Second, I conduct the test of the assumption that unobserved hospital quality $\eta = 0$ described in Section 2.5.2.2. Finally, the analysis is repeated using the formula in footnote\textsuperscript{44}

\textsuperscript{44}If POS plan premiums were $118 (84\%)$ higher than those of HMOs on average, this would explain the entire POS coefficient without any assumption regarding selection. Recent industry publications suggest a premium difference of around 30\% (see for example, the Ohio Chamber of Commerce Task Force Report on Provider Access 2000), implying that the selection effect may also be important. The estimated coefficients for the missing variable dummies, and for the insurer and market fixed effects in the relevant specifications, are not reported here. All but one of the missing variable dummies have insignificant coefficients in Specification 3. All the insurer fixed effects have negative coefficients, many of which are significant. Since insurer fixed effects are included only for the largest plans, the implication is that, all else equal, consumers prefer to enrol in local plans rather than those that operate at a national level.

\textsuperscript{45}See Appendix A for details on this issue.
26 to deal with the error term \( \varepsilon \). All three robustness checks yield results that are extremely similar to those reported.

The results from the full demand specification are reported in Table 8. Estimates from Specification 3 (including both large insurer and market fixed effects) are reported; the equivalent estimates from the logit framework are also reported here for ease of comparison. The two sets of results are very similar. The coefficient on expected utility is higher in the full model: this makes sense because the distribution of consumer characteristics is now being taken into account. The coefficient on premium/income is very small and not significant: higher-income consumers seem not to have a different price elasticity of demand for health plans from their lower-income neighbours\(^{46}\). Finally, some of the standard errors from the full specification are higher than those in the logit model, implying that too much may be being asked of the limited data available. Since the magnitudes of the coefficients are so similar, however, I take the full demand estimation results to be reasonable and use them in what follows.

\(^{46}\)Here premium is measured in $00 per member per month and income in $000 per year.
### Table 8: Plan Demand Results, Logit and Full Specifications

<table>
<thead>
<tr>
<th></th>
<th>Logit Specification</th>
<th>Full Demand Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Premium ($00 pmpm)</td>
<td>-0.92 (1.10)</td>
<td>-0.94 (1.13)</td>
</tr>
<tr>
<td>Expected utility from hospital network</td>
<td>0.55** (0.14)</td>
<td>0.59** (0.21)</td>
</tr>
<tr>
<td>Premium / Income</td>
<td>-</td>
<td>0.002 (43.9)</td>
</tr>
<tr>
<td>Physicians per 1000 population</td>
<td>0.21** (0.07)</td>
<td>0.21** (0.09)</td>
</tr>
<tr>
<td>Breast cancer screening</td>
<td>-0.36 (2.48)</td>
<td>-0.38 (2.66)</td>
</tr>
<tr>
<td>Cervical cancer screening</td>
<td>4.46** (1.75)</td>
<td>4.40** (2.09)</td>
</tr>
<tr>
<td>Check-ups after delivery</td>
<td>0.14 (1.03)</td>
<td>0.18 (1.38)</td>
</tr>
<tr>
<td>Diabetic eye exams</td>
<td>-1.20 (1.08)</td>
<td>-1.19 (1.60)</td>
</tr>
<tr>
<td>Adolescent immunization 1</td>
<td>-4.11** (1.16)</td>
<td>-4.11** (1.17)</td>
</tr>
<tr>
<td>Adolescent immunization 2</td>
<td>3.16** (1.40)</td>
<td>3.08 (3.76)</td>
</tr>
<tr>
<td>Advice on smoking</td>
<td>6.20** (1.80)</td>
<td>6.17** (2.08)</td>
</tr>
<tr>
<td>Mental illness check-ups</td>
<td>2.67** (1.25)</td>
<td>2.70** (1.30)</td>
</tr>
<tr>
<td>Care quickly</td>
<td>0.75 (3.93)</td>
<td>0.78 (5.63)</td>
</tr>
<tr>
<td>Care needed</td>
<td>0.81 (3.60)</td>
<td>0.85 (3.99)</td>
</tr>
<tr>
<td>Plan age: 0 - 2 years</td>
<td>1.33 (0.94)</td>
<td>1.36 (0.97)</td>
</tr>
<tr>
<td>Plan age: 3 - 5 years</td>
<td>-0.63 (0.42)</td>
<td>-0.64 (1.97)</td>
</tr>
<tr>
<td>Plan age: 6 - 9 years</td>
<td>-0.25 (0.22)</td>
<td>-0.25 (0.58)</td>
</tr>
<tr>
<td>POS plan</td>
<td>-1.11** (0.13)</td>
<td>-1.11** (0.13)</td>
</tr>
<tr>
<td>Constant</td>
<td>-10.94** (2.89)</td>
<td>-10.50* (5.65)</td>
</tr>
<tr>
<td>Large plan fixed effects</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Market fixed effects</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

N=559. Standard errors (adjusted for the three-stage estimation process) are reported in parentheses. ** significant at p=0.05; * significant at p=0.1.
2.7 Welfare Implications of Selective Contracting

The demand estimates provide information on the tradeoffs consumers are willing to make when choosing health plans: that is, on the extent to which hospital choice affects consumers’ welfare and choice of plans compared to other plan characteristics such as premium, size of the physician network and clinical quality. The results indicate that consumers significantly prefer plans that offer larger hospital networks.

The next step is to quantify the welfare benefits attached to an increase in network size. I consider the impact of a move from the observed set of networks in each market to a hypothesized equilibrium in which every plan offers every hospital. The two dimensions of welfare are considered in turn: first consumer surplus and then provider and insurer profits.

2.7.1 A Simple Measure of Consumer Surplus

The methodology used in this section follows that discussed by Nevo (2001(b)) and based on McFadden (1981). Consumer \(i\)'s expected gain from a change in a plan's hospital network, assuming that no other plan characteristics (including its premium) change, is:

\[
\Delta_i = u_i^t - u_i^{t-1}
\]

where \(u_i^t\) and \(u_i^{t-1}\) are defined by:

\[
u_i^t = \mathbb{E}_\omega \max_j (V_{ijm}^t + \omega_{ijm})
\]

\[
u_i^{t-1} = \mathbb{E}_\omega \max_j (V_{ijm}^{t-1} + \omega_{ijm})
\]

\footnote{There is only one market in my sample in which all plans already offer an unrestricted choice of hospitals: this is Buffalo NY. This market will therefore exhibit a zero welfare increase from the network change being considered.}

47
with \( w_{ijm} \) defined as in equation (12), so that:

\[
V_{ijm}^t = \xi_{jm} + z_{jm} \theta + \gamma_1 EU_{ijm}^t + \gamma_2 \frac{\text{prem}_{jm}}{y_i}
\]  

(15)

Note that this is the expected welfare gain from the perspective of the econometrician given the available data. A dollar-valued measure of welfare can be obtained using the method suggested by Hicks (1939) to create the equivalent variation (EV). The EV is the change in consumer wealth that would be equivalent to the change in consumer welfare due to the network modification. McFadden (1981) shows that:

\[
EV_{it} = \frac{1}{\alpha_i} (u_i^t - u_i^{t-1})
\]

where \( \alpha_i \) is the negative of the coefficient on premium in the plan utility equation. Integrating analytically over the extreme value distribution of \( \omega \) and summing over types of individual implies that:

\[
EV_m = \sum_i \frac{n_i}{\alpha_i} \left[ \ln \sum_{jem} \exp(V_{ijm}^t) - \ln \sum_{jem} \exp(V_{ijm}^{t-1}) \right]
\]

where \( n_i \) is the population in ZCTA-age-sex cell \( i \) and the difference between \( V_{ijm}^t \) and \( V_{ijm}^{t-1} \) comes solely from the change in the hospital network offered by each plan\(^{48}\).

The expression for \( EU_{ijm}^{t-1} \) is calculated using the specification in equation (10) and considering a network that includes all hospitals in the market. \( EU_{ijm}^t \) takes the value already calculated using equation (10) and the observed networks.

I also adjust for hospital capacity constraints. When a hospital is predicted to be over 85\% of its maximum capacity I reallocate patients randomly to non-capacity-constrained hospitals in the market. The adjustment affects patients’ hospital choices.

\(^{48}\)The utility offered by the outside option (being uninsured) is treated the same as that from any other choice: having normalized \( V_{0m} \) to zero for each market, we include a term \( e^0 = 1 \) in each log sum expression.
and therefore their values of $EU_{ijm}$; these values are used to adjust down each patient’s utility from each plan\textsuperscript{49}.

The results of this exercise are presented in Figure 3. Column 1 of the Table and Graph 1 give details on the EV per person per year: the dollar spending needed to compensate the average consumer in the market for a move back from unselective to observed networks. The variation in this measure across markets is high: from -$18 in Las Vegas NV to a very high $455 per person year in Austin TX. The median value is $9.89 per person per year. Aggregating over consumers, this implies a total welfare gain of $1.04 billion per year for the 43 major markets in the data\textsuperscript{50}.

\textsuperscript{49}The adjustment does not affect patients’ choices of plan; I assume that consumers expect to have access to every hospital on the plan’s list when they make their plan choice. Instead I account for the fact that some patients find, when they get sick, that they cannot access their preferred hospital and therefore receive less utility than expected from their chosen plan. The median equivalent variation falls only slightly as a result of this adjustment.

\textsuperscript{50}Median values are used to reduce the impact of outliers. The negative value in Las Vegas NV is caused because the move to unselective contracts displaces some high-valuation patients from capacity-constrained hospitals in favour of patients who value the hospital less. Only three of the 43 markets have negative EV per person values. Graph 1 of Figure 2 makes clear that Austin TX, which has an EV per person of $455 per year, is an extreme outlier. It includes a plan that is significantly different from others in the data. Scott and White Health Plan is managed by physicians, has strong links with the community, and is heavily focused on preventive rather than acute care. The demand model’s ability to identify the effect of this plan’s characteristics on consumer utility is limited: the specification implies that all attributes other than premiums and the hospital network affect all consumers equally, whereas Scott and White plan’s strengths are likely to appeal to some consumers more than others. The predicted average quality of the plan, and therefore the estimated welfare effects for this market, are therefore likely to be biased. However, this is an issue that probably only affects a few markets (Austin TX is clearly an extreme case), so reporting median results should be sufficient to control for the problem.
<table>
<thead>
<tr>
<th></th>
<th>Equivalent Value per person ($ per year)</th>
<th>Change in Producer Surplus per market ($ per year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$37.62</td>
<td>- $8.61 million</td>
</tr>
<tr>
<td>Median</td>
<td>$9.89</td>
<td>- $0.02 million</td>
</tr>
<tr>
<td>Maximum</td>
<td>$455.17</td>
<td>$288.2 million</td>
</tr>
<tr>
<td>Minimum</td>
<td>- $18.17</td>
<td>- $166.7 million</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>$77.87</td>
<td>$64.6 million</td>
</tr>
<tr>
<td>Total gain implied by medians</td>
<td>$1.04 billion</td>
<td>- $0.80 million</td>
</tr>
</tbody>
</table>

Graph 1: Equivalent Variation per person, $ per year

Graph 2: Change in Producer Surplus per market, $ million per year

Figure 3: Welfare Effects of a Move to Unrestricted Hospital Choice
One caveat should be mentioned here. The assumption that prices and premiums are fixed is unlikely to be realistic: part of the rationale for managed care was to allow plans to use the threat of exclusion from their networks as a lever to contain hospital prices. If this threat was removed, for example if plans were required by legislation to offer a free choice of providers, hospital prices might increase and bring premiums with them. Even if hospital prices were unchanged, plans might still increase their premiums when they moved to unrestricted hospital networks, simply in response to increased consumer willingness-to-pay. Since the analysis assumes away these issues the consumer surplus estimates represent an upper bound on the true impact of reducing selective contracting. In some cases a premium increase may simply represent a transfer from consumers to firms: this should therefore not have a significant effect on the calculation of total social welfare. However, there would be welfare effects in cases where the change prompted consumers to choose to be uninsured. Unfortunately the results in this chapter provide little guidance on the magnitude of these effects.

2.7.2 Welfare Analysis: Producer Surplus

The calculation of producer surplus too is fairly straightforward. I do not attempt to distinguish between hospital and plan profits since this would require estimating the prices paid to hospitals and therefore a full model of the supply side. Instead I consider the effect of the network change on the sum of hospital and plan profits (that is, on total producer surplus), again assuming that premiums remain fixed. The producer surplus generated by a given network depends on the number of consumers attracted to each plan and the cost of the hospitals they visit. I use the demand estimates to predict consumer flows to both plans and hospitals before and after the network change. These are used to calculate revenues (which are increasing in the number of consumers that choose high-premium plans) and costs (which increase if
consumers on average choose more expensive hospitals). The total producer surplus to be divided between plan $j$ in market $m$ and the hospitals in its network is:

$$PS_{jm} = \sum_i n_i s_{ijm} (prem_{jm} - \text{cost}_i)$$

where $n_i$ is again the population in ZCTA-age-sex cell $i$, $s_{ijm}$ is plan $j$'s share of type-$i$ people in market $m$ as specified in equation (14), and $\text{cost}_i$ is the cost of treating a person of type $i$\textsuperscript{51}. As before, define $p_{il}$ as the probability that consumer-type $i$ is admitted to hospital for diagnosis $l$. If $s_{ihl}$ is the probability that a type-$i$ person with diagnosis $l$ will visit hospital $h$ (the term in parentheses in equation (7)) and $\text{cost}_{hl}$ is the cost that hospital $h$ incurs by treating a patient with diagnosis $l$, then $\text{cost}_i$ is given by:

$$\text{cost}_i = \sum_l p_{il} E(\text{cost}_{il}) = \sum_l p_{il} \sum_{hH_{jm}} s_{ihl} \text{cost}_{hl}$$

The change in producer surplus when the plan switches to offering a free choice of hospitals, which will be divided between the plan and all the hospitals in its network, is given by:

$$PS_{jm}^{\text{change}} = PS_{jm}^{\text{choice}} - PS_{jm}^{\text{observed}}$$

where $PS_{jm}^{\text{observed}}$ is the producer surplus predicted under the observed contracts and $PS_{jm}^{\text{choice}}$ is that predicted when all plans in the market contract with all hospitals. The total change in producer surplus in the market is the sum of $PS_{jm}^{\text{change}}$ over all plans $j$ in market $m$. As in the consumer surplus calculation I take account of hospital capacity constraints by reallocating patients randomly from hospitals that

\textsuperscript{51}I assume that plans do not know consumers’ idiosyncratic error terms, $\varepsilon$ and $\omega$, when they make their network decisions. They therefore predict consumer flows using share equations (7) and (13), which imply taking an expectation over the error terms.
are predicted to be over 85% of maximum capacity to other providers in the market. This reallocation affects the calculation of treatment costs.

A number of assumptions are made to simplify the analysis. First, I assume that plans have no variable costs except the payments made to hospitals. Any other variable costs would affect the surplus change calculation, even if they are constant across plans, because adding a new hospital could persuade consumers to switch from the outside option (being uninsured) to being insured, creating new variable costs that are not captured by the analysis. However, I do not have access to data on plan variable costs and therefore cannot measure these effects. Data on total hospital expenses per admission are taken from the AHA 2001. Diagnosis-specific cost data would be preferable (since patient flows are predicted for each diagnosis separately) but were not available for this study. Finally, as in the consumer surplus calculations, I assume that plan premiums are fixed. As discussed above this will affect the total welfare calculation to the extent that premium increases, which are likely when plans expand their networks, lead consumers to choose to be uninsured.

The results of the analysis are set out in Figure 3. The hypothesized network expansion causes a median reduction in producer surplus of just $0.02 million per market per year. As with consumer surplus, there is considerable variation across markets, from an increase of $288 million per year to a decrease of $167 million per year. The median figure implies a total reduction in producer surplus of $0.80 million per year across the 43 markets in the sample.

2.8 Discussion and Conclusion

This chapter investigates the welfare effects of health plan restrictions on hospital choice. I use a three-stage method to estimate demand for health plans conditional on

52The analysis does allow for the existence of additional fixed costs, since these would cancel out when we consider the surplus change from a change in networks.
the hospital networks they offer. The results indicate that consumers place a positive and significant weight on their expected utility from hospital networks when choosing their plans. I find significant welfare effects of a move from plans’ observed networks to a hypothesized equilibrium where every plan offers access to every hospital in its market. The predicted results are shown in the following table.

<table>
<thead>
<tr>
<th>Welfare Effect</th>
<th>Predicted Change per year from Move to Unselective Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer Surplus</td>
<td>$1.04 billion</td>
</tr>
<tr>
<td>Producer Surplus</td>
<td>-$0.80 million</td>
</tr>
<tr>
<td>Total</td>
<td>$1.04 billion</td>
</tr>
</tbody>
</table>

The predicted gain in consumer surplus far outweighs the reduction in producer surplus: overall the results imply a benefit to society of just under $1.04 billion per year.

One particular caveat should be emphasized. It concerns the assumption of fixed prices. In reality a move to unselective contracting, particularly if it was achieved through legislation, would probably lead to an increase in premiums. This would affect the total welfare predictions: a rise in premiums would increase the uninsured proportion of the population with obvious negative welfare effects. An increase in the premium of one plan relative to another would also affect the allocation of consumers across plans. In addition, the fixed price assumption may bias the consumer surplus estimates up and the producer surplus estimates down; it therefore affects the certainty with which we can predict the distribution of total surplus. Given this caveat, it may be helpful to view the results as essentially considering just one effect of selective contracting: they estimate the loss to consumers from restricted hospital choice, ignoring the effect on costs, prices and health plan premiums. The welfare gain from reduced premiums may outweigh the loss from restricted hospital choice.
A model of the supply side is needed to understand these effects more fully.

The results given here raise an additional question: given that consumers prefer choice, why do plans in many markets choose selective networks? There are several reasons why plans may never agree to networks that offer unrestricted choice of hospitals. First, as already noted, the price-setting interaction between plans and hospitals creates incentives for selective contracting. A plan may choose to exclude hospitals that demand a high share of the surplus they generate (although we might expect the threat of exclusion to be enough to keep hospital prices down without plans actually acting on this threat). A capacity-constrained hospital may prefer to contract with only the plans that have the highest willingness-to-pay. Hospital networks also affect the level of premium competition between plans. It may be optimal for plans to "split the market" by contracting with different sets of hospitals and using the resulting product differentiation to avoid premium competition. Finally, insurers may select their hospital networks specifically to attract certain population segments (such as younger, healthier types) and deter others; this is essentially no different from other types of adverse selection where plans choose the range of services they cover to attract low-cost enrollees. All these strategic issues, combined with simple cost effects such as contracting costs and potential high costs of care at certain hospitals, may result in a set of potential equilibria where plans never offer unrestricted choice of hospitals. Again, a fuller model of the supply side of the market is needed to gain a better understanding of these issues. I introduce and estimate such a model in Chapter 3.
2.9 Appendix A: Details on Hospital and Plan Data

2.9.1 Hospital Dataset

The MEDSTAT MarketScan Research Database lists the date, the patient’s age, sex, zip code and primary diagnosis (defined using ICD-9-CM codes), the identity of the hospital, and the type of plan (managed care; indemnity; PPO etc.) for every admission. I include six diagnosis categories in my analyses: cardiac; cancer; labor; newborn baby; digestive diseases and neurological diseases. The specific 1998 ICD-9-CM codes included in each category are listed in Table 9. These six categories account for 55% of the full dataset. The seventh diagnosis category, "other diagnoses", comprises all other diagnoses included in the data. I also identify emergency admissions using the place of service and the type of service for each admission.

<table>
<thead>
<tr>
<th>Category</th>
<th>1998 ICD-9-CM codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiac</td>
<td>393-398; 401-405; 410-417; 420-429</td>
</tr>
<tr>
<td>Cancer</td>
<td>140-239</td>
</tr>
<tr>
<td>Neurological</td>
<td>320-326; 330-337; 340-359</td>
</tr>
<tr>
<td>Digestive</td>
<td>520-579</td>
</tr>
<tr>
<td>Labor</td>
<td>644, 647, 648, 650-677, V22-V24, V27</td>
</tr>
<tr>
<td>Newborn baby</td>
<td>V29-V39</td>
</tr>
</tbody>
</table>

The MEDSTAT data is supplemented with data from the American Hospital Association (AHA) for 1997 and 1998. This contains hospital characteristics, including details of location, ownership, accreditation, services provided, number of beds and numbers of admissions, for every hospital in the USA. I define distance from the patient’s home to the hospital using the five-digit zip codes of each; distance of the hospital from the city center is defined as distance from the hospital’s five-digit zip
code to that of the City Hall. Four "service" variables are created from the AHA data to summarize the services offered by each hospital: these will be interacted with consumer characteristics in the hospital choice model. The summary variables cover cardiac services; imaging; cancer; and birth services. Each hospital is rated on a scale from 0 to 1, where 0 indicates that no services within this category are provided by the hospital, and a higher rating indicates that less common (assumed to be higher-tech) service in the category is offered. The procedures included in each category are listed in Table 10.

Table 10: Procedures Included in each Hospital Service Category

<table>
<thead>
<tr>
<th>Cardiac</th>
<th>Imaging</th>
<th>Cancer</th>
<th>Births</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC laboratory</td>
<td>Ultrasound</td>
<td>Oncology services</td>
<td>Obstetric care</td>
</tr>
<tr>
<td>Cardiac IC</td>
<td>CT scans</td>
<td>Radiation therapy</td>
<td>Birthing room</td>
</tr>
<tr>
<td>Angioplasty</td>
<td>MRI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open heart surgery</td>
<td>SPECT</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PET</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The exact methodology for rating hospitals is as follows. If the hospital provides none of the services, its rating = 0. If it provides the least common service, its rating = 1. If it offers some service X but not the least common service, its rating = (1 - x) / (1 - y), where x = the percent of hospitals offering service X and y = the percent of hospitals offering the least common service.

The set of hospitals operating in each market is defined by the zip codes of consumers considered to reside in the market, and the distance they are likely to be willing to travel to hospital. I consider patients whose home zip code is within the Primary Metropolitan Statistical Area (PMSA). Previous papers have considered hospitals within boundaries such as counties or states, those within 30 miles or 50 miles of the city center, or those hospitals within 30 or 50 miles of the individual patient’s
home zip code. I limit the analysis to manageable proportions, while still including a reasonable sample of hospitals, by defining the market to include all hospitals within 30 miles of the city center. I include in the patient choice set every general medical/surgical hospital, other than those owned by the federal government, in the relevant market area.

I consider the 11 largest markets in the MEDSTAT data (those with over 1000 observations per market-year). The markets observed are located in Massachusetts, Illinois, Arizona, Washington, Florida, and Michigan States; five of the eleven markets are in Michigan. There are a total of 237 hospitals and 29,657 encounters in these market-years. A number of observations are lost from the analysis because of missing hospital or individual data: if a variable (such as home zip code) is missing for a given individual, that individual is excluded from the analysis. If a variable (such as services provided or location) is missing for a given hospital, the missing data is filled in using surrounding years of AHA data where possible; otherwise that hospital, and all individuals who chose it, are excluded. Thus the final dataset comprises 217 hospitals, 434 hospital-years and 28,666 encounters in total.

53 See, for example, Tay (2003)

54 According to this market definition, the Boston market contains 37 hospitals; Chicago contains 72. If I included every hospital within 50 miles of the city center, Boston would have 72 hospitals and Chicago would have 86.

55 The markets are Boston MA, Chicago IL, Dayton-Springfield OH, Orlando FL, Phoenix AZ, Seattle WA, and Detroit, Grand Rapids, Kalamazoo-Battle Creek, Lansing, and Saginaw-Bay, all in MI.

56 Each encounter is an individual admission; 51% of these are for PPO enrollees. If a single patient is admitted more than once in the two-year time period, I assume that the admissions represent independent choices.

57 This attrition will bias the results if it is non-random. The characteristics of consumers whose data is missing are unlikely to be biased in any direction. While it is possible that smaller hospitals are less likely to report data to AHA, this seems unlikely because the existence of missing data in one year does not predict whether data in the subsequent or previous year will be missing.
2.9.2 Plan Networks

The 516 plans for which network data was collected comprise all HMO/POS plans in 43 markets, as defined by the Atlantic Information Services data discussed in Section 2.4.3. The list of potential hospitals comprises all general medical/surgical hospitals listed by AHA 2001 that have more than 150 beds, are not owned by the federal government, and are located in the relevant PMSA. In smaller PMSAs, where there were fewer than 10 such hospitals, facilities with over 100 beds were included.

The expected utility calculation requires the different types of consumers to be defined using variables observed in the aggregate plan data; I therefore define types by sex, age group (0-17; 18-34; 35-44; 45-54; 55-64) and zip code tabulation area (ZCTA) of residence. There are 10 cells per ZCTA (5 agegroups and 2 sexes within each), and a total of 6363 ZCTAs across the 43 markets (an average of 148 in each market). The number of people in each ZCTA-age-sex cell is found using Census 2000 data from GeoLytics. Diagnosis probabilities given age, sex, and admission to hospital are estimated from the MEDSTAT data using probit analysis; probabilities of admission

---

58 The markets are: Atlanta GA, Austin TX, Baltimore MD, Boston MA, Buffalo NY, Charlotte NC, Chicago IL, Cincinnati OH, Cleveland OH, Columbus OH, Dallas TX, Denver CO, Detroit MI, Fort Worth TX, Houston TX, Indianapolis IN, Jacksonville FL, Kansas City MO, Las Vegas NV, Los Angeles CA, Miami FL, Milwaukee WI, Minneapolis MN, New Orleans LA, Norfolk VA, Oakland CA, Orange County CA, Orlando FL, Philadelphia PA, Phoenix AZ, Pittsburgh PA, Portland OR, Sacramento CA, St. Louis MO, Salt Lake City UT, San Antonio TX, San Diego CA, San Francisco CA, San Jose CA, Seattle WA, Tampa FL, Washington DC, and West Palm Beach FL.

59 This simple definition of the hospitals in the market was not used in the hospital dataset because, in some of the smaller markets included in that data, it seems likely that patients would choose to travel outside the MSA to go to hospital. The 30-mile radius definition increased the hospital choice set in these markets. The plan dataset covers only larger MSAs, for which a 30-mile radius would reduce the hospital choice set in almost all cases.

60 The analysis requires ZCTA-age-sex cells to be defined for an entire MSA. ZCTAs, rather than zip codes, were used because Census data from 2000 onwards uses the former rather than the latter. Some ZCTAs cross MSA boundaries, implying that this aggregation to MSA level is not exact. However, the difference in definitions accounts for a less than 8% difference between the MSA population defined by the MSA boundary and that defined by the boundaries of its constituent ZCTAS. Since residents of contiguous ZCTAs are likely to have similar characteristics (income, distance from hospitals and age/sex distribution), I assume that ZCTA and MSA boundaries are perfectly aligned.
AHA data for 2001 (the most recent year for which data was available) was used in the expected utility calculation. A number of hospitals have missing data for AHA 2001. To avoid dropping these from the choice set, the missing data was filled in using previous years of AHA data where possible, and if necessary (for 16 hospitals) using information provided in individual hospital websites\(^6^1\).

Two plans in my sample offer none of the hospitals in the defined choice set but are licensed in and have significant numbers of enrollees in the market. Each of these offers a number of smaller hospitals in the relevant area (too small to be included in the base list of hospitals used to define plans’ hospital networks). The expected utility from these plans’ networks cannot be calculated using my dataset. I assume that a substantially restricted choice set reduces expected utility compared to other plans in the market; I therefore assign each type of consumer the minimum possible utility from any hospital in the market, given each diagnosis, if he or she chooses one of these two plans.

### 2.9.3 Plan Characteristics

The two datasets from Atlantic Information Services are *The HMO Enrollment Report* and *HMO Directory 2002*. Both are based on plan state insurance filings. The enrollment data gives detailed enrollment (below the product level, e.g. there may be several HMO products within Aetna Boston) for every HMO and POS plan in 40 major markets in the USA\(^6^2\). The characteristic data cover all commercial health

---

\(^6^1\) Data was taken from hospital websites in 2003. The same data were not used to fill in 1997/98 characteristics since hospitals are likely to have changed the services offered over the intervening five year period.

\(^6^2\) AIS works with individual plans to disaggregate their base data. The data includes some Medicare-only and Medicaid-only insurers as well as commercial plans; my analysis excludes the former and examines only plans that accept commercial business. I also exclude plans with fewer
plans in the USA.

The unit of observation for the NCQA data is the NCQA plan identifier, which does not correspond exactly to the plan identifier for the AIS enrollment data. The NCQA data was matched to the AIS data at the insurer–product–market level; in cases where multiple NCQA plans correspond to one AIS plan, the mean rating over NCQA plans was used. Similarly, the Weiss data and AIS characteristic data do not correspond perfectly to AIS enrollment data plan identifiers: both contain more aggregated data (for example, characteristics are provided for Aetna Florida rather than Aetna Jacksonville; Aetna Miami etc.) and often covers only HMOs. The two datasets were matched to the AIS enrollment data at the insurer–product–market level where possible, and at the insurer–market level otherwise. Aggregate data was matched to all plans within the geographic area, and if no POS data was given separately, the plan’s HMO characteristics were matched to both HMO and POS plan types.

Missing data represents a significant issue. Of the 516 observations considered, 162 (31.4%) do not have HEDIS data and 212 (41.1%) do not have CAHPS data. Most of the plans without data did not respond to NCQA data requests; many did not provide information for any of the HEDIS and CAHPS categories used in this chapter. There are similar problems with the price measure used: premium earned per enrolled member per month. In most cases (354 observations) this measure is calculated from AIS data on both premium and enrollment. Both inputs come from Weiss for 120 observations (where one or both pieces of information was missing in the AIS data). Price data are missing for 42 of the 516 plans (8.2% of observations). Dropping plans with missing data (particularly NCQA data) could cause selection bias because the plans that failed to respond to NCQA requests are likely to be

---

than 100 enrollees and/or no hospitals in the relevant market; I assume that these plans primarily serve neighbouring areas. AIS publishes the data for 40 markets. I disaggregate this to 43 markets; see later in this section for methodology.
smaller or have lower quality than those that provided data. Instead these plans are
included and dummy variables that indicate missing premium and characteristic data
are added to the specification.\footnote{The other alternative would be to fill in the missing observations with previous years’ data. This is not attempted, both because plans with missing data in 2000 often had missing data in previous years and because cross-year correlations in reported data for a given plan are low (much lower than for the hospital data where this approach was used).}

The AIS enrollment data is converted to market share using the total non-elderly
population of the MSA as defined by the Census Bureau in 2000 as the denominator.\footnote{I exclude people aged over 64 from the plan demand equation in order to exclude Medicare enrollees. Patients over age 64 are not excluded from the hospital dataset. This is unnecessary since only PPO/indemnity enrollees are included: their preferences over hospitals are assumed to be identical to those of other consumers choosing private insurance conditional on diagnosis, income and location.}
The share uninsured and the share in PPO/indemnity plans in each market are also
needed for the analysis. Census Bureau data is used to find the number of non-
elderly uninsured; the difference between the total non-elderly population and the
sum of uninsured and insured by HMO/POS plans is assumed to be indemnity/PPO
coverage. One assumption is implicit in this methodology. Clearly, the publicly
insured, non-elderly (Medicaid) population should ideally be excluded explicitly from
these groups. When this was done using Census data, some markets had very low
or negative implied indemnity/PPO market shares. I therefore define the "plug"
between the total population and uninsured/HMO/POS as indemnity/PPO coverage,
assuming away the existence of the non-elderly publicly insured. The problem is
caused by error in the AIS enrollment data, which leads to error in the share of each
HMO/POS plan. I assume that the errors are randomly distributed across plans and
markets and therefore will not bias the results in any particular way.

Two properties of the enrollment data required attention. First, plan enrollment
in 17 of the 40 markets is reported at the Consolidated Metropolitan Statistical Area
(CMSA) rather than the PMSA level. Most of these\textsuperscript{65} are essentially one major city surrounded by very minor or inseparable suburbs: in these cases the analysis was performed as if the CMSA was a PMSA centered in the major city. However, for five CMSAs\textsuperscript{66} the population was sufficiently diffuse to warrant considering the individual PMSA/MSAs separately (for example, it seems unlikely that consumers in Miami would consider which hospitals are available in Fort Lauderdale when choosing a health plan). Enrollment in these five CMSAs was disaggregated to PMSA/MSA level using the populations of the counties where each plan was licensed; plans in the nine largest of these PMSA/MSAs were included in the plan choice model\textsuperscript{67}. The second issue is that, for 163 out of 516 plans (31.6%), enrollment data is reported at the state or (for three plans) national level rather than the MSA level. Two methodologies were used to handle these observations. First, the state or national level data was disaggregated to MSA level, again using the populations of the counties in which each plan was licensed, and the models estimated using all 516 observations. Second, the models were estimated with just the 353 plans for which data is provided; the missing HMO/POS plans were pulled into the "indemnity/PPO" option. The results obtained through the two different methodologies were very similar.

The final market share data were tested by comparing the implied total HMO share in each market to published data on market shares by state. The results were encouraging. For example, the predicted total HMO share in Boston is 57% of the non-elderly (and by assumption non-Medicaid) population; published data implies that HMOs have a 53% share of the non-elderly non-Medicaid population in Massa-

\textsuperscript{65}Examples are Cincinnati-Hamilton, Cleveland-Akron, and Sacramento-Yolo.

\textsuperscript{66}These are Dallas-Fort Worth, Los Angeles-Riverside-Orange County, Miami-Fort Lauderdale and San Francisco-Oakland-San Jose.

\textsuperscript{67}One market, New York, was included by AIS but excluded from my analysis because the market boundaries were difficult to define. 43 markets are therefore considered in total.
2.10 Appendix B: Estimating the Magnitude of Unobserved Hospital Quality

This appendix outlines the methodology used to estimate the variance of unobserved hospital quality $\eta$. Recall the regression defined by equation (8):

$$\hat{\delta}_h = x_h \alpha + \eta_h + \mu_h$$

where $\mu_h$ is the sampling error in $\delta_h$, defined by:

$$\hat{\delta}_h = \delta_h + \mu_h$$

Assume that $\mu$ and $\eta$ are independent, that $\eta$ has a Normal distribution with mean zero and $E(\eta \eta') = \sigma^2 I$, and that

$$E(\mu_i \mu_j) = \begin{cases} \omega_i^2 & \text{if } i = j \\ \omega_{i,j} & \text{otherwise} \end{cases}$$

The assumptions imply that the error variance for the regression analysis, defined as $u_h = \eta_h + \mu_h$, will not be homoscedastic. Following Hanushek (1974) and Lewis (2000), begin by running an OLS regression of $\hat{\delta}$ on $x$ and defining $\hat{\mu}$ as the residual vector. The expectation of the sum of squared residuals from this OLS regression can be written as:

$$E(\sum_h \hat{u}_h^2) = E(u'u) - tr((X'X)^{-1}X'\Omega X)$$

---

68 The source of the data is the website www.statehealthfacts.kff.org.
where $\Omega$ is the variance-covariance matrix of the vector of regression residuals $u$ and $tr$ is the trace operator\textsuperscript{69}. It can be shown that this implies:

$$
\sigma^2 = \frac{1}{H-k} \left[ E(\sum_h \hat{u}_h^2) - \sum_h \omega_h^2 + tr((X'X)^{-1}X'GX) \right]
$$

where $H$ is the number of observations in the sample, $k$ is the number of variables in $X$, and $G = \Omega - \sigma^2 I$ is the variance-covariance matrix of $\hat{\delta}$ in the first-stage hospital choice model. ($\sum_h \omega_h^2$ is the trace of this matrix.) An unbiased estimate of $\sigma^2$ can therefore be found using the following equation:

$$
\hat{\sigma}^2 = \frac{1}{H-k} \left[ E(\sum_h \hat{u}_h^2) - \sum_h \omega_h^2 + tr((X'X)^{-1}X'GX) \right]
$$

where $\sum_h \hat{u}_h^2$ is the sum of squared residuals from the OLS regression of $\hat{\delta}$ on $x$.

I follow this methodology and find that $\hat{\sigma}^2 < 0$. As noted by Lewis, this result is possible in small samples and can be assumed to imply that $\sigma^2 = 0$.

### 2.11 Appendix C: Adjusted Standard Errors

This appendix provides further details on the method used to adjust the standard errors for the three-step demand estimation process. Pakes (1997) proves consistency and asymptotic normality of two-step GMM estimators. The methodology used in this chapter is an extension to the three-step case. The necessary proofs are straightforward and are not given here; the following instead provides a derivation of the functional forms used in estimation. Consider the three steps of the estimation:

\textsuperscript{69}The trace of a square matrix is the sum of its diagonal elements.
1. In Step 1 the following equation is considered:

\[ u_{ihl} = \delta_h + x_h \nu_{ih} \beta + \varepsilon_{ihl} \]

MLE provides a consistent estimator of the true \( \beta \), which we call \( \beta_n \). The sample analogue of the moment condition is:

\[
F_n(\beta) = \frac{1}{N_1} \sum_{i,h,l} f_{ihl}(x, \beta) = \frac{1}{N_1} \sum_{i,h,l} I_{ihl} \frac{\partial \log s_{ihl}(x, \beta)}{\partial \beta}
\]

where \( s_{ihl} \) is the expression in parentheses in equation (7). Define:

\[ \Gamma_1 = \frac{\partial F}{\partial \beta} \]

and:

\[ L_1 = (\Gamma_1' \Gamma_1)^{-1} \Gamma_1 \]

where:

\[ F(\beta) = Ef(x, \beta) \]

2. In Step 2, a consistent estimate \( \alpha_n \) of the true \( \alpha \) in the following equation is found using OLS:

\[ \hat{\delta}_h = x_h \alpha + u_h \]

The sample analogue of the moment condition here is:

\[
M_n(\alpha) = \frac{1}{N_2} \sum_h m_h(x, \alpha) = \frac{1}{N_2} \sum_h x_h (\hat{\delta}_h - x' \alpha)
\]

Define:

\[ \Gamma_2 = \frac{\partial M}{\partial \alpha} \]
and:

\[ L_2 = (\Gamma_2^T \Gamma_2)^{-1} \Gamma_2 \]

where, analogous to Step 1:

\[ M(\alpha) = Em(x, \alpha) \]

3. Finally, consider Step 3. Considering the logit equation, and assuming a unit weight matrix for simplicity, we use two-stage least squares to find a consistent estimate, \( \hat{\vartheta}_n \), of the true parameter vector \( \vartheta \) in the following equation:

\[ y_{jm} = \log(s_{jm}) - \log(s_{om}) = z_{jm} \vartheta + \omega_{jm} \]

where:

\[ z_{1jm} = EU_{repjm} = h(z, x, v, \beta_n, \alpha_n) \]

The sample analogue of the moment condition here is:

\[ G_n(\vartheta, h(\beta_n, \alpha_n)) = \frac{1}{n} \sum_{j,m} g_{jm}(z, x, v, \alpha_n, \beta_n, \vartheta) = \frac{1}{n} \sum_{j,m} W_{jm}(y_{jm} - z_{jm} \vartheta) \]

where \( W_{jm} \) is a vector of instruments. Defining \( \Gamma_3 \) and \( L_3 \) analogously to the definitions in Steps 1 and 2 above, it can be shown that:

\[ avar(\sqrt{n}(\vartheta_n - \vartheta)) = (\Gamma_3' \Gamma_3)^{-1} \Gamma_3' V \Gamma_3 (\Gamma_3' \Gamma_3)^{-1} \]

where:

\[ V = avar \frac{1}{\sqrt{n}} \left[ \sum_{j,m} g_{jm}(\vartheta, h(\beta, \alpha)) + \frac{\partial g}{\partial h} \frac{\partial h}{\partial \beta} (\beta_n - \beta) + \frac{\partial g}{\partial h} \frac{\partial h}{\partial \alpha} (\alpha_n - \alpha) \right] \]

that is, the asymptotic variance of the first three terms of the Taylor series ex-
pansion for $\sqrt{n}G_n$. If we define:

$$M_1 = \frac{1}{n} \sum_{j,m} \frac{\partial g}{\partial h} \frac{\partial h}{\partial \beta}$$

and $M_2$ as the analogous expression for Step 2, we can write:

$$V = avar\left[\frac{1}{\sqrt{n}} \sum_{j,m} g_{jm} + \sqrt{n}M_1(\beta_n - \beta) + \sqrt{n}M_2(\alpha_n - \alpha)\right]$$

The fact that $\alpha_n$ and $\beta_n$ are GMM estimators implies that:

$$(\beta_n - \beta) = \frac{L_1}{N_1} \sum_{i,h,l} f_{ihi}(\beta) + o_p(1)$$

and similarly for $(\alpha_n - \alpha)$. This results in the following formula for $V$:

$$V = E(g'g') + \frac{n}{N_1} M_1 v_{c1} M_1' + \frac{n}{N_2} M_2 v_{c2} M_2' + \sqrt{\frac{n}{N_1}} E\left(\frac{g'f' L_1' M_1}{\sqrt{M_1}}\right)$$

$$+ \sqrt{\frac{n}{N_1}} E\left(M_1' L_1 f g'\right) + \sqrt{\frac{n}{N_2}} E\left(g m' L_2' M_2\right) + \sqrt{\frac{n}{N_2}} E\left(M_2' L_2 m g'\right)$$

$$+ \frac{n}{\sqrt{N_1 N_2}} E\left(M_1' L_1 m f' L_1' M_1\right) + \frac{n}{\sqrt{N_1 N_2}} E\left(M_2' L_2 m f' L_2' M_2\right)$$

where $v_{c1}$ and $v_{c2}$ are the variance-covariance matrices for the first and second steps respectively. The first three terms in this expression represent the impact of the variance-covariance matrices from each of the three steps on the final standard errors; those for the first two steps are increasing in the derivative of the third-step moment with respect to the first and second-step estimators respectively. The last six terms measure covariance effects, allowing for correlated errors across the three steps of estimation. The required standard errors are estimated by replacing expectations with sample averages in the above expression.
Chapter 3: Insurer-Provider Networks in the Medical Care Market

3.1 Introduction

This chapter continues my analysis of the restrictions imposed by each insurance plan on the network of hospitals from which its enrollees can choose. In Chapter 2 I estimate that a move from the existing hospital networks to an unselective outcome where all plans offer a free choice of hospitals would lead to a total gain in consumer surplus of $1.04 billion per year and a loss to producers of just $0.80 million per year assuming fixed premiums. In this chapter I demonstrate a second important consequence of the process used to determine insurers’ contracts with hospitals: an under-investment in capacity by hospitals that expect to fill their beds under a limited or selective network. I show that investment in new beds would generate a benefit to consumers of over $330,000 per bed per year assuming fixed premiums, far outweighing the effect on hospital and insurer profits. However, the bargaining process by which the prices paid to providers are agreed implies a negative incentive for capacity constrained hospitals to invest.
I base my analysis on the dataset, also used in Chapter 2, that defines the network of every managed care plan in 43 markets across the US. On average 10 per cent of insurer-hospital pairs in my data do not arrange contracts to provide care. I define selective markets as those in which at least four of the five major plans fail to reach agreement with at least one major hospital. By this definition roughly 20 per cent of observed markets are selective. 76% of enrollees in managed care plans do not have a free choice of hospitals. The data therefore show that, even in markets with reasonably small numbers of insurers and hospitals (12 and 15 per market on average, respectively), substantial numbers of potential contracts are missing. This contradicts a natural intuition that, if bargaining costs are small, we should observe agreement almost always. In addition to analyzing welfare effects, this chapter investigates the mechanism by which the observed selective equilibrium is determined.

I use demand estimates from the previous chapter to calculate the producer surplus (defined as plan revenues less hospital costs of care) generated by each potential hospital network for each plan in the data, taking into account patient flows across plans and hospitals. A simple analysis shows that around 50 per cent of contractual decisions are explained by this definition of producer surplus. A model of the price-setting negotiation between insurers and providers is needed to explain the rest of the data. I set out a simple motivating example to demonstrate how the negotiation can influence contractual decisions, even holding surplus fixed. I show that hospitals that do not need to contract with all plans in order to secure demand, such as those that expect to be capacity constrained even under a limited or selective network, can require plans to compete for contracts. This competition drives up hospital prices; as a result, some plans exclude these hospitals in equilibrium. In addition, if hospitals merge to form systems they can both demand higher prices than other providers and effectively force plans to contract with all or none of the hospitals in the system. Reduced form analysis is consistent with the model’s predictions.
The model also predicts welfare effects. I consider the example of capacity constrained hospitals in detail, and show that the predicted selective equilibrium can be inefficient even when the excluded hospital is full at equilibrium and the consumers with the highest value for it are the ones treated. The inefficiency is generated because consumers are forced to make sub-optimal choices across insurers in order to access their preferred hospital. The resulting loss of consumer welfare may outweigh the benefit derived when the highest-valuation patients are given preferential access to the hospital.

I estimate the profits secured by different types of hospital using insurers’ observed choices of hospital networks and data on the characteristics of both insurers and providers. The analysis is complicated by the existence of multiple potential equilibria and by possible problems with endogenous regressors. A number of existing papers have estimated models that address these issues\(^{70}\); however, their approaches are feasible only for problems involving small numbers of firms. This chapter adopts a different approach presented in Chapter 4 in which plans choose their networks in a simultaneous-moves game, conditional on their expectations regarding other plan choices and the prices demanded by hospitals. The equilibrium implicitly establishes markups for a hospital’s services that are functions of the characteristics of the hospital itself and the distribution of consumer, hospital, and plan characteristics in the particular market\(^{71}\).

I estimate the markups of three specific hospital types: those that expect to be capacity constrained, "star" hospitals (providers that are very attractive to consumers), and those that are members of hospital systems. I find that, as predicted by the

---


\(^{71}\) The identifying assumption used for estimation is the simple necessary condition that the profit each plan expects to earn from its chosen network is greater than the profit it could expect to earn from alternative networks. This assumption is sufficient to define bounds on the feasible values of the parameters to be estimated.
model, hospitals that can credibly threaten to turn down low price offers capture a high share of the surplus. Hospitals with exogenous characteristics predicting capacity constraints capture $1800 per patient treated more than other providers. This implies markups of approximately 15 per cent of revenues, in contrast to non-system, non-capacity constrained providers, whose markups are about two per cent. The profits of hospitals in systems are approximately $179,000 per month higher than other providers; this translates to a markup of about 14 per cent of revenues. They also charge high penalties from plans that contract with some but not all of the hospitals in their system. The results are therefore consistent with several recent papers that suggest that hospital mergers may prevent plans from using the threat of exclusions to control prices\textsuperscript{72}. Star hospitals may also capture high profits but these are imprecisely estimated. In addition, I find that hospitals with low costs have higher markups than their competitors, consistent with simple bargaining models.

Finally, I outline the implications of the results. The methodology used in this chapter does not fully detail how the multiplicity of possible equilibria is resolved. Instead my analysis provides a reduced form characterization of the markup equation that is generated by the equilibrium that does materialize. More precise details of the bargaining game would be needed to determine how that equilibrium is chosen, and this would be required before I could provide a detailed analysis of counterfactuals. I do, however, analyze the relationship between hospital characteristics and markups. This provides helpful information for assessing which bargaining models best describe the hospital-insurer price negotiation. To the extent that policy and environmental changes do not affect the reduced form relationship, the results can also be used to predict how changes in market characteristics are likely to affect hospital markups.

The welfare effects implied by the estimates are considered in this spirit. In

\textsuperscript{72}See, for example, Lesser and Grinsburg (2001), Mays, Hurley and Grossman (2003), and Capps and Dranove (2004).
particular I predict the impact of hospital investment to remove capacity constraints, assuming no effect of this change on plan networks or other market characteristics\textsuperscript{73}. The investment would result in a median benefit to consumers of $0.20 per person per new bed per year, or $338,800 per year per additional bed for the market as a whole. Insurer profits would increase as a result of the change. However, I find that the hospital markups generated by capacity constraints outweigh the additional revenues from new patients and imply that, at least in the short term, these providers have no incentive to invest to remove their constraints.

Several strands of the health economics literature are relevant to this chapter. A number of authors demonstrate that the prices paid by plans to hospitals are consistent with simple bargaining models\textsuperscript{74}. Gal-Or (1997, 1998) develops a Nash bargaining model in a two-plan, two-hospital setting. Vistnes (2000) has a model of two-stage competition between hospitals: providers compete first for preferential access to health plans and then for individual patients. Finally, Eggleston, Norman and Pepall (2004) use a similar theoretical framework in a market containing health plans, hospitals and physician groups to look at the effects of horizontal and vertical integration on prices. However, previous authors have not addressed the network allocation question empirically: that is the main contribution of this chapter.

In the next two sections I describe the contractual process between insurers and hospitals and introduce the dataset. Section 3.4 briefly outlines the demand estimates from the previous chapter and uses them to generate a measure of surplus. Section

\textsuperscript{73}I assume that hospital capacity is determined prior to the contracting process and that the relationship identified between capacity constraints and hospital markups is causal. See Section 3.2.2 for a discussion of the rationale for this assumption.

\textsuperscript{74}Most of these regress the prices paid to hospitals on measures of hospital and plan bargaining power. Examples are Brooks, Dor and Wong (1996), Zwanziger and Mooney (2000) and Feldman and Wholey (2001). In addition, Town and Vistnes (2001) and Capps, Dranove and Satterthwaite (2004) both investigate the effect of the hospital’s value to consumers on its profits. They estimate consumer preferences over hospitals and regress hospital profits or prices on variables that summarize consumer demand for the hospital.
3.2 Background on Contracting in Medical Care

3.2.1 The Contractual Process

Every HMO/POS plan contracts separately with every hospital in its network. The exact form of the contracts varies, but all specify a price to be paid to the provider per unit of care (for example a price per inpatient day or per diagnosis-related group (DRG)). Prices vary both across providers for a given insurer and across insurers for a given provider; contracts are usually renegotiated annually. Both parties in the negotiation need to balance consumer demand for services against the price agreed. A health plan would prefer to contract with the hospitals that are valued by its likely customers, particularly the customers on the margin of joining, but must also take into account the fact that hospitals in demand may seek higher prices than their less differentiated counterparts. Hospitals seek to maximize their returns by contracting with plans that both offer high prices and provide a steady flow of patients.

Interviews with plan and hospital representatives who are involved in contractual negotiations confirm that the form of the price negotiation, and therefore the effect of contracting decisions on final prices, is affected by the relative bargaining positions of hospitals and health plans in the market. As the Director of Operations Analysis in one hospital chain put it:

"There are counteracting effects here: the outcome [of any plan decision, like excluding a particular hospital] depends on where the balance of power lies."
In many markets, where managed care is strong and hospitals compete for contracts, the plan may begin negotiations by stating the maximum per cent increase in prices it can offer from the previous year and hospitals may generally accept that price increase. The Executive Director of another hospital system described one potential outcome in such markets:

"There are examples where there were too many hospitals in an area, and the plans played them off against each other to the point where the price paid was no more than marginal cost."

This idea is consistent with the theory that, when plans have better outside options than hospitals, they are able to exclude those that demand high prices. The negotiation can be very different in markets where a few hospitals have very strong reputations and high market shares: the hospital may be able to demand a price which the plan must pay in order to avoid losing the contract\(^\text{75}\). A hospital Director said the following:

"In market X [where hospitals are very strong], the prices [the best hospitals] charge are based on their very high patient satisfaction results and their strong reputation. They can get high prices from any plan in the market, and they don’t need them all."

The CEO of a small hospital in a different market had a similar story:

"Large [hospitals] in this market can dictate whatever prices they want. The bigger names can demand the highest prices."

The implication is that, in markets where hospitals expect to be full or can credibly threaten to turn down low price offers for some other reason, they demand a large share of the surplus they generate; this may prompt plans to exclude them.

\(^{75}\)Prices paid to hospitals were regulated at the state level in the 1960s and 1970s. However, since Medicare and Medicaid switched from cost-based to prospective payment systems, and managed care encouraged increased price competition between hospitals, rate regulation has virtually disappeared.
3.2.2 The Timing of Firm Decisions

In order to model the contractual process I need to specify the timing of the different hospital and plan decisions and make a number of other assumptions. These are summarized in the following section.

The stages of my model are as follows:

Stage 1: Plans and hospitals agree on contracts
Stage 2: Plans set premiums
Stage 3: Consumers and employers jointly choose plans
Stage 4: Sick consumers visit hospitals; plans pay hospitals per service provided

My main focus is on Stage 1. I assume fixed premiums throughout most of my calculations; I include a robustness test to consider the effects of potential premium adjustments in Section 3.8. I analyze Stages 3 and 4 in Chapter 2: my methodology is summarized briefly in Section 3.4.1 for ease of reference.

A few additional comments are in order. First, many insurers offer several types of product: for example Aetna health plan may offer both an HMO and a POS product in a given market. I assume that the choice of products, together with the hospital’s choices of capacity, location, services, and quality, are made prior to Stage 1. My analysis conditions on these decisions. I therefore do not explicitly model issues such as product-based price discrimination (the plan’s choice of products can be seen as a way of dividing the market into segments with different price elasticities of demand) and the hospital’s decision regarding investment in new capacity given that offered by its competitors. Similarly, I assume that hospital merger decisions are made prior to the contractual process. Second, I focus on inpatient care. According to the

\[\text{76}\text{These assumptions seem reasonable because the relevant variables change more slowly over time than hospital-insurer contracts. For example, over 90\% of hospitals did not alter their offerings of angioplasty, ultrasound, open heart surgery or neonatal intensive care units over the four-year period 1997-2001; 70 percent of hospitals changed their capacity levels by fewer than 20 beds over the same four-year period. The correlation between market-level bed capacity (beds per thousand population) in 1980 and that in 2001 is 0.63. Plan product offerings and hospital locations are similarly static.}\]
American Hospital Association, 65% of hospital revenues in 2001 were derived from inpatient care; the remainder came from outpatient services.

As noted in Section 2.2.2, my dataset contains no exclusive contracts and few vertically integrated organizations apart from Kaiser Permanente. Since vertical integration seems to be disappearing I do not attempt to explain the phenomenon in this chapter. I condition on the existence of Kaiser health plans and hospitals in my analysis of both the supply and demand sides of the market (since they are important members of the plan and hospital choice sets, particularly in California) but exclude them from my models of firm behavior.

The health plan must take state and federal legislation into account when choosing its providers. Many states have implemented Any Willing Provider laws which prohibit health insurers from excluding qualified health care providers that are willing to accept the plans’ terms and conditions. However, these regulations have been argued to remove the benefits of managed care, since they prevent plans from trading volume for lower provider prices. Perhaps for this reason they apply to hospitals in only seven states (in other areas they are largely limited to pharmacies). I have data covering two markets within these states; I find that plans are just as likely to exclude hospitals in these markets as elsewhere. I therefore assume that these regulations have no impact on plan decisions in the markets I consider. In addition, some states have implemented Essential Community Provider laws, which require insurers to contract with providers that offer "essential community services", such as public hospitals and teaching hospitals, and to contract with enough hospitals to serve the needs of the local population. I assume these regulations do not affect the decision of a plan to exclude any particular hospital since consumer demand forecasts would prevent it from dropping too many hospitals in any case.

---

Hospital-insurer contracts, in contrast, are usually renegotiated annually. My goal is to estimate the short-term effects of these hospital and plan characteristics on equilibrium contracts.
3.3 A Preview of the Data

The primary dataset analyzed in this chapter defines, for every HMO/POS plan in 43 major US markets, the network of hospitals offered to enrollees in March/April 2003. The dataset is also used in Chapter 2\textsuperscript{77}. It includes 516 HMO/POS plans and 665 hospitals in total\textsuperscript{78}. Figure 1 in Chapter 2 documents the observed variation across both markets and plans in the extent to which plans exclude major hospitals from their networks.

Table 11 compares the means of a number of market characteristics in selective and unselective markets. There are few significant differences. Selective markets do not have significantly smaller populations, higher managed care penetration, more hospitals, or more beds per capita than unselective markets and are not clustered geographically. There are no significant demographic differences. The only difference that is significant at $p=0.05$ (or in fact at $p=0.2$) is the standard deviation of the distances between hospitals in the market. Plans seem to be more willing to exclude hospitals in areas where hospitals are clustered into several groups, perhaps because each provider in a given group is a reasonable substitute for the others. The raw data therefore do not offer an obvious explanation for the observed variation; however, they do provide a hint that demand effects may be important. These are taken into account in the analysis described in Section 3.4.

\textsuperscript{77}See Sections 2.2.2 and 2.4.2 for details.

\textsuperscript{78}The markets are: Atlanta GA, Austin TX, Baltimore MD, Boston MA, Buffalo NY, Charlotte NC, Chicago IL, Cincinnati OH, Cleveland OH, Columbus OH, Dallas TX, Denver CO, Detroit MI, Fort Worth TX, Houston TX, Indianapolis IN, Jacksonville FL, Kansas City MO, Las Vegas NV, Los Angeles CA, Miami FL, Milwaukee WI, Minneapolis MN, New Orleans LA, Norfolk VA, Oakland CA, Orange County CA, Orlando FL, Philadelphia PA, Phoenix AZ, Pittsburgh PA, Portland OR, Sacramento CA, St. Louis MO, Salt Lake City UT, San Antonio TX, San Diego CA, San Francisco CA, San Jose CA, Seattle WA, Tampa FL, Washington DC, and West Palm Beach FL.
Table 11: Summary Data for Selective and Unselective Markets

<table>
<thead>
<tr>
<th></th>
<th>Unselective Markets (Category 1 and 2) Mean (std devn)</th>
<th>Selective Markets (Category 4 and 5) Mean (std devn)</th>
<th>p-value for difference in means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market population (million)</td>
<td>2.34 (1.08)</td>
<td>2.36 (1.96)</td>
<td>0.95</td>
</tr>
<tr>
<td>Number of HMO/POS plans with over 1% market share</td>
<td>6.75 (1.65)</td>
<td>6.57 (1.89)</td>
<td>0.76</td>
</tr>
<tr>
<td>Number of hospitals</td>
<td>19.88 (10.99)</td>
<td>21.24 (20.53)</td>
<td>0.79</td>
</tr>
<tr>
<td>Beds per 1000 population</td>
<td>2.82 (0.96)</td>
<td>2.94 (0.96)</td>
<td>0.79</td>
</tr>
<tr>
<td>Managed care penetration</td>
<td>0.33 (0.16)</td>
<td>0.35 (0.15)</td>
<td>0.61</td>
</tr>
<tr>
<td>Average age of population</td>
<td>34.97 (2.28)</td>
<td>34.31 (1.39)</td>
<td>0.32</td>
</tr>
<tr>
<td>% of under-65 population aged 55-64</td>
<td>0.09 (0.01)</td>
<td>0.09 (0.01)</td>
<td>0.73</td>
</tr>
<tr>
<td>Median total family income of population</td>
<td>$49,160 (8,244)</td>
<td>$46,130 (8,642)</td>
<td>0.29</td>
</tr>
<tr>
<td>Std devn of total family income of population</td>
<td>$54,197 (8,690)</td>
<td>$52,797 (6,511)</td>
<td>0.62</td>
</tr>
<tr>
<td>Mean distance between hospitals (miles)</td>
<td>11.62 (5.37)</td>
<td>12.54 (5.22)</td>
<td>0.31</td>
</tr>
<tr>
<td>Std devn of distances between hospitals (miles)</td>
<td>7.74** (3.26)</td>
<td>10.30** (4.06)</td>
<td>0.04</td>
</tr>
<tr>
<td>No. hospitals with open heart surgery</td>
<td>7.88 (3.63)</td>
<td>10.19 (8.59)</td>
<td>0.27</td>
</tr>
<tr>
<td>N</td>
<td>16</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>
A number of other datasets are introduced at various stages of the analysis. Hospital characteristic data is taken from the American Hospital Association (AHA) dataset for 2001. Plan characteristics come from two datasets from Atlantic Information Services\textsuperscript{79} for Quarters 3 and 4 of 2002, supplemented with information from the \textit{Weiss Ratings’ Guide to HMOs and Health Insurers} for Fall 2002. Data on plan performance comes from the \textit{Health Employer Data and Information Set} (HEDIS) and the \textit{Consumer Assessment of Health Plans} (CAHPS) 2000, both of which are published by the National Committee for Quality Assurance (NCQA). These data measure clinical performance and patient satisfaction in 1999. All these datasets are used in Chapter 2 of my dissertation: further details on the data, and the methodology used to create additional variables (such as plan market shares) that are employed again in this analysis, are given there. My demand estimation includes all 665 hospitals and all 516 managed care plans in the data. When I consider the supply side I restrict attention to non-Kaiser plans for which premiums are observed; I also exclude a few extremely selective insurers that I regard as outliers\textsuperscript{80}. The remaining data contain 451 plans in total\textsuperscript{81}. I model these plans’ contracts with the six largest hospitals in each market: these cover an average of 57 per cent of the total admissions to non-Federal general medical and surgical hospitals in the markets I consider.

\textsuperscript{79}These are \textit{The HMO Enrollment Report} and \textit{HMO Directory 2002}. Both are based on plan state insurance filings.

\textsuperscript{80}I exclude plans that drop more than four of the top six hospitals because these may have different reasons for their contracting decisions than other plans in the data. I also exclude two specific outliers: Scott and White Health Plan of Austin, TX and Group Health Cooperative of Puget Sound. These are different from most other plans in the market in that they are locally-based, consumer-driven insurers that are heavily focused on primary care.

\textsuperscript{81}The supply side analysis includes a prediction of plan market shares given the networks offered by every plan in the market. I condition on the existence of "excluded" plans (Kaiser plans, those that are very selective and those for which premiums are unobserved) when calculating the shares of the plans that are modelled explicitly. I also take account of indemnity and PPO plans, making assumptions about their characteristics, and allow consumers to choose the outside option of being uninsured, as described in detail in Chapter 2.
hospitals in the market.

Descriptive statistics for the plans and hospitals in the data are given in Tables 3 and 4 respectively, both in Section 2.4.3. The hospitals have 339 beds and 1.26 registered nurses per bed on average; 20% are teaching hospitals. The average market share of the HMO/POS plans in the dataset is 3%\(^\text{82}\). Premiums average $141 per member per month. 35% of insurers are POS plans; 76% have been in existence for over 10 years. HEDIS scores vary widely, from an average rating of 0.15 (for the percent of children receiving all required doses of MMR, Hepatitis B and VZV vaccines before their 13th birthday) to an average of 0.73 (the proportion of women aged 52-69 who had received a mammogram within the previous two years). The two most frequently-occurring plans are Aetna and CIGNA, with 15% and 10% of observations respectively.

Table 10 in Section 2.9.1 sets out details of the definitions of four variables that summarize the services offered by each hospital. The summary variables cover cardiac services, imaging, cancer and birth services: each hospital is rated on a scale from 0 to 1, where 1 implies that the hospital offers the least common of a list of relevant services and 0 implies that it offers none of the services. I interact these variables with consumer characteristics in the model of demand for hospitals. They can also be used to investigate which hospital characteristics are correlated with market share. Table 12 sets out the results of a regression of hospital market shares on hospital characteristics. All four service variables, and the indicator for teaching hospitals, are positively and significantly related to market share. Together with hospital location, they will be key determinants of hospital capacity constraints (and therefore market power and the ability to negotiate positive profit margins) later in the analysis.

\(^{82}\text{Shares are measured as percent of the nonelderly population in the market.}\)
Table 12: Regression of Hospital Market Shares on Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Coefficient estimate</th>
<th>Coefficient estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiac services</td>
<td>0.027** (0.006)</td>
<td>0.022** (0.005)</td>
</tr>
<tr>
<td>Imaging services</td>
<td>0.019** (0.007)</td>
<td>0.017** (0.006)</td>
</tr>
<tr>
<td>Cancer services</td>
<td>0.008** (0.005)</td>
<td>0.017** (0.004)</td>
</tr>
<tr>
<td>Birth services</td>
<td>0.017** (0.005)</td>
<td>0.009** (0.004)</td>
</tr>
<tr>
<td>Teaching hospital</td>
<td>0.019** (0.005)</td>
<td>0.030** (0.004)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.002 (0.006)</td>
<td>-0.011 (0.009)</td>
</tr>
<tr>
<td>Market FEs</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.16</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Notes: There are 665 hospitals in both specifications. Standard errors are reported in parentheses; **significant at p=0.05; *significant at p=0.1. Cardiac, imaging, cancer and birth services refer to the four hospital service variables defined in Section 2.9.1.

3.4 Effect of the Network on Producer Surplus

3.4.1 Demand Estimates

In order to understand the equilibrium network outcomes I need to analyze Stages 3 and 4 of the model, in which consumers choose their health plans taking into account the hospitals they expect to visit in the coming year. The parameter estimates generated in Chapter 2 are used as an input to this chapter’s supply side analysis. The details of the demand estimation process are set out in Chapter 2; I provide a brief overview here for ease of reference. The demand estimation process has three stages:
1. The first step is to estimate demand for hospitals using a discrete choice model that allows for observed differences across individuals. With some probability consumer $i$ (whose type is defined by age, gender, and zipcode tabulation area (ZCTA)) becomes ill. His utility from visiting hospital $h$ given diagnosis $l$ is given by:

$$u_{ihl} = \eta_h + x_h \alpha + x_h v_{il} \beta + \varepsilon_{ihl}$$

where $x_h$, $\eta_h$ are vectors of observed and unobserved hospital characteristics respectively, $v_{il}$ are observed characteristics of the consumer such as diagnosis and location and $\varepsilon_{ihl}$ is an idiosyncratic error term assumed to be iid Type 1 extreme value. Hospital characteristics include location, the number of beds, the numbers of nurses and doctors per bed, and details of services offered, ownership, and accreditation. This equation is estimated using standard maximum likelihood techniques and micro (encounter-level) data from the MEDSTAT MarketScan Research Database for 1997-98. The data provide information on the hospital admissions of indemnity plan and PPO enrollees.

2. Secondly, I use the estimated coefficients to predict the utility provided by each plan’s hospital network. Individual $i$’s expected utility from the hospital network offered by plan $j$ in market $m$ is calculated as:

$$EU_{ijm} = \sum_l p_{il} \log \left( \sum_{h \in H_j} \exp(\eta_h + x_h \hat{\alpha} + x_h v_{il} \hat{\beta}) \right)$$

where $p_{il}$ is the probability that individual $i$ will be hospitalized with diagnosis.

---

83 It would be preferable to estimate consumers’ hospital choices using data for managed care enrollees. However, this is not feasible because the available data do not identify the hospital networks offered by each managed care plan, so the choice sets of managed care enrollees are unobserved. Instead I consider the choices made by indemnity and PPO enrollees, whose choice set is unrestricted. I assume that indemnity/PPO enrollees have the same preferences over hospitals as HMO/POS enrollees, conditional on their diagnosis, income and location. I test this assumption using data for HMO/POS enrollees in Boston; see Chapter 2 for details.
3. Finally, I use aggregate data from Atlantic Information Services, the NCQA and the AHA to estimate the health plan demand model. I use a methodology similar to that first proposed by Berry, Levinsohn and Pakes (1995). The utility of individual $i$ from plan $j$ in market $m$ is given by:

$$\tilde{u}_{ijm} = \xi_{jm} + z_{jm} \lambda + \gamma_1 EU_{ijm} + \gamma_2 \frac{\text{prem}_{jm}}{y_i} + \omega_{ijm}$$

where $z_{jm}$ and $\xi_{jm}$ are observed and unobserved plan characteristics respectively, $\text{prem}_{jm}$ are plan premiums, $y_i$ is the income of individual $i$, and $\omega_{ijm}$ represents idiosyncratic shocks to consumer tastes, again assumed to be iid Type 1 extreme value. I consider HMO and POS plans only.$^{84}$ The characteristics included in $z$ are premium, the size of the physician network, plan age, a list of eight clinical quality variables (taken from the NCQA’s HEDIS dataset), and two variables summarizing consumer assessment of plans on dimensions such as availability of needed care and speed with which care is received (from their CAHPS dataset). The results of this stage of the analysis are set out in Table 8 of Section 2.6.2. I find that consumers place a positive and significant weight on their expected utility from the hospital network when choosing a plan. The coefficient magnitudes imply that a one standard deviation increase in expected utility is equivalent to a reduction in premium of $39 per member per month (a little less than one standard deviation).

$^{84}$I include two additional potential choices for each consumer: an indemnity/PPO plan option, defined using assumptions about the characteristics of these insurers in each market, and the outside option of being uninsured.
3.4.2 Producer Surplus Generated by the Network

With the demand estimates in hand, I now move on to consider the observed health plan-hospital contracts. The simplest model of contracting assumes that each insurer-provider pair bargains independently over the division of a surplus of size $M$. The implication (whatever the bargaining framework used) is that firms reach agreement if and only if $M > 0$. I investigate this theory by using my demand estimates to predict the producer surplus generated by each plan when it contracts with each potential hospital network: that is, the total profit to be divided between the plan and all the hospitals with which it contracts. The producer surplus generated by plan $j$ in market $m$ when it contracts with hospital network $H_j$ is:

$$S_{jm}(H_j, H_{-j}) = \sum_i \left( n_is_{ijm}(H_j, H_{-j}) \left[ prem_{jm} - p_i \sum_{h \in H_j} s_{ih}(H_j) \text{cost}_h \right] \right)$$ (16)

where $n_i$ is the population in consumer-type cell $i$ (defined by ZCTA, age, and gender), $p_i$ is the probability that a type-$i$ person will be admitted to hospital, $\text{cost}_h$ is the average cost of treatment at hospital $h$, and $\text{prem}_{jm}$ is plan $j$’s premium in market $m$. The quantities $s_{ijm}(H_j, H_{-j})$ and $s_{ih}(H_j)$ are plan $j$’s and hospital $h$’s predicted shares of type-$i$ people when networks $H_j$ and $H_{-j}$ are offered by plan $j$ and other plans respectively. These are predicted using the demand estimates and take account of the flow of consumers across plans, and across hospitals given their choice of plans, in response to network changes.

The surplus definition does not include plans’ non-hospital variable costs. Each plan faces a number of costs of enrolling consumers: these include payments to primary care physicians and prescription drug costs, for example, in addition to the costs of treatment at hospitals. If these non-hospital variable costs differ across consumer types they may affect the plan’s network choice since adding a new hospital could disproportionately attract certain types of consumer. Unfortunately, I do not have
access to data on plan variable costs and therefore cannot include these effects in the surplus term\textsuperscript{85}. However, I account for this issue later in the analysis by estimating the cost of enrolling certain types of consumers directly. The details of this robustness test are discussed in Section 3.7.

The calculation takes account of hospital capacity constraints. If any network combination implies that any hospital is over 85 per cent of its maximum capacity level, I reallocate patients randomly to non-capacity constrained hospitals in the market. I assume that patients are treated in the order in which they arrive and that the timing of sickness is random: each plan therefore has the same percentage of enrollees reallocated for any given capacity constrained hospital. The adjustment affects patients’ hospital choices and therefore their predicted costs of care but does not impact consumers’ choices of plan or premium levels\textsuperscript{86}.

Premiums are assumed fixed in this calculation. In reality, when plan $j$ considers a deviation from its observed network, it probably predicts that its own premium and those of other plans will adjust in response to the network change. Here again I encounter data limitations. I cannot estimate these adjustments accurately since I do not have access to panel data and so cannot observe the reaction of plan premiums to network changes over time. However, I include a robustness test for the fixed premium assumption; this is discussed in Section 3.8.

\textsuperscript{85}The analysis does allow for the existence of additional fixed costs, since these would cancel out when we consider the surplus change from a change in networks.

\textsuperscript{86}A hospital is predicted to be over 85\% of maximum capacity if predicted admissions $\times$ average length of stay at the hospital is greater than 85\% of the number of beds $\times$ 365 days. By using the surplus variable without adjusting consumers’ choices of plan, I am assuming that the plan does not expect consumers to predict their probability of treatment at each hospital in its network when choosing their insurer. Instead consumers are expected to assume they will have access to every hospital on the list. Consumers may update this belief if a hospital is consistently capacity constrained (although many of the non-Medicare, non-Medicaid enrollees considered in this paper will have little experience of seeking hospital treatment on which to base their updates). Unfortunately, without a panel dataset, there is no variation in the data to identify the extent of any such updating.
3.4.3 Does the Producer Surplus Term Explain the Observed Contracts?

The next step is to use the producer surplus estimates to identify the surplus generated by each observed contract decision. I repeat this calculation for each of the 2706 potential contracts between the 451 plans in the data and the six largest hospitals in each market, keeping all other plans’ networks fixed. The results are summarized in Table 13.

<table>
<thead>
<tr>
<th>Number of contracts</th>
<th>( \Delta Surplus &gt; 0 )</th>
<th>( \Delta Surplus = 0 )</th>
<th>( \Delta Surplus &lt; 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contract observed</td>
<td>2306</td>
<td>53.8%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Contract not observed</td>
<td>400</td>
<td>44.2%</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

I find that the estimated producer surplus generated by the contract (which I denote \( \Delta Surplus \)) is greater than zero for 54 per cent of the 2306 agreed contracts. The surplus that would be created by the contract is less than zero for 52 per cent of the 400 potential contracts that were not agreed upon. So the simplest hypothesis explains the data in just over 50 per cent of cases. One way to interpret the fit of this simple model is to calculate a pseudo-\( R^2 \) measure. If we place equal weight on correctly predicting the set of observed contracts and the set that are unobserved, the pseudo-\( R^2 \) is 0.46\(^{87}\).

It is worth noting here that my definition of producer surplus measures the effect of the contract on the profits to be divided between the plan and all the hospitals.

\(^{87}\)The pseudo-\( R^2 \) is defined as \( 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \), where \( y_i \) is the observed outcome, \( \hat{y}_i \) is its predicted probability, and \( \bar{y}_i \) is the mean value in the data. This is the same measure used to assess goodness of fit in Stata’s logit and probit calculations.
in its network. That is, I take account of the effect of a particular contract on the
plan’s profits from other hospitals with which it already has contracts. I do not,
however, account for the other relevant externality: the fact that if a hospital agrees
on a contract with one plan, this will affect consumer flows and therefore its revenues
and profits from other plans in its market. This interaction between the negotiations
of particular plans may well explain why my producer surplus measure imperfectly
predicts the data. The simple theoretical model in the next section considers this
issue in more detail\textsuperscript{88}.

3.5 The Price Negotiation

The producer surplus results provide us with a puzzle. We have to explain the 46
per cent of contracts agreed when the predicted surplus increase is negative and the
48 per cent not agreed when $\Delta Surplus$ is positive\textsuperscript{89}. I therefore move on to consider
the price negotiations which determine how the producer surplus is divided between
insurers and hospitals. The following simple motivating example takes account of the
externalities faced by hospitals: the fact that a hospital’s contract with one insurer
affects its profits from the others in its market. The model provides details on Stage
1 of the full game set out in Section 3.2.2. It demonstrates that, for certain types of
provider, the negotiation can lead to selective contracting even when each additional
contract would create positive incremental surplus to the plan, and to agreement
when the incremental surplus is negative. In this section I outline the intuition
and predictions of the model; details of the solution for a simple case are given in

\textsuperscript{88} The fact that my producer surplus measure does not fully explain the data does not necessarily
imply that the outcome does not maximize total producer surplus, nor that it is inefficient. However,
the model in Section 3.5 shows that an inefficient outcome is possible under certain parameter values.

\textsuperscript{89} Two assumptions made in the surplus calculation may help rationalize the contracts: the as-
sumptions of fixed premiums and zero plan non-hospital variable costs. Both assumptions will be
addressed in robustness tests of the full empirical model.
the Appendix. I make some fairly restrictive assumptions about the details of the bargaining game in order to solve for the equilibrium. However, the main intuition holds for more general models: I relax many of these assumptions when I return to the main empirical estimation.

Consider a market in which hospitals offer services to health plans. Insured consumers receive two types of service from their plan: acute care from the hospitals in the network and preventive services from the plan’s primary care physicians (PCPs). I assume the following order of actions:

1. Each plan announces its preferred hospital network
2. Plans agree on prices with their chosen hospitals
3. Plans set premiums; consumers decide whether to enroll and visit hospitals

Each plan-hospital agreement specifies the price to be paid to the hospital per treatment: the hospital is then required to treat every enrollee from the plan who requests care, provided it has spare capacity. I assume that plans make simultaneous, private take it or leave it offers to hospitals. All firms have complete information about other firms’ characteristics. Each plan offers a single product and sets a single premium level. Every consumer expects to need treatment once in the following year.

Each plan predicts the surplus created by every potential network, conditioning on its beliefs regarding other plan actions. Each also calculates the profits that would be earned by every hospital at equilibrium; these would reduce the share of the surplus captured by the plan. Plans therefore choose their networks to maximize:

---

90 This assumption simplifies the proofs for the model. A Nash bargaining model would not have qualitatively different results. In reality plans and hospitals probably bargain over contracts, implying that both insurer and provider capture a positive share of the surplus in each negotiation. The estimation procedure set out in Section 3.7 allows for flexibility regarding the type of bargaining model used.

91 Each plan is assumed to have perfect information regarding hospital characteristics. It can therefore perfectly predict which hospitals will accept its offers.
\[
\pi_{jm}(H_j, H_{-j}) = S_{jm}(H_j, H_{-j}) - \text{hospital profits}
\]  

where \( \pi_{jm}(H_j, H_{-j}) \) is the profit secured by plan \( j \) when it contracts with hospital network \( H_j \), given the networks of other plans \( H_{-j} \), and \( S_{jm}(H_j, H_{-j}) \) is the producer surplus generated by the network\(^{92}\).

In the simplest case consumers are unwilling to switch plans in order to gain access to hospital \( h \). The hospital therefore needs to contract with every plan to maximize its expected number of patients: if it is not expected to be full it will accept any price offer that covers its costs\(^{93}\). Plans capture 100 per cent of the surplus created and will include hospital \( h \) in their networks provided it generates positive producer surplus. The more interesting situation arises when consumers are willing to switch plans if necessary to ensure access to \( h \). In this case the hospital may credibly threaten to turn down low price offers for three reasons:

1. If all or most consumers with a positive value for \( h \) would switch plans to access it, turning down a low price offer would not significantly reduce demand for the hospital’s services. In addition, by turning down the low offer, the hospital can induce those enrollees who are willing to follow it across plans to move to an insurer that offers a higher price for their treatment. This increases the provider’s revenues from these "switchers". The hospital will optimally turn down the lower offer if the revenues lost from enrollees it cannot then access are outweighed by those gained from consumers who are willing to move across plans. This is most likely for hospitals that are very attractive to consumers: I describe them as "star" hospitals. I discuss in Section 3.7 the characteristics

\(^{92}\)I ignore plans’ non-hospital costs here, but account for them in the empirical estimation in Section 3.7.

\(^{93}\)Small hospitals, which would be full if they contracted with a single plan, are included in the category of capacity constrained providers discussed in point 2 below.
used to define these providers in the empirical estimation.

2. The second scenario is that a subset of consumers will switch plans to access $h$ and that the hospital expects to be capacity constrained: that is, it can fill its beds without treating all consumers who wish to access it. In this case the loss of patients from the low-priced plan has a smaller (perhaps even zero) impact on the hospital’s total volume, making it more likely to refuse the lower offer. A second effect also arises here. I assume that the timing of sickness is random and that hospitals with spare capacity cannot turn away patients. This implies that, if the capacity constrained hospital accepts all price offers, some enrollees from the lower-priced plan will displace higher-price patients from other insurers. Hospital $h$ therefore has an additional reason to accept only the highest price offers.$^{94}$

3. Finally, if a sufficiently large proportion of the hospitals in the market merge to form a single system, the combined organization may be very attractive to consumers. As in case 1., the proportion of consumers willing to switch plans to access the system may be very high, implying that each system member will optimally turn down low price offers.

In all three cases, insurers will be forced to compete for contracts with the provider: hospital $h$ will therefore capture a positive share of the surplus. A given plan may choose to exclude the hospital, focusing instead on those consumers whose low valuation for $h$ and higher valuation for its other services prevents them from switching, if other plans have a higher maximum willingness-to-pay (WTP) for the contract.$^{95}$

$^{94}$The feature that distinguishes capacity constrained hospitals from star providers is that the former might optimally accept the lower price offer if they had sufficient beds to treat all "non-switching" consumers who wished to access them. The latter have enough consumers willing to follow them across plans that they turn down low offers even when their beds are not full.

$^{95}$A given plan may have a lower WTP for the contract than other insurers for two reasons. First, it may have a better outside option than other plans due to variation in consumers’ preferences for
The model predicts, therefore, that selective equilibria may be observed, even when positive incremental producer surplus would be generated from additional contracts, when hospitals are extremely attractive; capacity constrained; or have merged to form systems\textsuperscript{96}.

The existence of systems can also explain the contracts that are agreed despite a negative incremental surplus. I observe in the data that some plans contract with some but not all members of a hospital system, but this practice is infrequent. I rationalize this observation with the idea that, if a hospital system has significant market power (as in point 3.), it will optimally impose penalties on plans that contract with some but not all of its members\textsuperscript{97}. Even systems with little market power may choose a bundling strategy, charging relatively more for contracts with individual hospitals than for those with the entire organization, to maximize the surplus captured from each plan. In both cases plans may be deterred from cherry-picking from the members of a system\textsuperscript{98}.

The model with capacity constraints is considered formally for the simple two-plan, other plan characteristics. Second, its non-switching enrollees may have a lower valuation for \( h \) than those in other plans: if so, the selective contract concentrates high-valuation consumers in the plans that contract with \( h \), so that a higher proportion of their WTP can be extracted in the form of premiums.

\textsuperscript{96}The plan that excludes the hospital does so not because the producer surplus that would be generated by the contract is negative, but because the price demanded by the hospital (and which other plans are willing to pay) is high enough to prevent the plan from capturing any profits from the contract. That is, a selective equilibrium emerges because of the interactions between the contracts negotiated by different plans in the market.

\textsuperscript{97}If this theory is correct, we should observe that all system hospitals generate a positive surplus with some plans in the market (otherwise the provider should be dropped from the system). In addition, all plans that agree on contracts with system hospitals should receive a positive surplus from the system as a whole. The data are consistent with these predictions.

\textsuperscript{98}The efficient outcome would result in the plan only contracting with hospitals with which it generated positive surplus. A system with high market power could demand a share of the profits generated from non-system hospitals. Some friction is required to prevent this outcome: this could be a cost of contracting that is paid once per non-system hospital and only once per system, or an inability of one hospital to transfer funds to another member of the same system to compensate it for lost revenues.
one-hospital case in the Appendix. I show, first, that a hospital’s ability to turn down low price offers forces plans to bid for contracts. Second, I demonstrate how variation in consumers’ underlying preferences can provide one plan with a lower WTP for the hospital than the other’s and can therefore lead to a selective equilibrium.

The model also considers welfare effects. Not only do the externalities faced by hospitals imply that my simple plan-level producer surplus measure is not enough to explain the data; they can also lead to an inefficient outcome. I demonstrate this possibility in the model in the Appendix. I choose the capacity constraints example for the detailed model because it raises the question: how can it be inefficient for a plan to exclude a hospital that is full in equilibrium? I show that an inefficiency can result even when no capacity is wasted and the consumers with the highest value for hospital $h$ are the ones treated. It is generated because consumers are forced to make suboptimal choices across health plans in order to gain access to the hospital. The resulting loss of consumer welfare, which may outweigh the gain derived when the highest-valuation patients are given preferential access to $h$, would be avoided if both plans contracted with it. The intuition is similar for the examples concerning systems and attractive hospitals.

The model would be much more complicated for markets that contained large numbers of asymmetric firms: analytic solutions may well not exist in these more realistic cases. In order to consider real-life markets I therefore turn to the data and an empirical estimation of the profits captured by specific types of hospitals.

### 3.6 Reduced Form Analyses and Identification

#### 3.6.1 Reduced Form Analyses

The theory outlined in Section 3.5 offers one possible reason why hospital-plan pairs which would generate positive producer surplus may not reach agreement: a hospital
that does not need contracts with all plans may charge a sufficiently high price that only those with the highest willingness-to-pay agree to the contract\textsuperscript{99}. In addition, the contracts that are agreed despite a negative estimated surplus may be explained by the penalties for excluding hospitals imposed by hospital systems. The full econometric model presented in Section 3.7 tests these theories by estimating the incremental profits captured by capacity constrained providers, star hospitals, and system members. In this section I conduct three straightforward analyses to demonstrate the sources of identification for the model.

The first analysis concerns hospital systems. Table 14 shows the results of a simple probit regression of the following form:

\[ \Pr(\text{contract}) = \Phi(\beta_0 + \beta_1 \text{system} + \beta_2 \text{samesysdrop}) \]

where \textit{system} is an indicator variable for membership of a hospital system and \textit{samesysdrop} measures the effect of the contract on the number of other hospitals in the plan’s network for which same-system members have been excluded. I also include market fixed effects. If the second system variable (\textit{samesysdrop}) is excluded the estimated coefficient \( \beta_1 \) is positive, indicating that system members are on average more likely to agree to contracts than other hospitals. When I add the second variable both coefficients become negative and significant. This implies that plans are less likely to agree to contracts with system hospitals than with other providers, and that they tend to contract either with an entire system or with none of its members.

\textsuperscript{99}The intuition is similar to that of a monopolist which restricts volume in order to maximize profits. Perfect price discrimination across consumers is impossible because enrollees are aggregated into plans, each of which charges a single premium, and may choose to move between plans. One effect of selective contracting is to concentrate high-valuation consumers into the plans that offer the highest prices to the hospital, reducing the negative impact of restricted volume on hospital profits. The main distinction between the monopolist example and the insurer-hospital case is that prices here are set by bargaining. The ability of a hospital to turn down low price offers also leads to an increase in its bargaining power with plans.
Table 14: Probit Analysis to Predict Contracts using Systems

<table>
<thead>
<tr>
<th></th>
<th>Coefficient estimate</th>
<th>Coefficient estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>0.079 (0.077)</td>
<td>-0.137* (0.080)</td>
</tr>
<tr>
<td>Samesysdrop</td>
<td>-0.769** (0.052)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.466** (0.144)</td>
<td>0.495** (0.151)</td>
</tr>
<tr>
<td>Market FEs</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Pseudo-R²</td>
<td>0.07</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Notes: N=2706 contracts. Standard errors in parentheses; **significant at p=0.05; *significant at p=0.1. "System" is an indicator variable for hospitals in systems. "Samesysdrop" measures the increase, when the contract is agreed, in the number of network hospitals for which a same-system hospital is excluded.

The probit analyses reported in Table 15 are similar. Here I consider capacity constrained hospitals: I use indicator variables for hospitals that were over 100 per cent and over 85 per cent of their maximum capacity in the previous year. The estimated effect of these measures on the probability of agreement is negative and significant, consistent with the theories set out above

---

100 Providers above their maximum capacity are defined as those with admissions > 365*number of beds/average length of stay. The results remain significant when I also add the change in surplus when the contract is agreed and the two systems variables used in the previous analysis. They are also robust to clustering the error terms by plan or by hospital.
Table 15: Probit Analysis to Predict Contracts using Capacity Constraints

<table>
<thead>
<tr>
<th>Coefficient estimate</th>
<th>Coefficient estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>cap_constr1</td>
<td>-1.008** (0.234)</td>
</tr>
<tr>
<td>cap_constr2</td>
<td>-0.411** (0.110)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.708** (0.145)</td>
</tr>
<tr>
<td></td>
<td>0.668** (0.142)</td>
</tr>
<tr>
<td>Market FEs</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Pseudo-R²</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
</tr>
</tbody>
</table>

Notes: N=2706 contracts. Standard errors in parentheses; **significant at p=0.05; *significant at p=0.1. "Cap_constr1" is an indicator variable for hospitals that were over capacity in the previous year; "cap_constr2" is an indicator variable for hospitals that were over 85% of capacity in the previous year.

The final reduced form analysis considers the overall relation of market-level capacity to the extent of selective contracting in the market. Not surprisingly, capacity constrained providers are most often observed in low-capacity markets. This clustering implies that the probability of disagreement should be high in these types of market. In addition, plans are obviously more likely to exclude hospitals in markets where this would have a lower effect on consumer demand. There is evidence that demand effects are lowest in markets with many beds per population: the

---

101 Low-capacity areas are defined here as markets with fewer than 2.9 beds per thousand population, the average value in the data. Low-capacity markets include Portland OR and San Jose CA. High-capacity markets include Pittsburgh PA and New Orleans LA. Beds per thousand population is a crude measure of hospital capacity; the full model also takes account of demographic information.

102 For example, 12% of hospitals in markets with less than the average bed capacity, and only 7% of hospitals in other markets, were capacity constrained in 2001; the correlation between capacity constraints and beds per thousand population is -0.48.

103 Hospitals in systems are not significantly clustered in either high- or low-capacity markets.

104 For example the correlation between the market’s capacity level and the average increase in expected utility when each hospital is added to a network containing all other hospitals is roughly -0.2.
second prediction is therefore that the probability of disagreement should rise with bed capacity once it reaches a point where few hospitals are expected to be full. Thus a U-shaped curve is predicted, with a high probability of disagreement in high- and low-capacity markets and a lower probability in intermediate markets. To test this prediction I conduct the following probit regression:

\[
\Pr(\text{nocontract}) = \Phi(\beta_0 + \beta_1 capacity + \beta_2 capacity^2 + \beta_3 capacity^3)
\]

The results are reported in Figure 4 together with a graph that summarizes them\textsuperscript{105}. The points on the graph show actual data: the per cent of hospitals excluded in each market plotted against actual bed capacity in the market. The fitted curve is U-shaped, as predicted by the model set out above.

\textsuperscript{105}When I include an additional term \(capacity^4\), its coefficient is insignificant and the overall picture for the other coefficients is unchanged.
Probit analysis to predict disagreement using hospital bed capacity

<table>
<thead>
<tr>
<th>Coefficient estimate</th>
<th>Coefficient estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bed capacity (beds per 1000 population)</td>
<td>-5.86** (1.08)</td>
</tr>
<tr>
<td>$Capacity^2$</td>
<td>1.78** (0.33)</td>
</tr>
<tr>
<td>$Capacity^3$</td>
<td>-0.17** (0.03)</td>
</tr>
<tr>
<td>Constant</td>
<td>4.93** (1.12)</td>
</tr>
<tr>
<td>Pseudo-R$^2$</td>
<td>0.02</td>
</tr>
<tr>
<td>N</td>
<td>2706</td>
</tr>
</tbody>
</table>

** significant at p=0.05; * significant at p=0.1

Effect of Bed Capacity on per cent of Hospitals Excluded

![Graph showing the effect of hospital capacity on the probability of exclusion](image)

Figure 4: Effect of Hospital Capacity on Probability of Exclusion
3.6.2 Identification

The reduced form analyses demonstrate the variation in the data that will be used to identify the full model. For example, the incremental profits captured by capacity constrained hospitals will be identified using variation in the probability of agreement across capacity constrained and non-capacity constrained providers, both within and across markets. The basic intuition is that, since we observe capacity constrained hospitals refusing to agree to contracts more frequently than other providers, they must demand higher profits than their competitors. The producer surplus generated by these hospitals when they agree on contracts with particular plans provides an upper bound on the profits they capture. The predicted producer surplus generated when they do not reach agreement offers a lower bound on their profits. We can estimate the average profit of capacity constrained hospitals by taking averages over these observations. A similar intuition applies to other types of hospital.

3.7 A Model for Estimation

3.7.1 The Plan Profit Equation

The reduced form results lend support to the theories discussed in Section 3.5. However, all the analyses so far have ignored the existence of multiple potential equilibria and the possibility of endogenous regressors. The next step is to estimate a model that takes these possibilities into account. In order to do so we need an equation to describe plan profits.

The profit of plan $j$ is the surplus generated given its chosen network $H_j$ minus its costs:

$$\pi_{jm}(H_j, H_{-j}) = S_{jm}(H_j, H_{-j}) - c_{jm}^{HOSP}(H_j, H_{-j}) - c_{jm}^{NONHOSP}(H_j, H_{-j})$$

(18)
where $c_{jm}^{HOSP}(H_j, H_{-j})$ is the cost of the plan’s contracts with hospitals (generated by hospital profits) and $c_{jm}^{NONHOSP}(H_j, H_{-j})$ represents its non-hospital costs.

I assume that each hospital receives a two-part payment: a fixed element and a per-patient markup, $f_{cj, h}(.)$ and $mk_{j, h}(.)$ respectively. If prices are set by bargaining both these quantities depend on hospital and plan threat points and are therefore functions of characteristics of the hospital, the plan, and the market as a whole. I would ideally use a model of the plan-hospital bargaining process to estimate their values directly. However, the fact that each firm’s threat point is endogenous (depending on the observed or expected outcome of all other pairs’ negotiations), together with the number of insurers and providers bargaining in each market, makes this approach infeasible. Instead I adopt a simpler methodology, closer to the reduced form analyses of the previous sections, by projecting hospital profits onto a set of hospital, insurer and market characteristics. That is, I estimate a reduced form function that summarizes the relationship of these variables to hospital profits. This approach allows me to estimate the parameters of the model without specifying the exact form of the price negotiation. I relax several assumptions made in the model in Section 3.5: in particular I no longer assume that insurers make take it or leave it offers or that they have complete information about other plans’ characteristics. The results from this relatively flexible model provide guidance for future modeling choices as well as information on the potential effects of policy and environmental changes on

106 Contracts in reality fall into at least three categories. Many plans pay hospitals on a per diem or case rate basis. The former involves a daily charge plus a separate charge for major procedures such as open heart surgery; the latter implies a single rate, usually for a surgery such as open heart surgery or organ transplants, that includes a specified number of inpatient days. Capitation contracts may also be used: here a hospital receives a fixed payment in return for which it provides or covers the cost of all hospital services needed by a designated population of enrollees.

107 The model in the Appendix demonstrates that the contract between a given plan-hospital pair is affected by the outcome of the negotiations between all other pairs in the market: for example the WTP of other insurers and the availability and price demands of other hospitals all depend on the outcomes of their own negotiations and all affect the threat points of both plan and provider. Modelling this set of negotiations explicitly would be very complicated given that there are on average 12 plans and 15 insurers in each market.
Markups$^{108}$. More specifically, the payment to each hospital is:

\[ pmt_{jhm}(.) = fc(x_j, x_h, x_m)\vartheta_1 + N_{jhm}(H_j, H_{-j})mk(x_j, x_h, x_m)\vartheta_2 \]  \hspace{1cm} (19)

where \((x_j, x_h, x_m)\) are plan, hospital and market characteristics, \(\vartheta_1\) and \(\vartheta_2\) are parameters to be estimated, and \(N_{jhm}(H_j, H_{-j})\) is the number of plan \(j\)'s enrollees treated by hospital \(h\)$^{109}$:

\[ N_{jhm}(H_j, H_{-j}) = \sum_i n_ip_is_{ijm}(H_j, H_{-j})s_{ih}(H_j). \]

I allow for two sources of randomness at this point. The first is an unobserved plan-specific fixed effect, \(\delta_j(\cdot)\), that is constant across choices. This can be thought of as a "large plan" effect that controls for the lower prices paid by plans with dominant market share. Secondly, there may be measurement error on the part of the econometrician: I denote this as \(u_{jH_j}\). A third error term, caused by observed differences in profitability across network choices that are not known to plans at the time decisions are made, will be introduced in equation (22).

Including \(\delta_j\) and \(u_{jH_j}\), and subtracting the sum across hospitals of the costs implied

---

$^{108}$Prior attempts to analyze the fraction of the surplus that goes to hospitals look at the marginal value of the hospital to a network conditional on assumptions regarding the networks in existence, but do not attempt to analyze the determinants of these networks. See for example Capps, Dranove and Satterthwaite (2004) and Town and Vistnes (2001).

$^{109}$Note that both this number of patients and the surplus term also depend on the \(x\)'s: the form of the dependence is modelled explicitly using the demand estimates from Ho (2004).
by equation (19), we obtain plan profits\(^{110}\) as:

\[
\pi_{jm}(H_j, H_{-j}, x, \delta_j, u_{jH_j}, \theta) = S_{jm}(H_j, H_{-j}) - \sum_{h \in H_j} fC(x_j, x_h, x_m)\theta_1 - \sum_{h \in H_j} N_{jhm}(H_j, H_{-j})mk(x_j, x_h, x_m)\theta_2 - c_{jm}^{\text{NONHOSP}} + \delta_j(H_{-j}, x_j, x_m) + u_{jH_j} \tag{20}
\]

\[
= \tilde{\pi}_{jm}(H_j, H_{-j}, x, \theta) + \delta_j(H_{-j}, x_j, x_m) + u_{jH_j}. \tag{21}
\]

The fourth term in equation (20) relates to non-hospital costs. These are likely to vary with the sickness level of the individual patient. I would ideally estimate the cost of enrolling each type of consumer, defined by age and sex, by including the following expression in the profit equation:

\[
c_{jm}^{\text{NONHOSP}}(H_j, H_{-j}) = \sum_i N_{jim}(H_j, H_{-j})c_i
\]

where \(N_{jim}(H_j, H_{-j})\) is the predicted number of enrollees of type \(i\) in plan \(j\) given the equilibrium hospital networks and \(c_i\) is the cost of insuring that type (to be estimated). Unfortunately the available data are not rich enough to estimate \(c_i\) in addition to the hospital cost parameters for more than one or two consumer types. In the main specification I set \(c_i = 0\) for all \(i\), assuming that non-hospital costs have little effect on plans’ network choices. As a robustness test I consider the effect of including the predicted total number of plan enrollees and the fraction of enrollees

\(^{110}\) A number of existing papers estimate the share of the surplus captured by a given hospital as a function of its characteristics and those of the plan and the market (conditional on the existence of the contracts; see Capps, Dranove and Satterthwaite (2003), Town & Vistnes (2001)). An analogous methodology would estimate the plan’s share of the incremental surplus created when each hospital was added to the network. The current approach is similar. If we write \(\pi_{jm} = \alpha(x_j, z_{jm}, \mu_q)S_{jm} = \sum_h N_{jh}\alpha_{jh}(\text{prem}_{jm} - \cos t_h)\), where \(\alpha_{jh}\) is the share of the surplus retained by the plan when negotiating with hospital \(h\), and ignore plan non-hospital costs, this is equivalent to: \(\pi_{jm} = S_{jm} - \sum_h N_{jh}(1 - \alpha_{jh})(\text{prem}_{jm} - \cos t_h)\). The value estimated by the current methodology’s markup term (if \(fC_{j,h}(.) = 0\)) is \(mk_{j,h}(.) = (1 - \alpha_{jh})(\text{prem}_{jm} - \cos t_h)\). That is, in the absence of a prediction for the hospital’s effect on premiums and therefore the surplus per patient, \(mk_{j,h}(.)\) estimates not \((1 - \alpha_{jh})\) but the average profit per patient captured by the hospital.
aged 55-64\textsuperscript{111}. The tests are discussed further in Section 3.8; they have little effect on the overall results.

The next step is to decide which variables to include in the expressions for fixed costs and markups. The list must be parsimonious: a large number of coefficients is unlikely to be identified given the limited data available and the fairly small variation in plan choice of networks observed. I use the theories discussed in Section 3.5 to inform the choice of variables. The main predictions are that hospitals that expect to be capacity constrained, those in systems, and those for which all or most consumers would switch plans should be most likely to fail to agree with plans (holding surplus fixed). System hospitals may also demand a higher price from a given plan if another same-system hospital is excluded than if it is not. Finally, even very simple bargaining models predict that lower-cost providers generate a higher total surplus, all else equal, and therefore earn higher markups than their competitors. I account for these predictions by including the following variables:

1. A measure of the extent to which particular hospitals are expected to be capacity constrained. I derive an exogenous predictor of this variable by calculating the number of patients treated at each hospital under the thought experiment that every plan contracts with every hospital in the market\textsuperscript{112}. This variable can therefore be thought of as an indicator for potentially capacity constrained hospitals\textsuperscript{113}.

\textsuperscript{111}I ignore the population aged over 64 because these consumers are eligible for Medicare.

\textsuperscript{112}I define a hospital to be capacity constrained if the predicted number of patients exceeds the number of beds * 365 / average length of stay in the hospital.

\textsuperscript{113}As noted in Section 3.3, the services provided by the hospital are key predictors of capacity constraints. Table 18 sets out the results of a probit regression which indicates that hospital cardiac and imaging services are positively and significantly related to the capacity constraints variable. The coefficients on teaching status and system membership are not significant. Similarly, distance from the city center is not significantly related to capacity constraints. Distance from consumers’ residences is an important predictor of demand for hospitals, but this does not imply that only city-center hospitals, or only those located in suburban areas, are predicted to be full.
2. Hospitals in systems, and those for which at least one same-system hospital is excluded.

3. Star hospitals: those that can expect many enrollees to switch plans to ensure access. I identify these hospitals using the US News and World Report’s hospital rankings for 2003, considering in particular each hospital’s reputation ranking among physicians\footnote{US News magazine publishes an annual report giving hospital rankings for 17 different specialties and overall. The overall rankings comprise scores for reputation; severity-adjusted mortality ratios; and other care-related factors such as the number of nurses per bed and the technology available. The reputation score was compiled by asking a random sample of board-certified physicians which five hospitals they believed to be the best in their specialty. The rating is the percentage of responding physicians who cited the hospital. 583 of the hospitals in the sample (88\%) have a reputation score of zero; the average nonzero value is 0.06 and the highest are 0.88 (Johns Hopkins Hospital, Baltimore) and 0.62 (Massachusetts General Hospital, Boston). I use the actual reputation rankings as a measure of star hospital status.}.

4. A measure of hospital costs per admission\footnote{I also tried using costs per bed per night rather than costs per admission; this generated very similar results.}.

5. I also include a constant term in $mk_{j,h}(\cdot)$: this identifies the average profit per patient received by non-system hospitals that are not capacity constrained\footnote{This term enables me to move away from the strict take it or leave it offers model to a framework where all hospitals receive positive profits.}.

In reality the profit received by a particular provider depends not just on its own characteristics but on those of the plan and the market. For example, the price demanded by a system hospital can be no higher than the maximum willingness-to-pay of other plans in the market, and this depends on the attributes of other plans, consumers, and hospitals in the area. I would ideally include plan and market characteristics, and interactions with network attributes, to identify these effects. However, I have difficulty in identifying the coefficients on these terms\footnote{Plan and market characteristics that are not interacted with network attributes do not vary across potential choices for a given plan and therefore cannot affect its choice. These characteristics therefore cannot be identified in the fixed cost term unless interacted with network attributes. It}.
unrealistic, given my limited data, to hope to estimate more than the most basic effects. The results reported therefore have no market characteristics: they identify only the average dollar profit per patient earned by each type of hospital\textsuperscript{118}.

There is not enough information in the data to allow for free interactions with both the fixed and the per patient component of the contracts. The results presented below are based on a specification where the fixed component of the contract depends on whether the hospital is in a system, whether another member of that system is excluded by the plan, and whether it is a star hospital, and the variable component depends on whether the hospital is capacity constrained and the cost per admission of the hospital. When I estimated models allowing both sets of variables to affect both the marginal and fixed components, the individual coefficients were insignificant but there was little difference in the implications of the estimates.

\textbf{3.7.2 Details on the Estimation Strategy}

The standard models that might be used to estimate the plan profit equation (such as the logit model) would use the profits for the different networks given by equation (20) and assume that plans chose networks to maximize these profits. We would then make the additional assumption of iid errors and estimate using maximum likelihood. However, the independence assumption may be difficult to accept for at least two reasons. First, econometrician measurement error leads to a correlation between the errors and the other right hand side variables of the plan profit equation (such as the surplus that is observed by the econometrician). In order to account for this we would need to know the joint distribution of the errors and the observed profit makes more sense to include these variables in the markup term, where they will be interacted with $N_{jk}$; however, in practice there was not enough variation in the data to generate significant coefficients.

\textsuperscript{118}The capacity constraints variable is calculated using the predicted allocation of patients across hospitals when all plans offer a free choice: it therefore incorporates information on market characteristics. The other variables, however, relate only to hospital characteristics.
determinants; we are unlikely to have information on this joint distribution. Second, plan prediction error causes analogous problems\textsuperscript{119}.

Instead I use the methodology presented in Chapter 4 which uses a method of moments approach with inequality constraints. I assume that plans choose their networks in a simultaneous-moves game, conditional on their expectations regarding other plan choices and the prices demanded by hospitals. The identifying assumption is the necessary condition that plan $j$’s expected profits from the observed $H_j$ are higher than the expected profits from any alternative $H_{j,a}$\textsuperscript{120}. That is, I assume that:

$$E(\pi_{jm}(H_j, H_{-j}, x, \delta_j, u_{jh}, \vartheta)|I_{jm}) \geq E(\pi_{jm}(H_{j,a}, H_{-j}, x, \delta_j, u_{jh}, \vartheta)|I_{jm})$$

for all alternatives $H_{j,a}$ where $I_{jm}$ is the information known to plan $j$ when it makes

\textsuperscript{119}One alternative is to ignore these two errors and assume a different source of randomness: that plans have access to information about profit differences across networks that is not available to the econometrician. If these errors are observed to all plans, it is still feasible to estimate equation (20) using maximum likelihood. Endogeneity issues arise again, however, because $H_j$, which is part of the right-hand-side variables, reacts to $H_j$ and will therefore be correlated with the error. Some previous authors have estimated models assuming this type of error. For example, Andrews, Berry and Jia (2004) and Ciliberto and Tamer (2003) address the problem by identifying every potential equilibrium for every parameter value evaluated, rather than assuming that each firm conditions on other firms’ expected choices. However, the number of plans and potential networks per plan in the problem considered here leads to combinatorics that make their approach infeasible. My current specification does not account for this type of random disturbance.

\textsuperscript{120}A number of bargaining models are consistent with the estimation procedure. We require only that a contract is observed if and only if it increases the plan’s expected profits. The simple model in Section 3.5 and Appendix A is consistent with this framework. (We can assume, for example, that plans make initial offers to all hospitals; capacity constrained hospitals make high counter-offers to all plans and only the plans with the highest willingness-to-pay accept these counter-offers. In the full information case plans may choose not to make initial offers to hospitals with which they predict they will never agree. In either case the condition required for estimation is satisfied.) Other possible models include the version where hospitals make take-it-or-leave-it offers to all plans in the market and any model of sequential offers (e.g. plans make offers to all hospitals; hospitals make counter-offers to all plans) provided there is an exogenous end-point, no contracts can be agreed before this point, and plans make the final contract choice. Finally, any model where plans choose which hospitals to bargain with and then reach agreement in every negotiation would meet the requirements. These "back and forth" models of bargaining are probably more realistic than the simplest take-it-or-leave-it offers models.
its choice\textsuperscript{121,122}. Taking expectations here admits the possibility of the third source of randomness mentioned in Section 3.7.1: that each plan may predict its profits from any given network with error, perhaps because of inaccurate predictions of other insurers’ network choices. I denote this error term $\phi_{jH_j}$, where:

$$\pi_{jm}(H_j, H_{-j}, x, \delta_j, u_{jH_j}, \vartheta) = E(\pi_{jm}(H_j, H_{-j}, x, \delta_j, u_{jH_j}, \vartheta) | I_{jm}) + \phi_{jH_j}$$

and $E(\phi_{jH_j} | I_{jm}) = 0$ by construction. We will require a set of instruments $z_{jm}$ such that $z_{jm} \in I_{jm}$, the plan’s information set, and

$$E(u_{jH_j} | z_{jm}) = 0. \quad (23)$$

Now consider a finite subset of alternatives $H^A_j = \{H_{ja}\}_{a=1}^A$. For any particular value of $\vartheta$ the value of $\pi_{jm}(H_j, H_{-j}, x, \delta_j, u_{jH_j}, \vartheta)$ is not observed (since the $\delta_j$ and $u_{jH_j}$ are not observed). However, the vector

$$\Delta\pi_{jm}(H^A_j, H_j, H_{-j}, x, \vartheta) = \tilde{\pi}_{jm}(H_j, H_{-j}, x, \vartheta) - \tilde{\pi}_{jm}(H_{ja}, H_{-j}, x, \vartheta)$$

where $\tilde{\pi}_{jm}(.)$ is defined by equation (21), is observable and can be computed. The assumptions imply that:

$$E \left[ \Delta\pi_{jm}(H^A_j, H_j, H_{-j}, x, \vartheta) \mid z_{jm} \right] \geq 0. \quad (24)$$

\textsuperscript{121}Several combinations of networks may satisfy this necessary condition; that is, there may be multiple potential equilibria. This does not prevent consistent estimation of the parameter vector $\vartheta$: I simply search for parameters consistent with the assumption that the observed set of networks constitute a Nash equilibrium, without attempting to model how that equilibrium was chosen from the set of potential equilibria.

\textsuperscript{122}One further assumption is needed. We have to assume that, when one plan deviates from its observed network, the others still succeed in securing the networks they bid for. This assumption seems reasonable since only small deviations are considered for just a single plan ($j$). In addition, the inequalities used for identification would still hold even if a plan was forced to drop one of the hospitals it chose to bid for, since such an action would increase the surplus generated by plan $j$. The parameter estimates should therefore still be consistent even in this case.
Note that the unobserved terms have dropped out of this equation. Equation (23) takes care of the econometrician measurement error $u_{jH_j}$. The plan’s prediction error, $\phi_{jH_j}$, drops out because $E(\phi_{jH_j} \mid z_{jm}) = 0$ for all $z_{jm} \in \mathcal{I}_{jm}$. The unobserved plan fixed effects are differenced out when we calculate the difference between the profits from observed and alternative choices. Translating expectations into sample means, the equation for estimation is therefore:

$$
\frac{1}{M} \sum_m \sqrt{n_m} \sum_{j=1}^{n_m} \left[ \Delta \pi_{jm}(H_j^A, H_j, H_{-j}, x, \theta) \otimes g(z_{jm}) \right] \geq 0 \tag{25}
$$

where $M$ is the number of markets in the sample, $n_m$ is the number of plans in market $m$, $\otimes$ is the Kronecker product operator and $g(z)$ is any positive-valued function of $z$. Each market is weighted by the square root of the number of plans in the market, since we expect less noise in the market average for markets containing many plans. All $\theta$ that satisfy this system of inequalities are included in the set of feasible parameters. If no such $\theta$ exists we find values that minimize the sum of the absolute values of the amount by which each inequality is violated.

Identification in this model comes from comparing the profits of each plan when it chooses its observed network to those from its alternatives. For example, the identifying assumption implies that, if the plan is observed to contract with a capacity constrained hospital, then the change in producer surplus it expects to result from the contract must be greater than the hospital’s expected profits. If the plan drops the provider, the expected change in producer surplus must be less than those profits. Any feasible alternative networks could be used to generate these comparisons. I consider six alternatives $H_{j,a}$: these are defined by reversing the plan’s contracts with each of the six largest hospitals in turn\textsuperscript{123}.

\textsuperscript{123}The methodology could easily be extended to more alternatives per plan. For example, each could consider reversing two contracts at a time rather than just one. I try including these additional alternatives as a robustness test, and find little change in the overall results. I limit my main analysis to considering only six alternatives because of concern that the reduced form function for hospital...
The instruments are required to be independent of the error terms $\phi_{jH_j}$ and $u_{jH_j}$; they must also be positive (to ensure that no inequalities are reversed by the interaction with $z$). I use the characteristics included in the fixed cost and markup terms (the $x$’s) other than the cost per admission, which I omit due to concerns about measurement error. I also include indicator variables for the following market and plan characteristics: a high number of beds per population, a high proportion of the hospitals in the market being in systems, a high proportion of the population aged 55-64, whether the plan is local, whether the plan has good breast cancer screening services and poor mental health services, and some of these characteristics interacted with the standard deviation of the distance between hospitals in the market (which as noted earlier is higher in selective than unselective markets because of its effect on consumer demand)\textsuperscript{124}. None of these instruments is a function of the observed equilibrium. Each is known to the plan when it makes its choice. Each is also correlated with $x$: for example, plans can more easily exclude hospitals in markets with a younger, less sick population or with more beds per population. System hospitals are more often excluded in markets with a high proportion of hospitals in systems. The logic is similar for the other instruments.

The final step is to calculate confidence intervals for the parameter estimates. I use the simulation methodology described in Chapter 4, estimating the limit distribution of the data used to define the inequalities, taking repeated draws on this distribution, and calculating a new estimate for each draw. The resulting vector of simulated values is used to find a 95 per cent confidence interval\textsuperscript{125}.

\textsuperscript{124}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.

\textsuperscript{125}The confidence intervals have not yet been adjusted to account for variance introduced by the estimated demand parameters. This is unlikely to significantly affect the results since the standard

profits could change after a major network change. I also tried allowing every plan to reverse its contract with every hospital in the market in turn rather than considering just the six largest hospitals; this too had little impact on the results.
3.8 Results

3.8.1 Overall Results

The results are reported in Table 16. The estimate of $\theta$ for every specification was a singleton: that is, there was no parameter vector that satisfied all the inequality constraints\textsuperscript{126}. The first column of the table reports results for the main specification. The point estimates all have the expected sign, where a positive sign implies a positive relationship to hospital profits, and have magnitudes which are consistent with the available information from other sources. Three of the five coefficients are significant at the traditional five per cent level, the "hospital in a system" indicator is significant at the 10 per cent level in a one-sided test, and the constant term in the markup is significant at the 12 per cent level, again in a one-sided test. However, the confidence intervals are reasonably large. The graphs in Figures 5 and 6 show the simulated distributions of four of the coefficients. They are clearly not Normally distributed. Each distribution is left-skewed, with most of its mass between zero and about a third of the upper bound to the confidence interval, but even with this caveat there is significant variance about the point estimates. This together with the robustness tests noted below makes statements about precise magnitudes difficult. The overall picture, however, is very clear. Hospitals in systems take a larger fraction of the surplus and also penalize plans that do not contract with all members. Capacity constrained hospitals also capture high markups, and hospitals with higher costs per patient receive lower markups per patient than other providers.

\textsuperscript{126} As discussed in Chapter 4, this does not imply that we should reject the specification. The result could easily be caused by the random disturbances in the inequalities. The probability that all inequalities are satisfied can be made arbitrarily small by increasing the number of inequality restrictions.
Table 16: Results of Full Model for Estimation

<table>
<thead>
<tr>
<th>Hospital Characteristics</th>
<th>Coefficient Estimate</th>
<th>Simulated 95 per cent CI</th>
<th>Coefficient Estimate</th>
<th>Simulated 95 per cent CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Component (Unit = $ million per month)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital in System</td>
<td>0.179</td>
<td>[-0.134, 0.624]</td>
<td>0.174</td>
<td>[-0.095, 0.543]</td>
</tr>
<tr>
<td>Drop Same System Hospital</td>
<td>0.595</td>
<td>[0.068, 1.810]</td>
<td>0.615</td>
<td>[0.057, 4.370]</td>
</tr>
<tr>
<td>US News Reputation</td>
<td></td>
<td></td>
<td>1.230</td>
<td>[-16.26, 2.579]</td>
</tr>
<tr>
<td>Per patient Component (Unit = $ thousand per patient)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capacity Constrained</td>
<td>1.867</td>
<td>[0.406, 11.39]</td>
<td>2.487</td>
<td>[0.130, 13.75]</td>
</tr>
<tr>
<td>Cost per Admission</td>
<td>-0.263</td>
<td>[-3.397, -0.20]</td>
<td>-0.131</td>
<td>[-2.471, -0.082]</td>
</tr>
</tbody>
</table>

Notes: There are 451 plans in each specification. 95 per cent confidence intervals are reported in parentheses. Positive coefficients imply positive relationships to hospital profits. "Hospital in system" refers to whether the hospital is in a system; "Drop Same System Hospital" refers to an indicator for hospitals for which a same-system hospital has been excluded; "US News Reputation" is the star hospital measure discussed in Section 3.7. Capacity constrained hospitals are those with predicted admissions (when all plans contract with all hospitals) > number of beds * 365 / average length of stay.
Figure 5: Simulated Distribution of Coefficients, Full Model
To help interpret the magnitudes of the results, note that the average cost per admission for hospitals in the data is around $11,000\textsuperscript{127}. The markups over these costs that I estimate vary by cost and type of hospital. For hospitals that are neither in a system nor capacity constrained the point estimates imply very low average markups of around two per cent of revenues. I estimate that capacity constrained hospitals receive an extra $1800 per patient which, when their costs are taken into account, translates into an average markup of approximately 15 per cent of revenues. Hospitals that are not capacity constrained but are in systems capture $179,000 in incremental

\textsuperscript{127}The cost variable, taken from the AHA survey 2001, is defined as total hospital expenses including items such as depreciation and interest expense.
profits per month per plan, which given their average patient load translates into a markup of about $1800 per patient. When costs per admission are also taken into account, system hospitals are predicted to have average profits of around 14 per cent of revenues, 12 percentage points higher than other hospitals. I also estimate a large penalty for excluding a hospital from a system: however, this happens only rarely. We can compare these results to a published figure: the average hospital markup for community hospitals has been estimated at about four per cent of revenues.\footnote{The Kaiser Family Foundation report "Trends and Indicators in the Changing Health Care Marketplace, 2004 Update" provides data on hospital costs and profits. The average profit margin for community hospitals was 4.2\% of revenues in 2001. This may be indicative that the estimated constant in the markup term is slightly inaccurate. The constant is the most imprecisely estimated of all the coefficients and if one looks at its distribution (shown in Figure 6) it is easy to see that the point estimate may well be different from its actual value. (A value of 2.50 rather than 2.61, for example, would translate to an average markup of one per cent of revenues for non-system, non-capacity constrained hospitals.)}

The second column of Table 16 adds the US News reputation measure of star hospitals to the specification. This coefficient has a very large confidence interval, possibly because there are very few star providers in the dataset. However, adding this measure changes the other coefficients only slightly.\footnote{I also tried using an indicator for teaching hospitals to measure star status, with similar results.}

### 3.8.2 Robustness Tests

The results in the previous section are consistent with the theory outlined in Section 3.5: that hospitals in systems and those that are expected to be capacity constrained seek rents and are optimally excluded by some plans in equilibrium. Hospitals in systems also seem to demand higher prices from plans that exclude their partners than from other plans. Could some other effect be causing these results? One possibility already discussed is that plans avoid contracting with hospitals that would attract enrollees for whom the non-hospital cost of insurance is high.\footnote{The producer surplus variable takes account of hospital costs. In particular it accounts for changes in the number of admissions to hospital when the age or gender profile of a particular plan’s enrollees changes. However, it does not allow the average cost per admission of a particular hospital} The first robustness
test therefore considers the effect of accounting for plan non-hospital variable costs. I add two variables to the plan profit equation: the number of enrollees and the fraction aged 55-64. The results are reported in Table 17. The addition had little effect on the overall results: the key parameters actually increased in magnitude. The costs per enrollee are very imprecisely estimated but the magnitudes make sense: they imply a $40 per month cost of insuring each young enrollee and an additional cost of $16,500 per month for every percentage point increase in the fraction of enrollees who are aged 55-64. This translates to an additional cost of insuring an older consumer of approximately $65 per month.

Inaccuracies in the estimated demand system could cause problems with the capacity constraints variable. For example, if the demand for hospital $h$ is biased up, so that the surplus increase when the hospital is added to plan $j$’s network is inflated, this would also imply an upward bias on the estimated hospital profit\textsuperscript{131}. This is an alternative explanation for the estimated profits of capacity constrained hospitals: hospital $h$ could be predicted to be full simply because its demand is biased up; both surplus and hospital profits would also then be (mechanically) overestimated\textsuperscript{132}. I test for this by replacing the indicator variable for predicted capacity constraints with a variable less closely tied to the demand estimates: an indicator for hospitals that were full in the previous year. I choose not to include this variable in the main specification because it is endogenous: any serial correlation in the disturbance from the model to vary with the age and sex of the patients admitted. This implies an assumption that the average patient mix of each hospital is fixed even when plan patient mixes change.

\textsuperscript{131}This is particularly likely to be a problem if plan $j$ is horizontally differentiated on a dimension not identified by the model. In that case the plan’s estimated average quality would be biased up (to explain its ability to exclude hospitals); its increase in surplus when excluded hospitals are added would also be inflated. I exclude the plans that are most clearly horizontally differentiated from the dataset: these are Kaiser Permanente, Group Health Cooperative of Puget Sound and Scott & White Plan of Austin TX. It is possible that this is still an issue; however, the robustness tests described in this paragraph are reassuring.

\textsuperscript{132}This argument would not, however, explain the reduced form results for capacity constrained hospitals.
would induce a bias in its coefficient. However, the endogeneity implies a negative bias, so a positive coefficient is still meaningful. The results are reported in Table 17. They are comparable to the main model: the capacity constraints coefficient is smaller than that in the main specification but still positive. This is reassuring; while the predicted surplus increase when hospitals are added to the network could still be biased up, it is not obvious why this should be more of a problem for hospitals that were capacity constrained the previous year than for other hospitals.

The final test concerns the assumption of fixed premiums. If premiums would in reality fall when certain types of hospital were dropped then both the surplus increase from adding them, and the hospital profit needed to explain plans’ unwillingness to agree contracts, would be biased down. As already mentioned, I cannot account perfectly for premium adjustments in response to network changes because I do not observe premiums over time. However, I can use my estimated results to perform a robustness test. I allow all plans to simultaneously adjust premiums to maximize their profits (revenues less prices paid) where prices are determined by the estimates from the main specification\(^\text{133}\). This premium adjustment is conducted as part of the producer surplus calculation for all networks considered; the supply side estimation is then repeated using the new measure of producer surplus. The results are reported in Table 17: they are fairly similar to those for the main analysis, although the magnitudes of some of the coefficients change.

Though none of the robustness tests change the qualitative nature of the results, some of the coefficients do change in magnitude. This is consistent with the results of a number of other robustness tests not reported here, and again implies that conclusions about overall effects can be drawn from the results but that it is difficult to make

\(^{133}\)I impose these predicted "optimal" premiums for both observed and unobserved contracts. The predicted values are quite different from those observed in the data. This is not surprising given the rudimentary nature of the model used: in particular, ignoring plans’ non-hospital costs and assuming fixed hospital prices across types of procedure probably generates significant noise.
statements about precise magnitudes.

**Table 17: Robustness Tests**

<table>
<thead>
<tr>
<th>Hospital Characteristics</th>
<th>Main Specification</th>
<th>Number Enrollees</th>
<th>Last Year Cap Con</th>
<th>Premium Adjustments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrollees</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.15, 0.16]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per cent Aged 55-64</td>
<td>1.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.47, 5.49]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fixed Component (Unit = $ million per month)**

<table>
<thead>
<tr>
<th></th>
<th>0.179</th>
<th>0.50</th>
<th>0.28</th>
<th>0.59</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[-0.134, 0.624]</td>
<td>[0.06, 0.71]</td>
<td>[-0.24, 2.26]</td>
<td>[0.02, 0.80]</td>
</tr>
<tr>
<td>Hospital in System</td>
<td>0.595</td>
<td>0.68</td>
<td>0.85</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>[0.068, 1.810]</td>
<td>[0.08, 2.10]</td>
<td>[0.07, 4.34]</td>
<td>[0.12, 2.12]</td>
</tr>
</tbody>
</table>

**Per patient Component (Unit = $ thousand per patient)**

<table>
<thead>
<tr>
<th></th>
<th>2.608</th>
<th>0.61</th>
<th>1.46</th>
<th>1.45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity Constrained</td>
<td>1.867</td>
<td>2.28</td>
<td></td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>[0.406, 11.39]</td>
<td>[-0.40, 2.79]</td>
<td></td>
<td>[-4.75, 14.7]</td>
</tr>
<tr>
<td>Last Year Capacity Constrained</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-2.72, 6.71]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost per Admission</td>
<td>-0.263</td>
<td>-2.09</td>
<td>-0.18</td>
<td>-1.67</td>
</tr>
<tr>
<td></td>
<td>[-3.397, -0.20]</td>
<td>[-2.28, -0.31]</td>
<td>[-5.94, 0.04]</td>
<td>[-3.07, -0.30]</td>
</tr>
</tbody>
</table>

Notes: N=451 plans. 95% confidence intervals reported in parentheses. The first test includes the number of enrollees and the fraction aged 55-64. The second replaces the predicted capacity constraints variable with an indicator for previous year capacity constraints. Finally, the surplus term is adjusted for premium changes in response to changes in network.
One further issue should be mentioned here: I have no data on plans’ physician networks and therefore cannot account for them in the model. It is possible that a plan might decide not to contract with a particular hospital because this would involve establishing new physician contracts. There is no obvious reason why these physician contracting costs should be higher for a hospital that expects to be capacity constrained than for other hospitals, but this point might go some way to explaining the result for hospital systems. If the physician networks associated with two hospital systems do not overlap, this provides an additional incentive for a plan to contract with all of one system or all of another rather than taking some hospitals from each. In the absence of relevant data it is difficult to say more on this issue; it may mean that the monetary costs of excluding a same-system hospital are overstated.

Are there other possible explanations for the results set out here? We could tell a story where teaching hospitals, or suburban providers, contracted with a subset of plans for reasons not related to bargaining. For example, teaching hospitals might prefer to concentrate on research rather than treating patients; suburban hospitals might only contract with the plans that covered their particular geographical areas. If these hospitals also tended to be capacity constrained this would confound our results. However, Table 18, which reports the results of a probit regression of an indicator variable for predicted capacity constraints on hospital characteristics, provides evidence against this idea. Neither teaching hospital status nor distance from the city center is significantly related to predicted capacity constraints. Alternatively, it is possible that capacity constrained hospitals turn down contracts with plans not because they wish to drive up prices but because their costs increase when they reach full capacity. However, in that case we should observe capacity constrained hospitals investing in new beds at a faster rate than other providers. In fact there is no significant difference between the investments made by the two types of hospitals in my
data\textsuperscript{134}. In addition, the data show that capacity constrained hospitals are clustered in markets with small numbers of hospitals, that is in low-competition markets. While this may simply be a mechanical correlation (hospitals are more likely to be full if there are few hospitals in the area), it is also consistent with the idea that hospitals limit their capacity only in markets where they have enough influence to capitalize on the resulting negotiating power\textsuperscript{135}.

Table 18: Probit Analysis to Predict "Predicted Capacity Constraints"

<table>
<thead>
<tr>
<th></th>
<th>Coefficient estimate</th>
<th>Coefficient estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiac services</td>
<td>0.967** (0.436)</td>
<td>1.334** (0.527)</td>
</tr>
<tr>
<td>Imaging services</td>
<td>0.748** (0.278)</td>
<td>0.924** (0.322)</td>
</tr>
<tr>
<td>Cancer services</td>
<td>-0.011 (0.217)</td>
<td>0.005 (0.244)</td>
</tr>
<tr>
<td>Birth services</td>
<td>0.078 (0.258)</td>
<td>0.038 (0.288)</td>
</tr>
<tr>
<td>Teaching hospital</td>
<td>0.142 (0.186)</td>
<td>0.128 (0.213)</td>
</tr>
<tr>
<td>System</td>
<td>0.109 (0.174)</td>
<td>0.116 (0.198)</td>
</tr>
<tr>
<td>Distance from city center</td>
<td>0.004 (0.007)</td>
<td>0.003 (0.008)</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.929** (0.464)</td>
<td>-3.515** (0.700)</td>
</tr>
<tr>
<td>Market FEs?</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pseudo-R\textsuperscript{2}</td>
<td>0.07</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Notes: N=665 hospitals. Standard errors are reported in parentheses; **significant at p=0.05; *significant at p=0.1. Cardiac, imaging, cancer and birth services refer to the four hospital service variables defined in Section 2.9.1. Distance from city center is measured in miles from the City Hall.

\textsuperscript{134}Capacity constrained hospitals increased their number of beds between 1997 and 2001 by 5.4 beds (3.7%); the equivalent value for other hospitals was 2.5 beds (3.2%). The difference is not significant at p=0.1.

\textsuperscript{135}The correlation of the percent of hospitals in the market that are capacity constrained to the number of hospitals in the market is -0.44.
3.9 Implications of the Results

3.9.1 Explaining the Observed Contracts

The producer surplus term considered alone explained just over 50 per cent of the observed contracts. We can now compare the performance of the full model to this benchmark. I use equation (20) to predict the change in plan profits when each hospital is excluded from each plan’s network, holding other plan choices fixed. I find that the new profit equation explains 62 per cent of the observed contracts overall: 66 per cent of the 400 contracts that were not agreed and 61 per cent of the 2306 contracts that are observed in the data. This is a considerable improvement on the 52 per cent and 54 per cent explained by the surplus term alone. The pseudo-$R^2$ value increases from 0.46 to 0.67.

When we aggregate to the market level the clustering of types of hospitals in specific areas also explains much of the cross-market variation in the data. The reduced form results show that markets with low and high bed capacity are more selective than intermediate areas. As already mentioned this can be explained by the higher proportion of capacity constrained hospitals in low-capacity markets and the lower demand effects of selective contracting in high-capacity areas. The estimates are also consistent with the interview evidence that the dominant influence on network decisions belongs to insurers in some markets and to providers in others: it makes sense that hospitals should dominate both in low-capacity markets and in areas where many hospitals have merged to form systems. (In Salt Lake city, for example, two systems own six of the nine largest hospitals; we would expect hospitals to have high leverage here.) Plan power should be high in high-capacity markets with few systems.
3.9.2 Investment Incentives for Capacity Constrained Hospitals

The benefit that hospitals derive from capacity constraints implies a potential disincentive to invest that may have negative welfare effects. I consider this issue both in general and for three specific hospitals that are predicted to be capacity constrained: St. Luke’s Medical Center in Milwaukee WI, SW Texas Hospital in San Antonio TX, and South Austin Hospital in Austin TX. For each I calculate the change in consumer surplus, plan profits and hospital profits that the model predicts would occur if the capacity constraints were removed.\(^{136}\)

Of course the removal of capacity constraints could affect plans’ network choices. The model in this chapter cannot predict the new equilibrium outcome for contracts, since as mentioned I do not fully detail how a single configuration of hospital networks is chosen from the multiple potential equilibria. In addition the change could lead to investment by other hospitals, or to entry or exit. The model cannot predict such developments. I therefore do not attempt a full analysis of the new equilibrium; I limit myself to a simple outline of the impact of the change if plans’ choices of networks and all other market characteristics were fixed. Even this requires an additional assumption: that the institutional changes do not affect the reduced form function used to describe hospital markups. While not entirely realistic this enables me to derive at least an approximate estimate of magnitudes. Finally, I also assume fixed premiums throughout the calculation.

The consumer surplus calculation finds the dollar value of consumers’ gain in util-

\(^{136}\)The calculations in this section assume that the relationship identified between capacity constraints and markups is causal. In terms of the model discussed in Section 3.5: if the hospital increased its capacity so that it was no longer capacity constrained, it might optimally accept all plan offers since the revenues lost from non-switching enrollees when it refused to agree to a low offer could now outweigh the increase in revenues from those who were willing to switch. Its threat to turn down low price offers would no longer be credible, so plans would revert to offering a price that just covered its costs and expecting the offer to be accepted.
ity when reallocation of patients from full hospitals to other, less-preferred providers is no longer necessary. The utility gain is defined as:

$$\Delta CS_m = \sum_j \sum_i n_i s_{ijm} \gamma_1 \lambda_i (EU_{ijm}^{\text{nocapcon}} - EU_{ijm}^{\text{capcon}})$$

where $EU_{ijm}^{\text{capcon}}$ is the utility a consumer with perfect foresight would expect to receive from the hospital network given the probability of reallocation away from the full hospital, $EU_{ijm}^{\text{nocapcon}}$ is the expected utility for the network when the capacity constraints are removed, and $\gamma_1$ and $\lambda_i$ are the coefficients on the expected utility from the network and premium in the plan demand equation respectively.\(^{137}\)

The plan profit calculation uses the expression for profits given by equation (20). The profit change from the removal of capacity constraints is given by:

$$\Delta \pi_{jm}^{\text{capcon}} = \pi_{jm}^{\text{nocapcon}}(H_j, H_{-j}) - \pi_{jm}^{\text{capcon}}(H_j, H_{-j})$$

where $\pi_{jm}^{\text{capcon}}(H_j, H_{-j})$ is calculated using the parameter estimates in Table 6 and $\pi_{jm}^{\text{nocapcon}}(H_j, H_{-j})$ sets the values for capacity constrained hospitals to zero. I adjust $S_{jm}(H_j, H_{-j})$ and $N_{jm}(H_j, H_{-j})$ for the reallocation of patients across hospitals when the capacity constraints are removed. The change in hospital profits is calculated similarly. The reported results for plan profits include all plans in the market; the hospital calculation includes just the capacity constrained hospital.

The results are set out in Table 19. The first row gives the median effect of investment to remove all hospital capacity constraints; the effects for three specific hospitals are then listed separately. The results are most easily understood by considering the

---

\(^{137}\) I assume that consumers do not have perfect foresight: they choose their plan under the belief that they can access any hospital on its list. The utility reduction from capacity constraints is caused when consumers realize a lower utility from the network than they expected. I multiply by $\frac{1}{\lambda_i}$ to convert from utils into dollars. Note that, in all three specific examples, only one hospital in the market is predicted to be capacity constrained.
specific examples which are reasonably representative of the overall data. The first two hospitals considered, St. Luke’s in Milwaukee and SW Texas Methodist Hospital in San Antonio, are large providers with significant high-tech services: each has over 700 beds. In both cases the model predicts that the removal of capacity constraints would have a substantial positive effect on consumer surplus: increases of $0.16 and $0.36 per person per new bed per year respectively. These figures translate to benefits of $232,000 and $469,000 per additional bed per market per year. The effects on producer surplus are smaller. The loss in hospital markups from reduced bargaining power outweighs the increased revenue from new patients and implies that both hospitals would lose money from the change (even if it involved zero investment): the losses are $18,618 and $25,780 per year per additional bed respectively. In the case of St. Luke’s the change would also reduce plan profits: the hospital has higher costs than others in the market, which are borne by the plan. Plans in San Antonio would see a positive but small profit increase of $19,000 per bed per year. The third example is somewhat different and is representative of a second type of hospital that the model predicts to be capacity constrained. South Austin hospital is a smaller suburban provider with fewer high-tech services, just 182 beds, and lower costs of care than its competitors. In this case the consumer surplus estimates translate to a $326,000 benefit to consumers per year per additional bed. The hospital’s low costs imply that plan profits would increase by a much higher $61,000 per year per new bed and hospital profits fall by just $5,300 per year per bed if the capacity constraints were removed.
## Table 19: Investment Incentives for Capacity Constrained Hospitals

<table>
<thead>
<tr>
<th>Example</th>
<th>CS per person ($ per bed per year)</th>
<th>CS per market ($ per bed per year)</th>
<th>Plan profit ($ per bed per year)</th>
<th>Hospital profit ($ per bed per year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>$0.20</td>
<td>$338,800</td>
<td>$19,728</td>
<td>- $19,797</td>
</tr>
<tr>
<td>St. Luke’s Milwaukee</td>
<td>$0.16</td>
<td>$232,270</td>
<td>- $2,779</td>
<td>- $18,618</td>
</tr>
<tr>
<td>SW Texas Methodist</td>
<td>$0.36</td>
<td>$469,000</td>
<td>$19,000</td>
<td>- $25,780</td>
</tr>
<tr>
<td>South Austin Hospital</td>
<td>$0.39</td>
<td>$326,400</td>
<td>$60,838</td>
<td>- $5,308</td>
</tr>
</tbody>
</table>

Notes: The first row gives the median effect (across markets) of investment to remove all hospital capacity constraints. For some markets this involves investment in more than one hospital. Rows 2-4 list the effects for three specific hospitals. All effects are given in $ per new bed per year. Plan profit effects are listed as a sum over plans in the market; hospital profits are given for the capacity constrained hospital alone.

Overall, then, the results have three implications. First, the benefit to consumers of removing hospital capacity constraints is large: a median benefit of over $330,000 per year for each new bed provided. The available data on the average cost of new hospital capacity implies a payback period of less than two years when the impact on consumers is taken into account\(^{138,139}\). Second, plan profits increase as a result of the  

---

\(^{138}\)Discussions with the COOs of several hospitals in the markets considered imply a capital cost of approximately $350,000 per new bed, assuming that a new wing is needed to house the new capacity, and staffing costs of around $65,000 per bed per year.

\(^{139}\)I conduct a robustness test for the consumer surplus results by recalculating the effect of the investment on consumers after setting the Type-1 extreme value error term in the demand equation.
change even if premiums are assumed fixed. Any premium increases would further increase the benefit to plans: for example, an increase of 10 cents per person per year for each new bed would generate a median total gain in plan revenues (across all plans in the market) of around $165,000 per new bed per year\textsuperscript{140}. Finally, however, in most cases hospitals (the organizations that are actually required to make the investment) face negative incentives to invest in this new capacity\textsuperscript{141}.

3.10 Discussion and Conclusion

The analyses in this chapter demonstrate some of the causes of the observed hospital-insurer networks at the firm level. Four factors are important: consumer demand for a choice of hospitals; hospital costs of care; the extent to which hospitals can credibly threaten to turn down low price offers and therefore push prices up; and the existence of hospital systems. Together these rationalize the majority of the observed contracts.

The full model for estimation sets few limits on the structure of the negotiation: I simply assume that firms bargain over a price that has both fixed and variable components and impose Nash equilibrium conditions to estimate the parameters. Previous papers have modeled the negotiation in more detail than the very simple framework used here but have not accounted for some of the hospital characteristics that this

\textsuperscript{140}The benefit to each individual plan is much smaller, however: an average of $14,000 per new bed per year if premiums increased by 10 cents per person. This small benefit, which implies a more than 20 year payback period for plans, explains why the prospect of profitable investment in new capacity does not in general lead to vertical mergers.

\textsuperscript{141}Of course these hospitals do not have negative incentives to invest overall: in fact they would benefit from increasing their number of beds as long as they remained capacity constrained. The results imply a disincentive to invest in new beds beyond this point, since the hospital’s threat to turn down low price offers then ceases to be credible. This implies an incentive to under-invest in capacity for three reasons: because the credibility of the threat increases with the extent of the capacity constraint; because hospitals are probably unable to predict future demand accurately; and because they prefer not to have enough beds to handle positive shocks to demand.
chapter shows are important. The results given here can therefore help determine which bargaining models best describe the hospital-health plan price negotiation; that is an additional contribution of this chapter.

The estimates also relate to a fairly substantial literature on the effectiveness of HMOs and POS plans in controlling costs. The original rationale for managed care was that the threat of selective contracting could be used as a lever to prevent hospitals demanding high prices. A number of recent papers have set out interview and other evidence suggesting that health plans’ leverage has declined in recent years prompting them to move away from selective contracting towards offering more choice\textsuperscript{142}. The major causes of the reduced leverage suggested by these papers are a rising consumer demand for choice and an extensive consolidation of hospitals resulting in increased provider market power. Capacity constraints are also mentioned as a source of hospital leverage. The evidence set out in Chapter 2 supports the first hypothesis: consumers do have a significant preference for choice. If this has developed recently, in response to experience of the restrictions imposed by managed care, it explains some of the move away from selective contracts. The results of this chapter are among the first to support the other two hypotheses. Without access to data on actual prices paid it is impossible to know whether the reduced form function estimated here has changed as a result of plans’ selective contracting; that is, whether high-priced hospitals would demand even more if no plans turned them down. However, I do show that hospitals in systems and those that expect to be full are the most often excluded when the surplus they generate is positive, consistent with the theory that they have the highest leverage. Further research would be useful, particularly in a setting where price data is available, to investigate these issues in more detail.

The final implication of this chapter relates to welfare. In Chapter 2 I estimated that a move from the observed hospital networks to an unselective outcome would

\textsuperscript{142}See for example Lesser and Ginsburg (2001) and Lesser, Ginsburg and Davis (2003).
lead to a total gain in consumer surplus of $1.04 billion per year and a loss to producers of just $0.80 million per year across the markets in the sample, assuming fixed premiums. This chapter demonstrates another important inefficiency caused by the contractual process: the high profits captured by capacity constrained hospitals imply an incentive to under-invest in capacity that translates to a significant loss to consumers. Plan incentives are somewhat better aligned with consumer preferences, but are less relevant given that hospitals, not insurers, are the organizations required to make the investment. While the analysis involves a number of assumptions that limit the accuracy with which these welfare effects can be measured, the distortion to provider incentives is clear. The results therefore demonstrate one avenue by which providers may modify their characteristics in order to increase their bargaining power. This subject merits attention from both policy-makers and researchers.

3.11 Appendix: A Model for the Price Negotiation

Consider a market in which one hospital, H1, offers services to two health plans, P1 and P2. Both plans offer preventive services through primary care physicians; those at P2 have higher quality than those at P1. The market contains three types of consumers. Half the population places a positive value on P2’s preventive services and the other half does not; of the consumers who value P2, half have a high valuation for H1 and half a low valuation\textsuperscript{143}. The three types of consumer are therefore defined as follows:

\textsuperscript{143}Extending the model to allow non-equal proportions of different consumer types does not affect the qualitative results: no new effects are generated. The assumption that consumer tastes for health plan services (separate from their hospital networks) differ across plans is intuitive. Women aged over 55, for example, are likely to have a higher valuation than other consumers for plans with high breast cancer screening rates.
2N of type C1: WTP $W_1$ for treatment at H1 and 0 for P2’s preventive services

$N$ of type C2: WTP $W_2$ for treatment at H1 and $V$ for P2’s preventive services

$N$ of type C3: WTP $W_3$ for treatment at H1 and $V$ for P2’s preventive services

Assume that $W_2 > W_1 > W_3$ and that $W_3 + V > W_1$. H1 has $3N$ beds and zero costs of care. Firms play the three-stage game described in Section 3.5. I assume that the timing of sickness is random. Therefore, if H1 agrees contracts with both plans and all consumers are insured, each consumer has probability $\frac{3}{4}$ of being treated at H1.

Finding the Nash Equilibrium of the game involves deriving each plan’s payoff in each of the four possible network combinations. The case where neither plan contracts with H1 is straightforward: $\pi_{P1} = 0$ and $\pi_{P2} = 2NV$ (since P2 charges $\text{prem}_2 = V$). If only P2 contracts with H1, the plan predicts that H1 will accept any price that covers its costs; P2 therefore pays H1 0 per treatment. Algebra implies that firm profits are $\pi_{P1} = 0$ and $\pi_{P2} = \max(3NW_1, 2N(W_3 + V), N(W_2 + V))$.

In the case where only P1 contracts with H1, P1 pays $p_1 = 0$ per treatment. For simplicity I assume parameter values such that P1 chooses to charge premium $W_1$ and P2 lets the C2 types switch to P1 rather than cutting its premiums in an attempt to keep them. (That is, I choose parameter values to rule out price wars.) This assumption requires, first, that if P1 sets $\text{prem}_1 = W_1$ (offering C2 types utility $W_2 - W_1$), P2 chooses not to cut premium to the point where $V - \text{prem}_2 = W_2 - W_1$.

The first assumption is therefore that $NV \geq 2N(V + W_1 - W_2)$, that is that:

$$2(W_2 - W_1) \geq V \quad (26)$$

The second requirement is that, given P2’s unaggressive reaction, P1 will choose to set premium $W_1$ and attract both C1 and C2 types, rather than setting $\text{prem}_1 = W_2$.
and hoping to attract the C2 types only. The assumption is therefore that:

\[ 3W_1 \geq W_2 \]  

Given these assumptions, firm payoffs are \( \pi_{p_1} = 3NW_1 \) and \( \pi_{p_2} = NV \).

Finally, consider the case where both plans make offers to H1. There are two scenarios under which H1 will choose to accept just one plan’s offer, implying an exclusive outcome. These are:

1. P1 has a higher WTP than P2, and if it wins, will choose to set \( \text{prem}_1 = W_1 \) so that both C1 and C2 types enroll in P1. H1 can then be certain of filling its beds even if it contracts exclusively with P1: if P1 makes a slightly higher offer than P2, H1 will accept just the higher offer.

2. P2 has a higher WTP than P1, and if it wins, will set \( \text{prem}_2 \) at a level to attract all three consumer types. Again, H1 will be certain of filling its beds by contracting exclusively with P2, and will do so if P2 makes a slightly higher offer than P1.

I consider here the parameter values needed to generate the more interesting case 1 since this potentially leads to an inefficient outcome. (Under case 2, all three types of consumer become enrolled in P2, implying that none had to make a suboptimal choice of plan to gain access to H1.)

We can determine whether P1 will choose to charge \( W_1 \), given any possible price \( p_1 \) paid to H1, by the following argument. If P1 pays price \( p_1 \) per treatment for an exclusive contract, it will choose to charge \( \text{prem}_1 = W_1 \) and attract C1 and C2

\footnote{It will never try to attract C3 types since in that case the hospital will be capacity constrained; the best it could do would be to set \( \text{prem}_1 = \frac{3}{4} W_3 \) and hope to earn \( 3NW_3 < 3NW_1 \).}

p1 paid to H1, by the following argument. If P1 pays price \( p_1 \) per treatment for an exclusive contract, it will choose to charge \( \text{prem}_1 = W_1 \) and attract C1 and C2
types if its profits from this option are greater than those from its alternatives\textsuperscript{145}. Algebra implies that the choice will be optimal provided $p_1 < \frac{1}{2}(3W_1 - W_2)$. So if $p_2 < p_1 < \frac{1}{2}(3W_1 - W_2)$ H1 will accept P1’s bid and reject that from P2.

P2’s maximum price can be calculated by considering its alternatives. If P1 wins an exclusive contract, P2 will charge $\text{prem}_2 = V$ and sell only to $C_3$ types, earning $\pi_{P2} = NV$. Conversely, P2 will win an exclusive contract if it offers a price higher than P1’s and chooses to set $\text{prem}_2 = \frac{3}{4}W_1$ (attracting all three consumer types)\textsuperscript{146}. P2’s maximum WTP for an exclusive contract can be found by equating profits under the win and lose scenarios and requiring a price such that P2 chooses $\text{prem}_2 = \frac{3}{4}W_1$:

\[ \text{algebra implies a value no higher than } p_2 = \frac{1}{2}(3W_1 - W_2 - V). \]

Finally, H1 may agree contracts with both plans\textsuperscript{147}. In this case P1 would set $\text{prem}_1 = \frac{3}{4}W_1$, attracting just $C_1$ types. P2 would set $\text{prem}_2 = \frac{3}{4}W_3 + V$, attract both $C_2$ and $C_3$ types, and earn profits $\frac{3}{2}N(W_3 - p) + 2NV$. Equating this with P2’s profits under the lose scenario implies that P2’s maximum WTP for a nonexclusive contract is $p_2 = W_3 + \frac{2}{3}V$\textsuperscript{148}.

P2’s maximum WTP for a contract with H1 is therefore given by $\tilde{p}_2 = \max(\frac{1}{2}(3W_1 - W_2 - V), W_3 + \frac{3}{4}V)$. The selective outcome will result provided that: $\tilde{p}_2 < p_1 < \frac{1}{2}(3W_1 - W_2)$. This is possible if the following inequality holds:

\[ W_3 + \frac{2}{3}V < \frac{1}{2}(3W_1 - W_2) \tag{28} \]

\textsuperscript{145}These are to charge $W_2$ and hope to attract just $C_2$ types or to charge $\frac{3}{4}W_3$ and hope to attract all consumers.

\textsuperscript{146}Its other options are to set $\text{prem}_2 = W_2 + V$ or $W_3 + V$ and attract just $C_2$-types or just $C_2$ and $C_3$ types respectively.

\textsuperscript{147}This will be the outcome if the two plans offer equal prices.

\textsuperscript{148}Neither plan will choose to start a price war under the nonexclusive contracts provided both plans pay H1 a price $p$ satisfying: $p > \max(2W_3 - W_1, 2W_1 - W_3 - \frac{2}{3}V)$. This inequality is satisfied at $p = W_3 + \frac{2}{3}V$. It is also possible that P1 would choose to charge premium $\frac{3}{4}W_1 + V$ and attract just $C_2$ types, leaving $C_3$ types uninsured. However, this is not the case for the parameters considered here.
The term on the left of this inequality represents P2’s WTP for a contract; the term on the right is P1’s incentive to sell to C1 as well as C2 types if it wins an exclusive contract with H1. The intuition is that P1 will secure the exclusive agreement provided P2 is willing to give up the contract and H1 expects P1’s enrollees to fill its beds.

If equation (28) holds the outcome is therefore that H1 agrees an exclusive contract with P1; C1 and C2 types enroll in P1 and C3 types enroll in P2. Plan profits are
\[ \pi_{P_1} = 3N(W_1 - \bar{p}_2 - \varepsilon) \] and \[ \pi_{P_2} = NV. \]

Can we assume that equations (26)-(28) can hold for reasonable parameter values? Simple examples demonstrate that this is possible: for example parameter values \( W_1 = 100, W_2 = 120, W_3 = 60 \) and \( V = 40 \) satisfy all three equations\(^{149}\). In these cases plan profits under the different network combinations (ignoring \( \varepsilon \) terms) are given by:

<table>
<thead>
<tr>
<th></th>
<th>P2 without H1</th>
<th>P2 with H1</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 without H1</td>
<td>0; 2NV</td>
<td>0; max(3NW_1, 2N(W_3 + V), N(W_2 + V))</td>
</tr>
<tr>
<td>P1 with H1</td>
<td>3NW_1; NV</td>
<td>min(3N(W_1 - W_3) - 2NV, ( \frac{3}{2}N(W_2 - W_1 + V) ); NV</td>
</tr>
</tbody>
</table>

The potential Nash equilibria are (H1,0) and (H1, H1). In both cases the equilibrium outcome is selective: P1 contracts with H1 but P2 does not\(^ {150} \).

Now consider the constrained efficient outcome. Given the institutional restriction that the hospital is committed, once it signs a contract, to treating patients in a

\(^{149}\)It is also easily confirmed that P1 earns positive profits in the (H1, H1) case for these parameter values.

\(^{150}\)In the second case, both plans bid for the hospital but only P1’s bid is accepted. Two types of equilibrium are possible for other parameter values. First, if \( W_2 \) is sufficiently high compared to \( W_1 \), P1 would choose to set premiums to attract solely C2 types if it won an exclusive contract. In that case the hospital will choose to contract with both plans. Second, if \( W_3 \) is high, P2 may have a higher WTP than P1, and may itself secure an exclusive contract.
random order, the social planner compares the following scenarios:

1. The selective equilibrium outcome, which results in social welfare $2NW_1 + NW_2 + NV$

2. The unselective outcome where both plans contract with H1, P2 enrolls $C_2$ and $C_3$ types and P1 sells just to $C_1$ types. Social welfare in this case is $\frac{3}{4}(2NW_1 + NW_2 + NW_3) + 2NV$.

The unselective outcome is therefore the constrained social optimum provided:

$$4V + 3W_3 > 2W_1 + W_2$$  \hspace{1cm} (29)

If this condition holds (as it does for the simple example given above), the benefit gained by $C_2$ types from enrollment in P2 outweighs the loss created when lower-value $C_3$ types displace other consumers from hospital beds. The inefficiency created by the bargaining process exists despite the fact that no capacity is wasted and that the consumers with the highest value for H1 are the ones treated. It is generated because H1 chooses which price offers to accept without internalizing the $C_2$ types’ benefit from enrollment in P2. A social planner takes this benefit into account when choosing the constrained efficient outcome.
Chapter 4: Moment Inequalities And Their Application

4.1 Introduction

This chapter is joint work with Ariel Pakes, Jack Porter and Joy Ishii. We consider the following problem. 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 alternatives 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 in-
teracting 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, and accept any value of the parameter vector 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 4.2 of this chapter provides conditions under which our vector of inequalities have positive expectation when evaluated at the true value of the parameter vector. Section 4.3 assumes this inequality condition and provides methods of inference for that parameter vector. Section 4.4 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 affect 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) classical measurement errors. We refer to unobservable determinants of profits that can affect the agent’s decisions as “structural disturbances” and assume that they 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 4.3 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. In particular, inequality tests are likely to be more robust to small deviations in modelling assumptions than tests of a point null hypothesis. Also we expect users to want to be able to test for the importance of allowing for structural disturbances. 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 analyzed 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 introduce 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 Chapter 3 and in 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 es-
timation subject to inequality restrictions. This subsection notes our debt to that
literature.

consider identification issues, derive properties of set estimators, and consider infer-
ential procedures for models with inequalities. Both these papers are primarily con-
cerned with the econometric issues surrounding a given set of inequalities, rather than
with how these inequalities might be obtained from an underlying model\textsuperscript{151}. 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.

\textsuperscript{151}Andrews, 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,
4.2 A Framework for the Analysis

We begin with a statement of the problem. Let \( d_{i,t} \) denote a decision made by agent \( i \) of type \( t \) for \( i = 1, \ldots, n_t \) and \( t \in T \). We index agents by types because some of our examples involve vertical relationships (e.g. buyer-seller networks) where a single observed outcome (e.g. a contract) requires decisions by two or more agents of different types. We assume \( d_{i,t} \) takes values in \( D_t \) for \( t \in T \).

When \( D_t \subset \mathcal{R} \) 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 \( d_{i,t} \) is vector valued then \( D_t \) is a subset of the appropriate product space\(^{152}\).

In addition there will be a set of variables \( y_{i,t} \in Y_t \) which are determinants of the returns (or the utility) the agent will make when its decision is \( d_{i,t} \) and other agents’ decisions are \( d_{-i,t} \) (this includes decisions of all other agents; those of the same type and those of other types). Not all components of \( y_{i,t} \) 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.

The agent’s returns conditional on \((d_{i,t}, d_{-i,t}, y_{i,t})\) are given by

\[
\pi_t(\cdot, \theta_0) : \Pi_{t \in T}[D_t^{n_t} \times Y_t] \rightarrow \mathcal{R}, \text{ for } t \in T,
\]

where \( \theta_0 \), the parameter of interest, is known to be contained in a compact subset of \( \mathcal{R}^K \). We let \( \mathcal{E} \) be the expectation operator, and \( \mathcal{J}_{i,t} \) be the agent’s information set at the time it makes its decisions.

\(^{152}\)For example \( D_t \) 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 contracts which ensure a single vendor per region, \( D_t \) becomes a proper subset of the product space of all possible two part tariffs. The reader who is interested in single agent problems need only set \( t = n_t = 1 \) and ignore both \( d_{-i,t} \) and the \( t \) subscripts in what follows.
Assumption 1 is our behavioral assumption.

**Assumption 1**

\[
\sup_{d \in D_t(d_{i,t})} E[\pi_t(d, d_{-i,t}, y_{i,t}, \theta_0)|J_{i,t}] \leq (1 + \delta)E[\pi_t(d_{i,t}, d_{-i,t}, y_{i,t}, \theta_0)|J_{i,t}]
\]

where \(D_t(d_{i,t}) \subset D_t\), for \(i = 1, \ldots, n_t\), and \(t \in T\).

In the special case where \(\delta = 0\) and \(D_t(d_{i,t}) = D_t\), Assumption 1 is typically derived from optimizing behavior. If \(n = t = 1\) and \(D_t\) is a finite set then Assumption 1 is an implication of a standard discrete choice problem. If \(D_t\) 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 interacting agents these optimality conditions are necessary for a Nash equilibrium. Note that in this case there may be different combinations of the \(\{d_{i,t}\}\) of the various agents that satisfy our conditions; i.e. there may be multiple equilibria. Throughout we will only require the necessary conditions that any of these Nash equilibria must satisfy.

By allowing for \(\delta > 0\) and \(D_t(d_{i,t}) \neq D_t\) we can allow for various types of non-optimizing behavior. When \(\delta > 0\) and \(D_t(d_{i,t}) = D_t\), Assumption 1 allows agents to choose non-optimal strategies, as long as their consequences are not “too sub-optimal”. For example, if \(\delta = .5\) non-optimal choices would be allowed provided they did not, on average, reduce profits more than 50% from the profits that would be earned from optimal strategies. Complementary reasoning applies to the actions per se when \(\delta = 0\) but \(D_t(d_{i,t}) \neq D_t\). For example, if there was a continuous control, and we specified that \(D_t(d_{i,t}) = \{d : |d - d_{i,t}| \geq \alpha d_{i,t}, d \in D_t\}\) for some \(\alpha > 0\), then we would be specifying that though small deviations about optimal behavior can occur (deviations that leave the choice within \(\alpha\%\) of the optimal decision), at least on average large deviations do not occur. For now assume \(\delta\) and \(D_t(d_{i,t})\) are set
4.2.1 Functional Form and Orthogonality Conditions

If $(d, d') \in D_t^2$ are two different feasible choices of a type $t$ agent, let

$$\Delta \pi_t(d, d', d_{-i,t}, y_{i,t}, \theta_0) \equiv (1 + \delta) \pi_t(d, d_{-i,t}, y_{i,t}, \theta_0) - \pi_t(d', d_{-i,t}, y_{i,t}, \theta_0).$$

Then our functional form and stochastic assumptions are given by Assumption 2.

Assumption 2

For $i=1,...,n_t$, and $t \in T$, assume $y_{i,t}$, the vector of variables which determine profits conditional on the controls, has a component, labelled $y_{i,t} \in Y_t$, that is observed by the econometrician, and two unobserved components (or disturbances), labelled $(\nu_{1,i,t}, \nu_{2,i,t})$, which are defined as follows. For each $(d, d') \in D_t \times D_t$, let

$$\Delta \pi_t(d, d', d_{-i,t}, y_{i,t}, \theta_0) = \mathbb{E}[\Delta \pi_t(d, d', d_{-i,t}, y_{i,t}, \theta_0) | \mathcal{J}_{i,t}] + \nu_{1,i,t}$$

so that

$$\mathbb{E}[\nu_{1,i,t} | \mathcal{J}_{i,t}] = 0.$$

In addition, we allow expected returns of agent $i$ of type $t$ to differ from the expectation of a known observable function, denoted by $\Delta r_t(\cdot)$, by an unobserved random variable, $\nu_{2,i,t} \in \mathcal{J}_{i,t}$, that is mean independent of $x_{it} \in \mathcal{J}_{i,t}$. That is, for each

An alternative would be to begin the analysis assuming optimal behavior, i.e. assuming $\delta = 0$ and $D(d_i) = D$, 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 $\delta$ 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 forms, the relaxation of optimizing behavior needed to rationalize the data is too large to be a priori reasonable.
\[(d, d') \in D_t \times D_t,\]

\[\mathcal{E}[\Delta \pi_t(d, d', d_{-i,t}, y_{i,t}, \theta_0)|J_{i,t}] = \mathcal{E}[\Delta r_t(d, d', d_{-i,t}, y_{i,t}, \theta_0)|J_{i,t}] + \nu_{2,i,t} f_t(x_{i,t}), \quad (31)\]

for some (possibly unknown) function \(f_t(\cdot),\) and

\[\mathcal{E}[\nu_{2,i,t}|x_{i,t}] = 0, \quad \text{for an observable } x_{i,t} \subset J_{i,t}. \quad (32)\]

Note that equations (30) and (31) imply that \(\nu_{1,i,t}\) and \(\nu_{2,i,t}\) will generally depend on \((d, d')\) (though that dependence is suppressed in the notation).

Note also that Assumption 2 puts conditions on the difference in profits, but not on the level of profits per se. As a result in the optimizing examples (i.e. when \(\delta = 0\)), unobserved factors that have the same effect on the agents profitability regardless of the \(d\) chosen (i.e. agent-specific “fixed effects” that are additive in the level of profits) are differenced out of the specification. So the unobservables (or \(\nu = (\nu_1, \nu_2)\)), need only capture the effects of omitted variables that affect how profits change when \(d\) changes.

The functions \(r_t(\cdot, \theta)\) and \(f_t(\cdot)\) depend only on observables. \(r_t(\cdot, \theta)\) provides the measured incremental returns minus costs of the change from \(d\) to \(d'\) (at least up to a parameter vector to be estimated). In contrast, \(\nu_{1,i,t}\) and \(\nu_{2,i,t}\) are determinants of profits that are not directly observed by the econometrician.

The distinction between \(\nu_1\) and \(\nu_2\) is of some importance. From equation (30), \(\nu_{1,i,t}\) is mean independent of \(J_{i,t}\). Since the actual decision (or \(d_{i,t}\)) is measurable \(J_{i,t}\), it follows that \(\nu_{1,i,t}\) is mean independent of \(d_{i,t}\). In contrast \(\nu_{2,i,t}\) is contained in \(J_{i,t}\), so \(\mathcal{E}[\nu_{2,i,t}|d_{i,t}]\) is not likely to be zero.

The importance of accounting for one or both of \((\nu_1, \nu_2)\) will vary with the problem being studied. Differences between \(\nu_1\) and zero do not change the agent’s expected
profits at the time decisions are made, so $\nu_1$ realizations can be caused by either expectational errors, or measurement errors. Note that there are two sources of expectational errors; (i) environmental variables that evolve exogenously over time and are unknown to the agents at the time they make their decision, and (ii) asymmetric information, or any component of the $y$’s or the $\nu_2$’s of the firm’s competitors that is not known to the firm when it makes its decision (this causes uncertainty in the realization of $d_{-i,t}$).

In contrast $\nu_2$ is a “structural” disturbance, i.e. a source of variance in the difference in profits that the agent conditions its decisions on, but the econometrician does not observe. We assume $\nu_2$ enters the difference in returns linearly and that it is mean independent of a subset of the variables that the agents know when they make their decisions, a subset which will become our “instruments"\textsuperscript{154}. Variation in $\nu_2$ will be important when the econometrician’s profit measure does not account for determinants of the variance in profits that affect agent’s decisions.

A few comments on other aspects of Assumption 2 are in order. As noted the $y$ must be observed by the econometrician but may include random variables whose values are not known to the agents when they make their decisions. Accordingly the distribution of the $y_{i,t}$ can depend both on $d_{i,t}$ and $d_{-i,t}$. This enables us to analyze sequential games in which the profits generated by a particular decision depends on a subsequent stage game which conditions on all prior decisions, or on $\{d_{i,t}\}_{i,t}$. For example after the firm’s know the outcome of their contracting or investment decisions they might engage in a pricing game which determines profits. Note that when this occurs the computation of the profits from the alternative choices (i.e. from $d' \neq d_{i,t}$) must allow for the induced changes in the subsequent controls; see the empirical

\textsuperscript{154}At the cost of complicating the notation we could weaken these assumptions for various special cases. In particular there are cases in which; mean independence can be replaced by a weaker zero covariance assumption, the linearity assumption can be replaced with simulation or numerical integration techniques, and the $x_{i,t}$ need only be known by the $i^{th}$ agent.
examples below for details.

Note also that when we deal with the multiple agent case we are making the implicit assumption that the \( \nu_2 \) of the firm’s competitors only affects the agent’s profits through their effects on \((y_{i,t},d_{i,t},d_{-i,t})\). On the other hand we have not had to specify whether the \( \nu_{2,-i,t} \) is in the agent \( i \)’s information set at the time decisions are made. In particular \( \mathcal{J}_{i,t} \) could contain the values of the \( \nu_{2,-i,t} \), contain a signal on their likely values, or not contain any information on their values at all. Relatedly, we are not making a particular assumption on the relationship of the \( \{\nu_{2,i,t}\} \) draws of the different agents.\(^{155}\)

So far these assumptions are not very stringent. In particular we have not specified a particular form for the distribution of \( \nu_1 \) or \( \nu_2 \). Moreover, and perhaps more importantly, the mean independence assumption need not apply to all the variables that affect the returns from a choice. So the model allows for discrete choice sets and endogenous regressors, and does not need to specify a parametric form for the disturbance distribution. We will, however, also require one of the two assumptions described in our Condition 1.

The first assumes we can form a linear combinations of profitability differences which; (i) has components which have positive expectation (from Assumption 1), and (ii) does not depend on the structural errors (on our \( \{\nu_2\} \)). Assumption 2 will then ensure that the linear combination will have positive expectation conditional on \( x \) at \( \theta = \theta_0 \) provided \( x \) is in the information sets of all the agents whose profitability differences enter the linear combination.

The second case requires an ability to form unconditional moments, or more pre-

\(^{155}\) Different papers in the Industrial Organization literature have used one or the other of the assumptions regarding whether or not \( \nu_{2,-i,t} \in \mathcal{J} \). Bresnahan and Reiss, 1991, Berry, 1992, assume full information; Seim 2002 and Pakes Ostrovsky and Berry 2003, assume knowledge of \( \nu_{2,i,t} \) but not of \( \nu_{2,-i,t} \); and Fershtman and Pakes 2004, allow for signals. Of course if we knew (or were willing to assume) more on the properties of the \( \nu_{2,i,t} \) we might well be able to provide more precise estimators of \( \theta \) (see, for example, Bajari, Hong and Ryan, 2004).
cisely moments that do not depend on the decisions any of the agents made; i.e. on \( d \equiv \{d_{i,t}\}_{i,t} \in D \equiv \Pi_{t \in T} \mathcal{D}^{\alpha_t} \). It assumes that no matter \( d \) there is a linear combination of profitability differences which; (i) has components which have positive expectation (from Assumption 1), and (ii) is linear in the structural errors (our \( \nu_2 \)). Assumption 2 will then ensure that this linear combination of profitability differences will have positive expectation conditional on \( x \) at \( \theta = \theta_0 \) provided; (i) \( x \) is in the information set of all the agents whose profitability differences are included in the linear combination, and (ii) \( x \) satisfies equation (32) of Assumption 2 (it is an instrument). To make this precise we assume there is a mapping from \((i, t, d, d')\) to an index \( q \) such that \( \nu_{2,i,t}(d, d') = \nu_{2,q} \) for some \( q \in Q \), and we can form “unconditional” sample averages of the \( \{\nu_{2,q}\} \).

**Condition 1**

Assume either that:

(i) for a subset of the data that are selected by a function of their \( J_{i,t} \), there is a linear combination of profitability differences from the observed choices that does not depend on \( \nu_{2,i,t} \), i.e. there is a set of weights \( \{\chi_{i,t}(k, l, d_{k,l}, d')\} \) with \( d' \in \mathcal{D}(d_{k,l}) \), such that

\[
\sum_{k,l,d'} \chi_{i,t}(k, l, d_{k,l}, d') \Delta \pi_{l}(d_{k,l}, d', d_{-k,l}, y_{k,l}, \theta_0) = \]

\[
\sum_{k,l,d'} \chi_{i,t}(k, l, d_{k,l}, d') \mathbb{E}[\Delta r_{l}(d_{k,l}, d', d_{-k,l}, y_{k,l}, \theta_0) | J_{k,l}] + \sum_{k,l,d'} \chi_{i,t}(k, l, d_{k,l}, d') \nu_{1,k,l}
\]

and there is an unobservable

\[
x \in \cap_{(k,l) \in j \cap \mathcal{J}_{k,l}} \mathcal{I}_{k,l}, \]

where

\[
\mathcal{I}_{i,t} = \{(k, l) \in \mathcal{J}_{i,t} \cap \mathcal{J}_{k,l} : \sum_{d'} |\chi_{i,t}(k, l, d_{k,l}, d')| \neq 0\};
\]
(ii) there is a set \( \{ \nu_{2,q} \}_{q \in Q} \), such that no matter \((i,t,d,d')\), \(\nu_{2,i,t} = \nu_{2,q}\) for some \(q\), and there is a known linear combination of differences from observed profitability that is linear in \(\nu_{2,q}\), i.e. for any realization of \(d \in D\) there is a set of weights \(\{ \chi_q(k,l,d_{k,l},d') \} \) with \(d' \in D(d_{k,l})\), such that

\[
\sum_{k,l,d'} \chi_q(k,l,d_{k,l},d') \Delta \pi_1(d_{k,l},d',d_{-k,l},y_{k,l},\theta_0) = \quad (35)
\]

\[
\sum_{k,l,d'} \chi_q(k,l,d_{k,l},d') \mathbb{E}[\Delta r_l(d_{k,l},d',d_{-k,l},y_{k,l},\theta_0) | J_{k,l}] + \sum_{k,l,d'} \chi_q(k,l,d_{k,l},d') \nu_{1,k,l} + g_q(x) \nu_{2,q},
\]

for a (possibly unknown) function \(g_k(.)\), and that

\[
\mathcal{E}[\nu_{2,q}|x] = 0, \quad (36)
\]

for an observable \(x\) that satisfies equation 34 with the index \(q\) substituting for the index \((i,t)\) in that equation.

We now show how some familiar models generate profitability differences of the form in equation (33) or (35) (some will require us to form linear combinations of differences in profits of more than one agent). Before doing so, however, we want to stress two points. First Assumptions 1 and 2 are not, in general, sufficient for our Condition 1, and hence that condition will have to be checked separately in each case studied. Second Condition 1 is part of a set of sufficient, but not necessary, conditions for the properties of the estimators we shall consider. For example if a subset of \(Q\) satisfies (35) all that we will require is that the subset not depend on the realization of \(d\).

Here are some familiar examples that satisfy Condition 1.

- \(\pi(.)\) 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 \( f_t(\cdot) \) in Assumption 2 is identically zero (equivalently \( \nu_{2,i,t} = 0 \)). Then the difference between the profits the \((i,t)^{th}\) agent did earn and those it would have earned had it chosen any \( d' \in \mathcal{D}_t(d_{i,t}) \) satisfies part (i) of Condition 1. I.e. to form equation (33); pick a \( d' \in \mathcal{D}(d_{i,t}) \), set \( \chi_{i,t}(i,t,d_{i,t},d') = 1 \), and set \( \chi_{i,t}(k,l,d_{k,t},d') = 0 \) elsewhere. This generates

\[
\Delta \pi_t(d_{i,t}, d', -d_{i,t}, y_{i,t}, \theta_0) = E[\Delta r_t(d_{i,t}, d', -d_{i,t}, y_{i,t}, \theta_0) | \mathcal{F}_{i,t}] + \nu_{1,i,t}.
\]

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 for agents whose choices are not always exactly optimal conditional on a prespecified information set (\( \delta \) need not be zero, and \( \mathcal{D}(d) \) need not equal \( \mathcal{D} \)), and (iii) allows explicitly for interacting agents (clarifying the conditions that must hold in that case). 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). Auctions are an example where this is often easy to do, but these assumptions are relevant for any problem where we can measure profits up to a mean zero error. As a result they constitute a special case we might often want to test for.

- **Agent specific fixed effects.** If there are multiple decisions involving the same \( \nu_2 \) then we can satisfy part (i) of Condition 1 by forming moment conditions that difference out the \( \nu_2 \). 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.
For example, suppose agents (of a single type) make two simultaneous decisions
\(d = (d_1, d_2)\) with \(d_w \in \{0, 1\}\) for \(w = \{1, 2\}\) and that the structural error is the same across decisions. For simplicity assume also that the profit function is additive across choices\(^{156}\), so that \(\Delta \pi(d, d', \cdot) = \sum_w \Delta \pi_w(d_w, d'_w, \cdot)\), and

\[
\Delta \pi_w(d_{i,w}, d'_{i,w}, \cdot) = \mathcal{E}[\Delta r_w(d_{i,w}, d'_{i,w}, \cdot)| \mathcal{J}_i] + (d_{i,w} - d'_{i,w})\nu_{2,i} + \nu_{1,w,i},
\]

for \(w = \{1, 2\}\). To obtain equation (33) in Condition 1, set \(\chi_i(k, d_k, d') = 0\) whenever \(k \neq i\) or \(k = i\) and \(d_i = (0, 0)\) or \(d_i = (1, 1)\). When \(d_i = (1, 0)\) set

\[
\chi_i(i, d_i, d') = 1 \text{ if } d' = (0, 0) \text{ or } d' = (1, 1),
\]

and zero elsewhere, while if \(d_i = (0, 1)\) set

\[
\chi_i(i, d_i, d') = 1 \text{ if } d' = (0, 0) \text{ or } d' = (1, 1),
\]

and zero elsewhere. Then if \(1\{\cdot\}\) is notation for the indicator function

\[
\sum_{k, d'} \chi_i(k, d_k, d') \Delta \pi(k, d', \cdot) = \sum_{d'} \chi_i(i, d_i, d') \left[ \Delta \pi_1(d_{1,i}, d'_{1,i}, \cdot) + \Delta \pi_2(d_{2,i}, d'_{2,i}, \cdot) \right]
\]

\[
= 1\{d_i = (1, 0)\} \mathcal{E} \left\{ [\Delta r_1(d_{1,i} = 1, d'_{1,i} = 0, \cdot) + \Delta r_2(d_{2,i} = 0, d'_{2,i} = 1, \cdot)] | \mathcal{J}_i \right\}
\]

\[
+ 1\{d_i = (0, 1)\} \mathcal{E} \left\{ [\Delta r_1(d_{1,i} = 0, d'_{1,i} = 1, \cdot) + \Delta r_2(d_{2,i} = 1, d'_{2,i} = 0, \cdot)] | \mathcal{J}_i \right\}
\]

\[+ \sum_{d'} \chi_i(i, d_i, d') \nu_{1,i},\]

and the last term is mean independent of any \(x \in \mathcal{J}_i\). Note that this approach allows estimation of panel data discrete choice models with fixed effects; and it

\(^{156}\)To analyze the nonadditive case simply assume that for any \((d, d')\), \(\Delta \pi(d, d', \cdot) = \mathcal{E}[\Delta r(d, d', \cdot)| \mathcal{J}_i] + [(d_1 - d'_1) + (d_2 - d'_2)]\nu_{2,i} + \nu_{1,i}.\) Note also that we could allow choice specific effects by including dummy variables for the choices in the profit specification.
does not require distributional assumptions on either the effects or the disturbances. Moreover similar ideas can be used in many different problems\textsuperscript{157}.

- **Ordered choice.** Consider any discrete choice with an order to the choice set and a determinant of the choice which is known to the agent but not to the econometrician. Lumpy investment decisions (say in the number of stores, or machines...) are one example of an ordered choice problem, and our first empirical example is a case in point. It has markets consisting of sets of interacting firms (again of a single type) each of whom decides how many units of a machine to purchase and install. The model allows for a component of marginal cost which varies across firms, is known by the firms when they make their decisions but not to the econometrician, and is mean independent of a set of observables. In this example \( d \) is a positive integer representing the number of machines built and installed, and \( \nu_{2,i} \) is the unobserved firm specific portion of that agent’s marginal cost schedule. Since \( \nu_{2,i} \) is individual (but not choice) specific, we have \( q = i \) and \#\( Q \) equal to the number of active firms. In the simplest case when marginal costs are constant the difference of profits at \( d \) and \( d' \) includes a cost difference equal to \((d - d')(\theta + \nu_{2,i})\). We use part (ii) of Condition 1 to form our moment inequality in this case and set \( \chi_i(k, \cdot) = 0 \) whenever \( k \neq i \). One example of equation (35) can be obtained by setting \( \chi_i(i, d_i, d') = 1 \) if \( d' = d_i - 1 \) and zero otherwise. This generates the difference between the expected incremental profits from the last machine bought and its marginal cost, or

\[
\Delta \pi(d_i, d_i - 1, \cdot) = \mathcal{E}[(\Delta r(d_i, d_i - 1, \cdot)|\mathcal{J}_i] - (\nu_{2,i} + \theta) + \nu_{1,i}.
\]

\textsuperscript{157}For example Pakes, 2004, considers the case of buyer/seller networks with buyer and seller specific fixed effects. In the network example the fixed effect assumptions imply more inequalities than difference in difference inequalities alone. In applications where there are repeated observations on exactly the same set of choices, one can use the repetitions in even richer ways; see Pakes, Porter, and Wolfram, in process, for an electric utility bidding example.
Assumption 1 implies that $\mathcal{E}[\Delta \pi(d_i, d_i - 1, \cdot)|\mathcal{J}_i] \geq 0$, which, from Assumption 2, generates $\mathcal{E}[\Delta r(d_i, d_i - 1, \cdot)|\mathcal{J}_i] \geq (\theta + \nu_{2,i})$. This gives us an upper bound on marginal cost for each firm (that can then be averaged across firms). Alternatively if we set $\chi_i(i, d_i, d') = 1$ only if $d' = d_i + 1$ we get the negative of the difference in expected incremental profits of the $d_i + 1^{th}$ machine and marginal cost. This will put a lower bound on marginal cost. Clearly other values of $d - d'$ could be used also\textsuperscript{158}.

- **Buyer-seller networks with unobserved transfers.** Here $\mathcal{T} = \{b, s\}$ for buyers and sellers respectively. Type $b$’s incremental cost is the cost of purchase, and its incremental expected returns are the expected profit from resale. Type $s$’s incremental returns are $b$’s purchase cost and its incremental costs are the costs of production. Assume that sellers make take it or leave it offers to all buyers. These offers are proprietary and hence not public information. So it is their properties that we want to investigate empirically. The offers are assumed to be a parametric function of observables (e.g. an unknown markup per unit purchased) and an error ($\nu_2$). For now assume there is only one seller and one buyer in each market studied. $\mathcal{D}_s$ is a set of contracts which we assume includes a null contract, say $d_s = \phi$, that is never accepted. $\mathcal{D}_b = \{0, 1\}$ with $d_b = 1$ indicating the contract was accepted. Note that any transfer cost to the buyer is a revenue for the seller, so $\nu_{2,s} \equiv -\nu_{2,b}$. As a result in the simple case with one buyer and one seller per market there is only one value of $\nu_2$ per market (so we omit the $q$ index from the notation), and it enters the profits of the buyer and the seller with opposite signs ($f_s = 1$ and $f_b = -1$ in our equation (31)).

Then Condition 1 can be satisfied in two ways (more details are provided in

\textsuperscript{158}The discussion of this case 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.). We consider ways of weakening this assumption in the discussion of our first empirical example below.
Section 4.4.2).

- Assumption 1 implies that (i) the expected profits to the seller conditional on acceptance of the offer by the buyer are positive, and (ii) if the buyer rejects the offer it is because profits without the contract are higher then profits with the contract. Let $\Delta \pi_s(\cdot)$ be the increase in seller profits if there is a contract. Then the actual seller’s profits from offering the contract $d_s$ is $\mathbb{1}\{d_b = 1\} \Delta \pi_s(\cdot)$. We now derive an inequality which satisfies part (ii) of Condition 1. There are two mutually exclusive and exhaustive possible outcomes; either there is, or there is not, a contract. To form equation (35) set; $\chi(s, d_s, d'_b = \phi) = 1$, $\chi(b, d_b = 0, d'_b = 1) = 1$, and the rest of the $\chi(\cdot)$ equal to zero. This generates the seller’s profits from contracting when the contract is accepted (which is additive in the transfer disturbance or $\nu_2$), and it picks up the savings of the buyer from not contracting when there is no contract (which is also additive in the disturbance from the transfers or $\nu_2$). So, no matter the $d$, the equation contains $\nu_2$ with a coefficient of one.

- Assumption 1 also implies that if $d_b = 1$ then the sum of the expected incremental profits of the two agents must be positive (there must be a positive surplus to divide between them). Since the unobserved transfers cancel out when computing this sum, this implies that we obtain an equation that satisfies part (i) of Conditions 1 or equation (33) if we set; $\chi(b, d_b = 1, d'_b = 0) = 1$, $\chi(s, d_s = 1, d'_s = \phi) = 1$, and the rest of the $\chi(\cdot) = 0$.

Our second empirical example is a generalization of this model. It allows $n_b > 1, n_s > 1$, and bilateral contracts between any subset of the possible buyer-seller couples (each contract can affect the transfers between all agents). Then
every possible buyer-seller combination will generate a structural error so \( \#Q = n_s \times n_b \).

4.2.2 Inequality Conditions

We now derive a set of moment inequalities that can be used for estimation and inference. Parts (i) and (ii) of Condition 1 differ slightly in notation; part (i) subscripts the weights by \((i, t)\) for the different observations for which equation (33) holds [i.e., the weights are \( \chi_{i,t} (\cdot) \)], and part (ii) indexes them by \( q \) for the distinct \( \nu_{2,q} \). For notational ease we will stick with the \( q \) index, but it is to be understood that we implicitly switch to the \((i, t)\) indexes when we use part (i) of Condition 1.

Take \( x \in \cap_{(k,l) \in \mathcal{J}_k,l} \mathcal{J}_k,l \) and let \( h : X \rightarrow \mathcal{R}^+ \) be a vector of positive functions of \( x \). Then provided each \( d' \in D_{k,l} \) Assumption 1 implies

\[
\mathcal{E} \left[ \mathcal{E} \left[ \sum_{k,l,d'} \chi_q(k,l,d_{k,l},d') \Delta \pi_t(d_{k,l},d',d_{-k,l},y_{k,l},\theta_0)|x] \otimes h(x) \right] \right] \geq 0.
\]

\( \Delta \pi(\cdot, \theta) \) is not observed, but \( \Delta r(\cdot, \theta) \) is. Moreover if either the conditions of part (i) or part (ii) of Condition 1 are satisfied, Assumption 2 implies

\[
\mathcal{E} \left[ \mathcal{E} \left[ \sum_{k,l,d'} \chi_q(k,l,d_{k,l},d') \Delta r_t(d_{k,l},d',d_{-k,l},y_{k,l},\theta_0) \otimes h(x) \right] \right] = \mathcal{E} \left[ \sum_{k,l,d'} \chi_q(k,l,d_{k,l},d') \Delta r_t(d_{k,l},d',d_{-k,l},y_{k,l},\theta_0) \otimes h(x) \right] \geq 0. \tag{37}
\]

Our estimator forms the sample analog of equation (37) and looks for values of \( \theta \) that satisfy this matrix of moment inequalities.
4.3 Estimation and Inference

We provide details for the case where there is data on $J$ markets indexed by $j = 1, \ldots, J$. A market is a draw on $(y_j, x_j, d^j)$ where $y_j \equiv \{y_{i,t}^j\}_{i=1:t \in T}$, 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 $P$, 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 (37) and finding the set of $\theta$, say $\Theta_J$, that minimize over $\theta \in \Theta$ (a compact subset of $\mathbb{R}^K$).

\[
\left\| \left( \frac{1}{J} \sum_{j=1}^J \frac{1}{\#Q} \sum_{q \in Q} \sum_{k,l,d'} \chi_q(k, l, d_{k,l}, d') \Delta r_j^\tau(d, d', d_{-k,l}^j, y_{k,l}^j, \theta) \otimes h(x_j) \right) \right\|\]  

(38)

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 $\theta$. That is, though $\Theta_J$ must contain at least a single point, it may well be equal to a larger set of points (more on this distinction below).

Let $\Theta_0$ denote the set of parameter values that satisfy equation (37). In the literature on estimation subject to inequality restrictions, $\Theta_0$ is often called the identified set. Under regularity conditions, one can show set-consistency of $\Theta_J$ for $\Theta_0$; see Andrews, Berry and Jin (2004) and Chernozhukov, Hong, and Tamer (2003).

For our general case we only consider the problem of constructing confidence intervals that asymptotically cover (functions of) the true parameter ($\theta_0$) 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 a more powerful test is presented 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.

4.3.1 Confidence Regions

In this subsection, 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 \( \theta \) that exceeds the moments when both are evaluated at \( \theta_0 \) 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

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

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 4.2.2, the assumptions imply that

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

Define the variance of the population moments to be

\[ \Sigma(\mathcal{P}, \theta) = Var(m(\cdot, \theta)) \]

and denote its sample analogue by

\[ \Sigma(P_J, \theta) = \frac{1}{J} \sum_{j=1}^{J} (m(y_j^i, d_j^i, x_j, \theta) - m(P_J, \theta))(m(y_j^i, d_j^i, x_j, \theta) - m(P_J, \theta))^\prime. \]

Finally note that with this notation the estimation problem in (38) 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

\[ \Theta^{CI} = \{\theta : m^*(P_J, \theta) \geq 0, \ \theta \in \Theta\}. \tag{39} \]

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

\[ m^*(P_J, \theta_0) \geq m(\mathcal{P}, \theta_0), \tag{40} \]

then \( \theta_0 \in \Theta^{CI} \). The confidence region is built by finding functions \( m^*(\cdot) \) such that
the sufficient condition (40) occurs with probability approaching (at least) 1 − α, and using them to build the $\Theta^{CI}$ in equation (39). The test simply asks whether $\Theta^{CI}$ is the empty set.

All we need to construct our $m^*()$ 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. More formally, we assume

**Assumption 3**

(a) $\sqrt{J}[m(P_j, \theta_0) - m(P, \theta_0)] \rightarrow_d N(0, \Sigma(P, \theta_0))$, and

(b) $\Sigma(P_j, \theta_0) \rightarrow_d \Sigma(P, \theta_0)$.

We now construct $m^*(\cdot)$. For a fixed $\theta$, suppose $Z^*(\theta) \sim N(0, \Sigma(P_j, \theta)).$ Find a vector $\tilde{z}_{J,\alpha}(\theta)$ that satisfies $\Pr(Z^*(\theta) \geq -\tilde{z}_{J,\alpha}(\theta)|P_J) = 1 - \alpha$. 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 $\tilde{z}_{J,\alpha}(\theta)$, set

$$m^*(P_j, \theta) = m(P_j, \theta) + \frac{1}{\sqrt{J}}\tilde{z}_{J,\alpha}(\theta).$$

(41)

and substitute this expression into the definition of $\Theta^{CI}$ in (39). Then $\Theta^{CI}$ is an asymptotic $(1 - \alpha)$ level confidence interval for $\theta_0$.

**Theorem 1** Let

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

\[159\text{Note that we do not require the weak convergence of the empirical process } \{m(P_j, \theta)\}_{\theta \in \Theta}. \text{ In particular, stochastic equicontinuity of the empirical process is not needed. Similarly, consistency of } \Sigma(P_j, \theta) \text{ is only required at the point } \theta_0. \text{ Of course, any other consistent estimator of the variance would suffice.}

\[160\text{Construction of } Z^*(\theta) \text{ is discussed at the end of section 4.3.2.}
and suppose Assumptions 2 and 3, and Condition 1 hold. Then

$$\lim_{J \to \infty} \Pr\{\theta_0 \in \tilde{\Theta}^{CI}\} \geq 1 - \alpha.$$ 

**Proof.** Define $m^*(P_J, \theta)$ as in (41),

$$\Pr(\theta_0 \in \tilde{\Theta}^{CI}) = \Pr(m^*(P_J, \theta_0) \geq 0) \geq \Pr(m^*(P_J, \theta_0) \geq m(P, \theta_0)) = \Pr(\sqrt{J}[m^*(P_J, \theta_0) - m(P, \theta_0)] \geq 0) = \Pr(\sqrt{J}[m(P_J, \theta_0) - m(P, \theta_0)] \geq -\bar{z}_{J,\alpha}(\theta_0)) \longrightarrow 1 - \alpha$$

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

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

$$\lim_{J \to \infty} \Pr\{\tilde{\Theta}^{CI} = \phi\} \leq \alpha,$$

where $\phi$ is the empty set.

**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.\(^{161}\)

\(^{161}\)One 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 $\bar{\xi}(\theta)$, choose $\bar{z}_{J,\alpha}(\theta)$ to have the form $\sigma(P_J, \theta)\bar{\xi}(\theta)$. 

158
• It should not be terribly surprising if \( \Theta_J \) is a singelton, say \( \theta_J \); i.e. if there is no (or only a single) value of \( \theta \) that satisfies \( m(P_J, \theta) \geq 0 \). For example, if \( m(P, \theta_0) = 0 \) 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 \( \theta \) 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 \( \theta_0 \), 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.

### 4.3.2 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 ($\nu_2$). 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 focus 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 $\delta$ 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, allows 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 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

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

So for any $\epsilon$,

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

If $\theta_0$ 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 $\theta_0$), and then computing the norm of its negative part.
Since \( \theta_0 \) 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 \( \theta \) (a stochastic process in \( \theta \)), say \( \{Z_J(\theta)\} \), with approximately the same distribution at each \( \theta \) as \( \{\sqrt{J}[m(P_J, \theta) - m(P, \theta)]\} \). Let \( z_{\alpha,J} \) be defined by

\[
Pr\left\{ \sup_{\theta \in CI_{1-\alpha/2}} \| (Z_J(\theta))^- \| \leq z_{\alpha,J} \right\} = 1 - \alpha/2.
\]

Then,

\[
Pr\{T_J \geq z_{\alpha,J}\} \leq Pr\{\theta_0 \notin CI_{1-\alpha/2}\} + Pr\{T_J \geq z_{\alpha,J}\ | \theta_0 \in CI_{1-\alpha/2}\} \leq \alpha,
\]

so \( z_{\alpha,J} \) is an \( \alpha \) level confidence interval for \( T_J \). More formally we have the following theorem.

**Theorem 2** Suppose (a) Assumption 3 holds; (b) \( CI_{1-\alpha/2,J} \) is such that

\[
\lim_{J \to \infty} Pr(\theta_0 \in CI_{1-\alpha/2,J}) \geq 1 - \alpha/2; \text{ and (c) } Z^*_J(\theta) \text{ is a stochastic process such that at each } \theta, \ Z^*_J(\theta)|P_J \sim \mathcal{N}(0, \Sigma(P_J, \theta)).
\]

Now define \( z_{\alpha,J} \) by \( Pr^*(sup_{\theta \in CI_{1-\alpha/2,J}} \| Z^*_J(\theta)^- \| \geq z_{\alpha,J}|P_J) \leq \alpha/2 \). Then under \( H_0 : m(P, \theta_0) \geq 0 \),

\[
\lim_{J \to \infty} Pr\left\{ \min_{\theta} \| (\sqrt{J}m(P_J, \theta))^- \| \geq z_{\alpha,J} \right\} \leq \alpha.
\]
PROOF:

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

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

The result follows by taking limits.

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 \( \sup_{\theta \in CI_{1-\alpha/2}} \|Z_j^*(\theta)\| \).\(^{162}\) As 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.\(^{163}\)

\(^{162}\)Note 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 \( \theta_0 \)). We impose no conditions on the covariances of \( \{Z_j^*(\theta)\} \) at different \( \theta \)'s, i.e. \( \text{Cov}(Z_j^*(\theta), Z_j^*(\theta')) \) 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 \( \theta \).

\(^{163}\)Another direction not pursued here is to compute the \( \bar{z}_{\alpha,J} \) from the draws on \( Z^*(\hat{\theta}_J) \), where \( \hat{\theta}_J \) is the estimate of the parameter value (recall that if \( \Theta_J \) is a set, the test is uninteresting).
4.3.3 Inference for Linear Models

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

\[
m(P_J, \theta) = Z_J \theta - W_J, \quad \text{so} \quad \Theta_J = \argmin_{\theta \in \Theta} \| (Z_J \theta - W_J)_- \| \quad (42)
\]

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

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

Note that we have assumed the sign of \( W \), 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 \( W \) is a normalization which chooses that scale.

In this setting \( \Theta_0 \) is the identified set and \( \Theta_J \) is the corresponding set estimator. Under the assumption that \( \Theta \) is compact and convex, \( \Theta_0 \) and \( \Theta_J \) 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 \( \theta \) (though, with a bit more notation, one could use analogous reasoning to find joint confidence regions for smooth functions of \( \theta \)).

Let \( \theta_k \) denote the \( k^{th} \) component of \( \theta \), and \( \Theta_{k,0} = \{ \theta_k : \theta \in \Theta_0 \} \). Then the closed convexity of \( \Theta_0 \) implies that \( \Theta_{k,0} \) is a closed interval on \( \mathcal{R} \). We denote that interval by \( \Theta_{k,0} = [\underline{\theta}_{k,0}, \overline{\theta}_{k,0}] \). Sample estimates of the \( k^{th} \) component bounds, \( \underline{\theta}_{k,0} \) and \( \overline{\theta}_{k,0} \), are available from \( \Theta_J \) (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

\[^{164}\text{It is straightforward to generalize the results in this section to models which are linear in non-linear functions of a set of parameters, provided those functions are homogenous of some degree.}\]
corresponding upper and lower bounds:

\[
\underline{f}_k(Z, W) = \min \{ \theta_k : \theta \in \arg \min_{\tilde{\theta} \in \Theta} \| (Z \tilde{\theta} - W) - \| \}
\]

\[
\overline{f}_k(Z, W) = \max \{ \theta_k : \theta \in \arg \min_{\tilde{\theta} \in \Theta} \| (Z \tilde{\theta} - W) - \| \}.
\]

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

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 \( \overline{f}_k \). 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 \( \overline{f}_k \) 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 \( \overline{f}_k \) 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 \( \alpha \) level confidence interval for \( \theta_{k,0} \) we look for
numbers, \((d^-_k(\alpha), d^+_k(\alpha))\) such that

\[
Pr \left\{ \theta_{k,0} \notin [\theta_{k,J} - d^-_k(\alpha), \bar{\theta}_{k,J} + d^+_k(\alpha)] \right\} \leq \alpha,
\]

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

\[
Pr \left\{ \theta_{k,0} \notin [\theta_{k,J} - d^-_k(\alpha), \bar{\theta}_{k,J} + d^+_k(\alpha)] \right\} \leq
Pr\{\theta_{k,0} \leq \theta_{k,J} - d^-_k(\alpha)\} + Pr\{\bar{\theta}_{k,0} \geq \bar{\theta}_{k,J} + d^+_k(\alpha)\}.
\]

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

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

Assumption 4 (a) The parameter space \(\Theta\) 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 \(f_k\) and \(\bar{f}_k\) and the duals to these programs, as defined in the Appendix, have unique nondegenerate solutions at \((Z, W)\).

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

\[\text{Note that this provides a confidence interval for the interval } [\theta_{k,0}, \bar{\theta}_{k,0}] \text{ and hence for } \theta_{k,0}. \text{ In this sense, one might find a shorter interval for just } \theta_{k,0}. \text{ For more on this distinction see Imbens and Manski (2004).}\]
bound functions, $f_k$ and $\overline{f}_k$, can be expressed as

$$f_k = \arg\min \theta_k \text{ s.t. } Z\theta \geq W \text{ and } \theta \in \Theta,$$

and

$$\overline{f}_k = \arg\max \theta_k \text{ s.t. } Z\theta \geq W \text{ and } \theta \in \Theta,$$

So the bound functions are optimum values of linear programs, and the boundedness of the space ensures 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 in a neighborhood of $(Z,W)$, but differentiability will sometimes hold under weaker conditions on the linear programs (see below).

Finally we need to ensure 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 5** For some $\zeta > 0$, $E \| (Z_{j,J}, W_{j,J}) \|^{2+\zeta} < \infty$.

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

**Theorem 3** Given Condition 1, and Assumptions 2, 4 and 5,

(a)

$$\sqrt{J} \left( \begin{pmatrix} f_k(S_{j,J}) - f_k(S) \\ \overline{f}_k(S_{j,J}) - \overline{f}_k(S) \end{pmatrix} \right) \xrightarrow{d} N(0,V)$$

where $V = \Gamma V_S \Gamma'$ and $\Gamma = (\nabla f_k(S)', \nabla \overline{f}_k(S)')'$ the stacked partial derivatives of $f_k(S)$ and $\overline{f}_k(S)$; and
(b) Let

$$S^*_j = S_j + \frac{V_{S^{1/2}}^c}{\sqrt{j}} + o_{as}(1/\sqrt{j})$$

where $\epsilon^*$ 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 $P_J$)

$$\sqrt{j} \left( \bar{f}_k(S^*_j) - \bar{f}_k(S_j), \overline{f}_k(S^*_j) - \overline{f}_k(S_j) \right) \xrightarrow{d} N(0, V).$$

Proof of Theorem.

Let $T_\eta$ denote the $\eta$ neighborhood of $S$ given by the conclusion of the Lemma in the Appendix. Also, let $f_k(\cdot) = (\bar{f}_k(\cdot), \overline{f}_k(\cdot))$. All partial derivatives of $f_k$ exist (and are continuous) on $T_\eta$. Let $\Gamma(S) = \left( \begin{array}{c} \nabla \bar{f}_k(S) \\ \nabla \overline{f}_k(S) \end{array} \right)$ (so $\Gamma = \Gamma(S)$).

We prove conclusion (b), since conclusion (a) is standard and follows by analogous reasoning. A superscript $\omega$ will be used to denote a particular sample sequence. Let the $o_{as}(1)$ term in (44) be denoted $\tau_j$. Let $A = \{ \omega : \lim_{j \to \infty} S^*_j = S, \lim_{j \to \infty} \tau^*_j = 0 \}$. 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 $S^*_j \in T_\eta$ for all $J \geq \bar{j}$. Let $P^*$ denote the probability for $S^*_{j\omega}$ along the sample sequence given by $\omega$. Then, $|S^*_{j\omega} - S_j| \xrightarrow{as*} 0$ and $S^*_{j\omega} \xrightarrow{a} S_j \xrightarrow{a} S$. Therefore, $S^*_{j\omega} \xrightarrow{a} S$. 

168
For $J \geq \bar{J}$,

\[
\sqrt{J}(f_k(S_j^{\omega}) - f_k(S_j^{\omega})) \\
= \sqrt{J}(f_k(S_j^{\omega}) - f_k(S_j^{\omega}))\mathbf{1}_{\{S_j^{\omega} \in T_\eta\}} \\
= \sqrt{J}(f_k(S_j^{\omega}) - f_k(S_j^{\omega}))\mathbf{1}_{\{S_j^{\omega} \in T_\eta\}}\mathbf{1}_{\{S_j^{\omega} \in T_\eta\}} + o_{as^*}(1) \\
= \sqrt{J}[\Gamma(S_j^{\omega})] - S_j^{\omega})\mathbf{1}_{\{S_j^{\omega}, S_j^{\omega} \in T_\eta\}} + o_{as^*}(1) \\
= \sqrt{J}\Gamma(S_j^{\omega})\mathbf{1}_{\{S_j^{\omega}, S_j^{\omega} \in T_\eta\}} + o_{as^*}(1) \\
\rightarrow N(0, \Gamma V_S \Gamma')
\]

where the first equality follows by the choice of $\omega$ and $J$; the second from the fact that $S_j^{\omega} \overrightarrow{\text{as}} S^{166}$; the third from the differentiability proven in the Lemma in the Appendix; the fourth from an argument similar to the first, the continuity of $\Gamma(\cdot)$ at $S$ 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 the conditions of the theorem and substitute them into the linear program in (43), then the distribution of the solutions to that program will be identical to the limit distribution of $(\hat{\theta}_{k,j}, \overline{\theta}_{k,j})$. If $\hat{V}_S$ is the sample covariance of $S$, then $\hat{V}_S \overrightarrow{\text{as}} V_S$. So we can set $S_j^*$ to be $S_j + \frac{\sqrt{J}}{2}$. I.e., there is a natural way to construct the distribution of moments substituted into the linear program; take draws from

\footnote{\textit{Let $A^* = \{\omega^* : \lim_{J \rightarrow \infty} S_j^{\omega^*}(e^{\omega^*}) = S\}$. Then $P(A^*) = 1$. By the definition of $A^*$, for a given $\omega^* \in A^*$ there exists $\bar{J}$ such that $|S_j^{\omega^*}(e^{\omega^*}) - S| < \eta$ for all $J \geq \bar{J}$. Hence, $\mathbf{1}_{\{S_j^{\omega^*}(e^{\omega^*}) \notin T_\eta\}} = 0$ for all $J \geq \bar{J}$ and $\sqrt{J}(f_k(S_j^{\omega^*}(e^{\omega^*})) - f_k(S_j^{\omega^*}))\mathbf{1}_{\{S_j^{\omega^*} \in T_\eta\}}\mathbf{1}_{\{S_j^{\omega^*} \in T_\eta\}} = 0$ for all $J \geq \bar{J}$. Then, $P(\{\omega^* : \lim_{J \rightarrow \infty} \sqrt{J}(f_k(S_j^{\omega^*}(e^{\omega^*})) - f_k(S_j^{\omega^*}))\mathbf{1}_{\{S_j^{\omega^*} \in T_\eta\}}\mathbf{1}_{\{S_j^{\omega^*} \in T_\eta\}} = 0\}) \geq P(A^*) = 1$, i.e. $\sqrt{J}(f_k(S_j^{\omega^*}) - f_k(S_j^{\omega^*}))\mathbf{1}_{\{S_j^{\omega^*} \in T_\eta\}}\mathbf{1}_{\{S_j^{\omega^*} \notin T_\eta\}} = o_{as^*}(1).}$
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 es-
timated in a preliminary stage and that $S_J = S_J(\tilde{\beta})$ where $S_J(\tilde{\beta}) = S_J(\beta_0) + 
\Psi_{\beta_0}(\tilde{\beta} - \beta_0) + o_p(1/\sqrt{J})$. Then, the expression for $V_S$ is the asymptotic variance
of

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

The asymptotic variance of the right-hand side expression above is usually ob-
tained 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 up-
per 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 approxima-
tion by linearizing the solution to the linear programming problem. We suggest
the simulated estimator, because we think it likely that the normal approxima-
tion to means of data moments, particularly to moments of differences that are
unlikely to have large tails, 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 4 (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 (43) still define $f_k$ and $\overline{f}_k$, but these bound functions may not be continuously differentiable in a neighborhood of $(Z, W)$. 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 them$^{167}$. However even in these cases directional derivatives may still exist at $(Z, W)$, 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 4.3.2 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 $U^*_J = \frac{\nu^{1/2} x^*}{\sqrt{J}}$, which is just $S^*_J$ centered at zero. Note that $U^*_J$ implicitly gives simulation draws on $(Z_J, W_J)$, i.e. $U^*_J = vec(Z^*_J, W^*_J)$, 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 $\bar{z}_{\alpha, d}$ in Theorem 2.

### 4.4 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

---

$^{167}$In particular if it does occur for one choice of $(D(d_i))$, it is unlikely to occur for another.
substantive problem. Next we describe our method of moments inequality estimators and discuss their properties. We conclude with the empirical results which contain both our inequality estimators and alternative estimators that have been used in the past for similar problems. We are currently working on Monte Carlo experiments and on a more detailed set of empirical results than are currently available; so this section of the chapter should be viewed as work in progress.

4.4.1 Ordered Choice

This subsection 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 chapter can be used to empirically analyze “lumpy” investment decisions, or investment decisions subject to adjustment costs which are not convex for some other reason\textsuperscript{168}. The analysis allows for interacting agents as then the signs of the expected profitability differences follow from the conditions for a Nash equilibrium.

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

\textsuperscript{168} Actually Ishii’s problem has two sources of non-convexities. One stems from the discrete nature of the number of ATMs choice, the other from the fact that network effects can generate increasing returns to increasing numbers of ATMs.
and its competitors, i.e. on \((d_i, d_{-i})\) (since there is only one type of agent in her model, we ignore the \(t\) index in this section). 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 consumers pay to an ATM owner when that owner is not the consumer’s bank). Interest rates are set in a simultaneous moves Nash game. This setup provides Ishii (2004) with the parameters needed to compute the banks’ earnings conditional on the ATM networks in existence\(^{169}\).

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 chapter provides initial estimates of those costs, while Ishii (2004) provides robustness tests and considers the implications of the results.

4.4.1.1 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 \(d_i \in D \subset \mathbb{Z}^+\), the non-negative integers. In the simplest 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

\(^{169}\text{These earnings are calculated as the earnings from the credit instruments funded by the deposits minus the costs of the deposits (including interest costs) plus the fees associated with ATM transactions. The ATM fee revenue is generated when non-customers use a bank’s ATMs and revenue is both generated and paid out when customers use a rival’s ATMs.}\)
\[
\pi(y_i, d, d_{-i}, \theta) = r(y_i, d, d_{-i}) - (\nu_{2,i} + \theta)d + \nu_{1,d,i},
\]

where \( r(y_i, d, d_{-i}) \) is the profits that would been earned in the second stage if the firm chose \( d \) and its competitors chose \( d_{-i} \) in the first stage, \( \theta \) is the average (across banks) of the marginal cost of purchasing and installing ATM’s, the \( \nu_{2,i} \) is the unobserved bank specific deviation from that marginal cost, and \( \nu_1 \) and \( \nu_2 \) have the properties listed in our Assumption 2.

\( r(\cdot) \) in equation (45) 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 \( \Delta \pi(\cdot) \).\(^{170} \) So \( m = 2 \) and \( \mathcal{E}\Delta \pi(\cdot) \) is the vector consisting of

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

and

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

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 \( \Delta r(\cdot) \) with \( h(x) \) (since \( \nu_2 \) 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 \( \Delta r(\cdot) \otimes 1 \). Then the

\(^{170}\)These conditions will also be sufficient if the expectation of \( \pi(\cdot) \) is (the discrete analogue of) concave in \( d_i \) for all values of \( d_{-i} \), a condition which can not be checked without more detail on the model. On the other the realizations of profits evaluated at the estimated value of \( \theta \) were concave in \( d_i \) for almost all banks.
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

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

\[
= \frac{1}{J} \sum_j [\Delta \tau_L^j(\cdot) - \theta] = \Delta \tau_L - \theta,
\]

where

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

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

\[
m_R(P_j, \theta) = \Delta \tau_R + \theta.
\]

The set of \(\theta\) that minimize of the objective function in equation (38) are the values of \(\theta\) that make both equations (46) and (47) positive. If \(-\Delta \tau_R < \Delta \tau_L\) then this set is

\[
\Theta_J = \{\theta : -\Delta \tau_R \leq \theta \leq \Delta \tau_L\}
\]

while if \(-\Delta \tau_R \geq \Delta \tau_L\) there is a single \(\theta\) which minimizes (38) and it is given by\(^{171}\)

\[
\Theta_J = \{\theta_J = \frac{1}{2}[-\Delta \tau_R + \Delta \tau_L]\}.
\]

Either way we obtain the conservative confidence interval in Theorem 1 as

\(^{171}\)In the simple case where \(h(x) \equiv 1\), if \(r(\cdot)\) is concave in \(d_i\) then, at least in expectation, \(\Delta \tau_L \geq -\Delta \tau_R\), so we do not expect the set of minimizers to be a singleton. Once we add instruments, however, concavity no longer ensures that the inequalities will be satisfied by a set of \(\theta\) values.
\[ \Theta^{CI} = \{ -\Delta \bar{\tau}_R - z_\alpha \leq \theta \leq \Delta \bar{\tau}_L + z_\alpha \}, \tag{48} \]

where \( z_\alpha \) is defined as in that theorem.\(^{172}\) Interestingly, in this simple case it is easy to show that provided \( \Delta \bar{\tau}_L > -\Delta \bar{\tau}_R \), the confidence interval in equation (48) is identical to the confidence interval from Theorem 3.

**Comparing the Two Confidence Intervals.**

The coincidence of the two confidence intervals is because the model produced only one upper and one lower bound for \( \theta \); the Theorem 1 confidence interval is only “conservative” in the way it treats non-binding constraints. 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), so then there will be constraints that do not bind. Indeed if \( h \) indexes instruments

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

where \( \Delta \bar{\tau}_{h,R} \) provides the upper bound from the \( h^{th} \) instrument and so on. So \( \Theta_J \) becomes shorter (weakly) as the number of instruments increases.

Now consider the two confidence intervals. The confidence interval from Theorem 1 is based on the probability that at \( \theta = \theta_0 \) *none* of the inequalities generated by the instruments will be violated, regardless of whether they bind. The confidence interval from Theorem 3 is based on the probability that only a random selection of those constraints, those that bind for the simulation draw, are not violated.

\(^{172}\)Note that this is a case in which the variance covariance of the moment restrictions do not depend on \( \theta \) and, as a result, neither does \( z_\alpha \). This makes the confidence interval particularly easy to construct, but is a property which does not survive most complications to the model. For example it does not extend to the case where there are additional determinants of the costs of ATMs or where we weight the averages from different markets before averaging over markets, see the discussion below.
As a result the $z_{\alpha,h}$ of the Theorem 1 confidence interval in equation (48) must increase with the number of instruments (and hence moments) as indexed by $h$. To see this consider three moment conditions. Then the $z_{\alpha,3}$ in (48) satisfies

$$1 - \alpha = Pr\{\epsilon_3 \geq -z_{\alpha,3}, \epsilon_2 \geq -z_{\alpha,3}, \epsilon_1 \geq -z_{\alpha,3}\} =$$

$$Pr\{\epsilon_3 \geq -z_{\alpha,3} | \epsilon_2 \geq -z_{\alpha,3}, \epsilon_1 \geq -z_{\alpha,3}\} \times Pr\{\epsilon_2 \geq -z_{\alpha,3}, \epsilon_1 \geq -z_{\alpha,3}\}.$$

When we only have two moments, $1 - \alpha = Pr\{\epsilon_2 \geq -z_{\alpha,2}, \epsilon_1 \geq -z_{\alpha,2}\}$, which implies that $z_{\alpha,3} \geq z_{\alpha,2}$. An inductive argument proves the analogous result for any increase in the number of moments. The increase in $z_{\alpha,h}$ with $h$ depends on its correlation with existing constraints.\(^{173}\)

The Theorem 3 or simulated confidence interval for this problem can always be written in the same form as the conservative confidence interval, i.e. as the interval between $\max(-\Delta \tau_{1,R}, -\Delta \tau_{2,R})$, minus an adjustment factor and $\min(\Delta \tau_{1,L}, \Delta \tau_{2,L})$ plus an adjustment factor. However the adjustment factors for the simulated confidence intervals need not increase as we increase the number of moments. For example, assume $\Delta \tau_{1,L} < \Delta \tau_{2,L} < \Delta \tau_{1,L} + z_{\alpha}^2$, and $\Delta \tau_{2,L}$ has no variance. Since $\Delta \tau_{2,L}$ has no variance we know that the upper bound to the simulated confidence interval when we move from $h = 1$ to $h = 2$ falls from $\Delta \tau_{1,L} + z_{\alpha}^2$, to $\Delta \tau_{2,L}$. Here the simulated confidence interval accounts for the fact that some constraints can be superfluous.

**Properties of Parameter Estimates**

When $h > 1$ the estimate of the lower bound for $\theta_0$ 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, it should not be surprising if, in finite

\(^{173}\)One way of seeing this is to recall that since the $\epsilon$’s have a joint normal distribution, $\epsilon_3 = \rho_1 \epsilon_1 + \rho_2 \epsilon_2 + \sigma \xi$ with $\xi$ standard normal, and substituting this into the formula above.
samples, there is a positive bias in the binding constraint (if $h$ binds, $\hat{\theta}_h$ will tend to be larger than $\theta_h$). This bias should increase with the number of inequalities. So when there are a large number of inequalities it should not be overly surprising if the estimated lower bound is greater than $\theta_0$. Analogously, since the estimate of the upper bound is a minimum, it should not be overly surprising if the upper bound estimate is less than $\theta_0$. Of course, if the lower bound is greater than $\theta_0$ and the upper bound is less than $\theta_0$, then the estimate $\Theta_J$ is just a point (even if the true $\Theta_0$ is an interval). This situation accentuates the need for a test with good small sample properties, and underlies our interest in the Monte Carlo results below (to come).

**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

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

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

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

If we plot these two inequalities on a graph, their boundaries will be given by two parallel lines. If $\Delta \tau_L > -\Delta \tau_R$, then $\Theta_J$, the estimate of $\Theta_0$, 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 $\Theta_J$ 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). Moreover if we construct moments that drop those observations we violate our Condition 1 (in particular then the sample mean of the structural error would converge to the expected value of that error conditional on not being at the boundary, and not to zero).

For this example, consider a market in which a number of banks chose not to purchase ATM’s \((d = 0)\), and so do not have a change from the left. If we let \( \Delta r_{L,d,i} \equiv r(y_i, d, d_{-i}) - r(y_i, d - 1, d_{-i}) \) and assume that it decreases in \( d \), then if we simply drop the \( d = 0 \) banks and form the average of \( \Delta r_{L,d,i} \) among the firms with \( d_i \geq 1 \), Assumptions 1 and 2 only ensure that

\[
\mathcal{E}[\Delta r_{L,d,i}] \geq \theta + \mathcal{E}[\mathcal{E}[\nu_{2,i} \mid \mathcal{J}_i] \leq \Delta r_{L,d=1,i} - \theta].
\]

Since \( \mathcal{E}[\nu_{2,i} \leq \Delta r_{L,d=1,i} - \theta] \leq \mathcal{E} \nu_{2,i} = 0 \), the last term in the above expression is negative, and the estimated upper bound from the selected subsample need not converge to a number larger than \( \theta \).

The severity of the boundary problem is likely to depend on how many observations are at the boundary, and we should be able to get an indication of its magnitude by using a function of an \( x \in \mathcal{J} \) to select a subsample of firms for which \( \mathcal{E}[\Delta r_{L,i} \mid \mathcal{J}_i] \) is likely to be large and rerunning the inequality tests. A large difference in the estimates indicates a need to modify the estimator to account for the boundary problem. One way to circumvent this selection problem would be to substitute any other variable which the researcher is willing to assume is greater than \( \theta + \nu_{2,i} \) for the \( r(y_i, d_i, d_{-i}) - r(y_i, d_i - 1, d_{-i}) \) of firms with \( d_i = 0 \), and form \( m_L(\cdot) \) by averaging
across the whole sample. For example we could assume that the change in revenues from another decision by the same agent is larger than the cost of that decision which, in turn, is larger than the cost of the first unit of the good in question (we might know that the change in revenue from adding the last teller is larger than the tellers’ cost, which, in turn, is larger than the cost of an ATM; for related examples see the empirical results below).

4.4.1.2 Alternative Estimators and Empirical Results

There are two estimators that have been traditionally used for ordered problems (i) an ordered discrete choice model (e.g. ordered probit or logit), and (ii) a first order condition estimator that ignores the discrete nature of the control and forms moment conditions from the optimality conditions for a continuous control (as in Hansen and Singleton, 1982). We introduce each of these procedures before turning to the empirical results.

Alternative Estimators

Typically, ordered choice models do not allow for firm specific fixed effects, for expectational errors, or for measurement errors. In our notation the ordered choice model sets \( \nu_1 \equiv 0 \) in equation (45), assumes a particular distribution for \( \nu_2 \), and forms the likelihood of the observed \( d \).

Regardless of the distribution chosen, the likelihood of any \( \theta \) in this model on our data is minus infinity; so the 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). We note that as long as there is some uncertainty when decisions are made, or some measurement error in measured
returns, we should expect one agent’s difference from the left to be less than its
difference from the right even if all agents are behaving optimally.

The stochastic assumptions required by the first order condition estimator 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 differ-
ences in outcomes not explained by observables to $\nu_1$. Then if agents are maximizing,
the first order condition for agents with a $d > 0$ must have an expectation of zero
conditional on $J$. As a result, the estimator finds a $\theta$ that minimizes

$$
\left\| \frac{1}{J} \sum_j \frac{1}{n^j} \sum_i \{d^j_i > 0\} \left( \frac{\partial r(y^j_i, d, d^j_{i-1})}{\partial d} \big|_{d=d^j_i} - \theta \right) \times h(x^j_i) \right\|.
$$

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).\(^\text{174}\) Whether or not this causes a substantial
difference in the estimates is likely to depend on the “lumpiness” of the investment
good.

More conceptually interesting, in the model with $\nu_2 = 0$ we simply drop the agents
at a boundary (with $d = 0$ in our context). We do this because (i) the “derivative”
at $d = 0$ need not be zero, and (ii) when $\nu_2 = 0$ the model allows us to select
observations with $d$ values in any subset of $D$. I.e. with $\nu_2 = 0$ the difference
between the observable realization of the first order condition and its expectation is
entirely determined by $\nu_1$, and $\nu_1$ is mean independent of $d$. As noted above when
$\nu_2$ is nonzero (and nondegenerate), the decision will generally depend on its value.

\(^{174}\)This 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 condi-
tion 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.
Then selecting on $d$ could violate Condition 1. We come back to this distinction in discussing our empirical results below.

**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 contains a total of 291 banks in 10 markets. 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 has a standard deviation of 40.1)\(^{175}\). This is a small sample, at least by modern I.O. standards, and as a result we will focus on parsimonious specifications for the cost function.

Both the inequality and the first order condition estimators require instruments (the $h(x)$ above). The instruments used include a constant term, the market population, the number of banks in the market, and the number of branches of the bank (this has a mean of 6 and a standard deviation of 15). These instruments are all highly, and positively, correlated with one another. Also 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)\(^{176}\).

Table 20 contains the inequality estimators.\(^{177}\) The first row provides the results

---

\(^{175}\)The 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.

\(^{176}\)It also implies that some of the properties of the simple model discussed in the last section do not necessarily hold here. For example the $z_\alpha(\theta)$ now depends on $\theta$, and the two confidence intervals need not be exactly equivalent for the single instrument case.

\(^{177}\)All estimators for both empirical problems analyzed in this paper were obtained using the “fmincon” algorithm in Matlab. In the linear case, “fmincon” finds the $\text{argmin}$ 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 $\theta$ we obtain the vertices of $\Theta_J$. For details on the search method used in fmincon see
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. As expected, when there is only one upper and one lower bound, the simulated and the conservative confidence intervals are very close to one another, placing the true $\theta_0$ between $23,000$ 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 $38,444$. Note that though $\Theta_J$ reduces to a point, the conservative confidence interval increases with the number of moments – though not dramatically (since the instruments are highly and positively correlated).

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 moments for the inequality 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 assume the larger banks would have had at least one ATM at any cost in the support of the cost distribution, and use an average of the returns from the first ATM’s for the banks that did, 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 having only a minor impact. This result is preliminary. More detailed results with corrections for the boundary problems in all of the subsequent rows of this table should be available shortly.

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 \( \pi_i(\cdot, d) = r_i(\cdot) - d_b \theta_b - d_r \theta_r - \nu_i(d_b + d_r) \). We get point estimates for each cost, with \( \theta_r \) about 5% higher than \( \theta_b \). However the confidence interval for \( \theta_r \) covers that for \( \theta_b \), which, in turn, is similar to the confidence intervals we obtained when we did not differentiate branch locations\(^{178}\).

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.\(^{179}\) When we considered richer models the confidence intervals do widen but the estimates do not change much. However the sample is small and we did not expect to be able to estimate a more detailed specification for the cost function.

There are a number of other implications of these estimators that are worth noting.

\(^{178}\)We initially expected a larger convenience and rental cost advantage of in-branch locations. However on going back to the data we found that 16 banks own remote ATMs 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.

\(^{179}\)These estimates of costs are for costs over a six month period. So the estimated ATM cost per location is about $6,030 per month, with a confidence interval of $6,140 to $5,330. There are, on average, 1.3 ATMs per location. The only other estimate of costs we are aware of is from a survey by Dove Consulting which estimates the average monthly cost per ATM for a large bank in 2001 was around $1,500 (Dove Consulting (2002)). There are, however, reasons to expect this figure to underestimate true ATM costs. First, it is based on a nation-wide survey, but average commercial rental rates and other costs are likely higher in Massachusetts than in other parts of the country. Second, many of the survey respondents are huge ATM deployers whose per-ATM purchase, maintenance, and processing costs is likely to be lower than the average bank. Third, the survey cost figure does not include costs of constructing, installing, and programming an ATM, which presumably can be high.
First there was no specification in which the confidence interval in Corollary 1 was the null set; so the model passes that conservative test. We have not implemented the specification test in Theorem 2 yet. Second the confidence intervals are not symmetric about the point estimates when there are point estimates, a fact that comes from the non-normality of the estimator’s small sample distribution. Finally the fact that the estimates from the model when \( D(d_i) = \{d : |d - d_i| = 2\} \) are close to the results when \( D(d_i) = \{d : |d - d_i| = 1\} \), indicates that there is no reason to doubt that firms, at least on average, act optimally.

This last result bodes well for the first order condition estimator that we now turn to, as that estimator cannot be modified to allow for non-optimal choices. Table 21 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 we allow for a separate \( \theta_r \) and \( \theta_b \), the estimates are larger than those obtained from the inequality estimator, but they are also much less precise.

These estimators seem to behave as would the inequality estimator with an extra source of error (perhaps due to using derivatives when it is inappropriate to do so), and possibly an upward bias. In this context, we note that the inequality estimator is no more difficult to compute than the first order condition estimator, and we expect the extra error in the latter to be larger the more lumpy the investments.\textsuperscript{180}

Finally, both the policy and the descriptive implications of our estimators are examined in some detail in Ishii(2004).

\textsuperscript{180}Note also that this estimator drops the \( d = 0 \) observations entirely, while even if we do not correct the inequality estimator for dropping the \( d = 0 \) observations, that estimator still uses those observations in constructing the inequality from the right.
Table 20: Inequality Method, ATM Costs

<table>
<thead>
<tr>
<th></th>
<th>( \Theta_J )</th>
<th>Simulated 95% Conf. Interval</th>
<th>Conservative 95% Conf. Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h(x) \equiv 1 )</td>
<td>[32,006, 32,492]</td>
<td>23,301 41,197</td>
<td>22,535 42,368</td>
</tr>
<tr>
<td>( h(x) = \text{full, } d &gt; 0 )</td>
<td>32,492</td>
<td>29,431 38,444</td>
<td>15,501 45,368</td>
</tr>
<tr>
<td>( h(x) = \text{full, } d \geq 0 )</td>
<td>32,522</td>
<td>29,571 38,478</td>
<td>16,944 44,157</td>
</tr>
</tbody>
</table>

Alternative Inequality Estimators of \( \Theta_0 \) (\( h(x) = \text{full} \))

\[
D(d_i) = \{d : |d - d_i| = 2\} \\
D(d_i) = \{d : |d - d_i| = 1, 2\}
\]

Extending the Model (\( h(x) = \text{full} \))

\[
\begin{array}{llllll}
\theta_b \text{ (in-branch constant)} & 36,649 & 32,283 & 38,871 & 10,317 & 58,244 \\
\theta_r \text{ (remote constant)} & 38,348 & 26,179 & 47,292 & 5,617 & 71,387 \\
\end{array}
\]

Table 21: First Order Conditions, ATM Costs

<table>
<thead>
<tr>
<th>( \theta_{01} ) (constant)</th>
<th>Coeff. 38,491</th>
<th>Std. Error 7,754</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_0 ) (in-branch constant)</td>
<td>50,132</td>
<td>11,102</td>
</tr>
<tr>
<td>( \theta_1 ) (remote constant)</td>
<td>55,065</td>
<td>12,010</td>
</tr>
</tbody>
</table>

Notes: There are 291 banks in 10 markets in every specification. 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 equilibrium for interest rates.
4.4.2 Discrete Choice and Buyer/Seller Networks

This subsection is based on Chapters 2 and 3. The model analyzes the interactions between privately insured consumers, health insurers, and hospitals in the US medical care market. A summary of the relevant points are given here for ease of reference. The model described here is more general than that in Chapter 3 in that it allows for a structural error term $\nu_2$ in addition to the expectational errors already permitted (these are denoted $u$ and $\phi$ in Chapter 3; they are rolled into $\nu_1$ in this chapter). It also considers the hospital’s profit-maximization problem as well as that of the plan.

The analysis can be put in the framework of a three stage game. In the last stage consumers choose a health insurance plan given the hospital networks and the premiums made available by the plans in the market. In the second stage plans set their premiums in a Nash equilibrium which conditions on both consumer preferences and the established hospital networks. The first stage consists of a bargaining game between the hospitals and the plans and it generates a set of hospital networks; one for each plan. The bargaining equilibrium establishes contracts which govern how the surplus of premiums over costs are split between the hospital and the plan.

Chapter 2 obtains estimates of the demand system and of hospital costs. Chapter 3 provides descriptive results on the formation of hospital networks. It finds that certain markets tend to have insurer networks which exclude certain hospitals, while other markets tend to have “inclusive equilibria”, equilibria in which each plan contracts with all major hospitals, and that markets which are not inclusive are socially costly. As a result we would like to understand the circumstances which generate non-inclusive equilibria, and how they impact other plan and hospital decisions (e.g. investment choices).

In order to analyze these questions we need 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” (a similar
set of issues arise in the analyzing many markets where vertical relationships are important). This section provides one such framework, shows how to apply our econometric techniques to make inferences on its parameters, and uses it to investigate the nature of the contracts for a hospital’s services.

We assume a Nash equilibrium; that is, the profits each agent expects to earn as a result of its choices are greater then the profits it would expect to earn from alternative feasible choices (Assumption 1). There are likely to be many configurations of hospital networks that satisfy this condition, and we will not investigate how this multiplicity of possible equilibria gets resolved (and, therefore, will not be able to perform counterfactual experiments). We simply assume that the contracts that materialize have a fixed fee and a per patient component (i.e. two-part tariffs), and then provide a “reduced form” characterization of how these components relate to plan, hospital, and market characteristics. This reduced form characterization ought to help determine the relevance of alternative bargaining models, and, to the extent policy and environmental changes do not affect the reduced form relationship per se, provide some idea of how changes in market characteristics are likely to impact on hospital markups. Chapter 3 provides a more detailed investigation of the implications of the results presented here.

4.4.2.1 The Model

We begin with the determinants of returns, and then develop the inequalities to be taken to the data.

The Returns to Different Networks.

A consumer’s utility from a hospital network conditional on the consumer having a given diagnosis is estimated from the observed consumer choices of hospitals. Here Chapter 2 uses a discrete choice demand system which depends on observed individual
characteristics and observed and unobserved hospital characteristics. The consumer’s expected utility from a given network is 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 plan as a function of this expected utility, the premiums the plan charges, and other observed and unobserved plan characteristics. Chapter 2 estimates the parameters of the function determining the utility from an insurer’s characteristics using market level data on consumers’ plan choices (as in Berry, Levinsohn, and Pakes 1995).

Given these estimates, the surplus of plan $i$, if its network of hospitals is $H_i$ and the competitor’s networks are given by $H_{-i}$, is derived as

$$S_i(H_i, H_{-i}) = \sum_q \left( n_q s_{qi}(H_i, H_{-i}) \left[ prem_i - p_q \sum_{h \in H_i} s_{qh}(H_i)c_h \right] \right)$$

(49)

where $n_q$ is the population of consumer-type $q$ (detailed by zip code, age, and gender), $p_q$ is the probability that a type-$q$ person will be admitted to hospital, $s_{qi}(H_i, H_{-i})$ and $s_{qh}(H_i)$ are plan $i$’s and hospital $h$’s predicted shares of type-$q$ people when networks $H_i$ and $H_{-i}$ are offered by plan $i$ and other plans respectively, $c_h$ is the average cost of treatment at hospital $h$, and $prem_i$ is plan $i$’s premium. The shares and premiums are obtained from the demand estimates and the Nash in premiums assumption.\(^{181}\)

The profit of plan $i$ is this surplus minus the costs of its contracts with hospitals, while the profit of hospital $h$ is the payments from the contracts minus their costs. We assume that the transfers implicit in the contracts can be expressed as two part tariffs with fixed cost $fc(\cdot)$, and per patient markup $mk(\cdot)$. The actual contracts are largely proprietary, and when accessible are too complicated to summarize in a small

\(^{181}\)We note that the surplus calculation also takes account of hospital capacity constraints. If any network combination implies that any hospital is over 85% of its maximum capacity level, patients are reallocated randomly to non-capacity constrained hospitals in the market.
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 been a Nash equilibrium in the sense of Assumption 1. As a result we write $f_c(\cdot) = f_c(x_{i,h,j}, \beta)$ and $m_k(\cdot) = m_k(x_{i,h,j}, \beta)$ where $x_{i,h,j}$ is a vector of plan, hospital, and market characteristics, and $\beta$ is a parameter vector to be estimated.

More specifically, if $T_{i,h}$ denotes the realized transfers from plan “$i$” to hospital “$h$,” we assume

$$T_{i,h} = f_c(x_{i,j,h}, \beta) + m_k(x_{i,h,j}, \beta)N_{ih}(H_i, H_{-i}) + \nu_{2,i,h} \quad (50)$$

where $\nu_2$ has the properties as given in Assumption 2, and

$$N_{ih}(H_i, H_{-i}) \equiv \sum_q n_q p_q s_q(H_i, H_{-i}) s_{qh}(H_i),$$

i.e. $N_{ih}(H_i, H_{-i})$ is the number of plan $i$’s enrollees treated by hospital $h$.

With this notation, plan profits can be expressed as the difference between the surplus in equation (49) and the transfers made to the hospitals in its network, or

$$\pi^M_i(H_i, H_{-i}, \beta) = r^M_i(H_i, H_{-i}, \beta) - \sum_{h \in H_i} \nu_{2,i,h} + \nu_{1,i}, \quad (51)$$

where

$$r^M_i(H_i, H_{-i}, \beta) \equiv \mathcal{S}_i(H_i, H_{-i}) - \sum_{h \in H_i} f_c(x_{i,j,h}, \beta) - \sum_{h \in H_i} N_{ih}(H_i, H_{-i}) m_k(x_{i,h,j}, \beta).$$

Hospital profits are simply the net transfers they receive from their network of hospitals, say $M_h$, or

$$\pi^H_h(M_h, M_{-h}, \beta) = r^H_h(M_h, M_{-h}, \beta) + \sum_{i \in M_h} \nu_{2,i,h} + \nu_{1,h}, \quad (52)$$
where

\[ r^H_h(M_h, M_{-h}, \beta) \equiv \sum_{i \in M_h} f_c(x_{i,j,h}, \beta) + \sum_{i \in M_h} N_{i,h}(H_i, H_{-i})m_k(x_{i,h,j}, \beta). \]

Note that because \( \nu_{2,i,h} \) is a disturbance in the transfers between plans and hospitals, each \( \nu_{2,i,h} \) that appears in (52) for a given hospital appears with an opposite sign in (51) for the respective plan. Of course there may also be expectational errors in both the surplus, and in the patient flows (in \( N_{i,h} \)), provided they satisfy the properties of \( \nu_1 \) in Assumption 2. What 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. Below we show how to generalize and allow, in addition, plan and hospital specific fixed (or random) effects.

**Constructing Moment Inequalities**

In order to determine which inequality constraints are satisfied in equilibrium, we need assumptions on the nature of the bargaining game. There are a number of possibilities here, but we shall focus on a model where hospitals make simultaneous take it or leave it offers to plans\(^{182,183}\). Then Assumption 1 implies that hospitals expect to increase their profits from any offer that is accepted, and plans accept or reject contract offers according as the offers increase or decrease their expected profits.

\(^{182}\)As noted in Chapter 3, a number of other bargaining models are consistent with the estimation procedure set out here. For example, any model of sequential offers (e.g. plans make offers to all hospitals; hospitals make counter-offers to all plans) would be fine provided there is an exogenous end-point, no contracts can be agreed before this point, and plans make the final contract choice.

\(^{183}\)In related work Pakes (2004) considers two models that specify the strategies of the players and the timing of the game (and are in that sense “structural”), and one model that is reduced form. In one structural model hospitals make take it or leave it offers to plans, and in the other plans make take it or leave it offers to hospitals. In both these cases the empirical results explore the characteristics of the offers. Pakes’ third possibility is more reduced form in nature in that it tries to characterize the contracts that were established (instead of those that were offered). Here the assumption is that if a contract is established it leads to an increase in the expected profits of both parties to the contract, and if no contract is established the increment to the total expected profits that would have been split between the two sides had the contract been established was negative.
We begin with the assumption that there is no structural error (i.e. $\nu_2 \equiv 0^{184}$). In particular we assume the six largest hospitals (they cover an average of 57% of admissions to hospitals) simultaneously make take-it-or-leave-it offers to all insurers. We start by using only the inequalities implied by the plan decisions; so the $\Delta \pi(\cdot)$ required for the objective function in equation (38) is a six element vector obtained by computing the difference between our estimate of the profits for the observed networks and our estimate of the profits for the six alternative networks obtained by reversing the plan’s contract choice with each of the six largest hospitals in the market (if the plan did contract with the hospital we compare to a situation in which it did not contract with the hospital, and vice versa). We have not yet added the inequalities implied by the hospitals’ decisions; i.e. that hospitals expect to make positive returns from accepted contracts.

Those are the only empirical results we have to date. We present the theory required to allow for the structural error, and for firm and hospital effects, directly after our empirical results. However this section is incomplete in that we have not yet done the tests for the structural errors or estimated the model which allows for them.

A Traditional Model for Comparison

We also compare the results to those from a logit model. For the logit model, we use the profits from the different networks given by equation (51), and assume that plans both know the value of any unobserved disturbance in that equation at the time it makes its decisions and includes that value in its measure of profits. In addition the logit model assumes that those disturbances have i.i.d. extreme value distributions, and then estimate using maximum likelihood.

The assumptions on the disturbance are likely to be problematic for a number of

---

184This is the assumption made in Chapter 3.
reasons. Most importantly these assumptions imply that there is no expectational or measurement error in the profit measure (in our terminology the assumption is \( \nu_1 \equiv 0 \)). Moreover if we interpret the disturbance as structural (in the sense of \( \nu_2 \)), then it can not be both independent of profits and a determinant of the firm’s decision (since those decisions determine profits), so maximum likelihood should lead to inconsistent estimates.\(^{185}\)

### 4.4.2.2 The Data and Empirical Results

The primary data set contains every insurance plan in 43 major US markets, and considers the network of hospitals these plans offered to enrollees in March/April 2003. This data includes characteristics of 451 plans and 665 hospitals in these markets carefully put together from a number of different data sources (for more detail, see Chapters 2 and 3). As in the ATM example, we find 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.

Plan profits are estimated as plan membership times the plan’s premium minus the plan’s transfers to the hospitals (fixed cost plus markup per patient times the predicted number of patients). The plan membership and the patient flows needed for these calculations have to be recomputed every time we evaluate a different network.\(^{186}\) Below we specify the fixed cost and the markup terms in the transfers as

---

\(^{185}\)Note 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 plan, or independent across different plans. 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.

\(^{186}\)In principle premiums should also be adjusted when plan \( i \) considers a deviation from its observed network, but they depend on non-hospital related plan costs which we do not know. However, since the premium changes are for changes of only one hospital in the network we expect them to be small, and we examine the robustness of our results with respect to the fixity of premiums shortly after

193
a function of observables and a parameter vector to be estimated. All but one of these observables are also used as instruments (hospital cost per admission is used as a possible determinant of markups, but not as an instrument due to concerns over measurement error in this variable). The additional instruments used to construct $h(x)$ in equation (38) include other market and plan characteristics known to the agents at the time of the contracting decision.\textsuperscript{187}

This is not a large data set and we have no theoretical restrictions on the form of the equations determining the transfers in equation (50). So again we have to be somewhat parsimonious. We have focused on determinants of $fc(\cdot)$ and $mk(\cdot)$ suggested by standard IO and/or bargaining models. These variables, which are motivated and described in detail in Chapter 3, include\textsuperscript{188}:

- A constant term in $mk(\cdot)$ which picks up a “standard” markup per patient treated.

- A measure of the extent 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.\textsuperscript{189}

\textsuperscript{187}They include indicator variables for high proportion of population aged 55-64, 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. Note that, unlike in the ATM example, the instruments used here are neither particularly highly nor all positively correlated with one another.

\textsuperscript{188}Recall that since the estimator is based on the changes in profits that results from changes in choices, we should not expect to be able to analyze how variables that do not vary with plan choices affect profits, and so such variables are omitted from the analysis.

\textsuperscript{189}The 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.
• A measure of hospital costs per admission.

• Hospitals in systems, and hospitals for which at least one same-system provider is excluded by the plan.

• "Star" hospitals: as identified by the US News Hospital Rankings for 2003.\footnote{US News magazine publishes an annual report giving hospital rankings for 17 different specialties and overall. We considered two different definitions of "star" hospitals: those that feature on the magazine’s overall Honour Roll, implying a ranking among the top 17 hospitals in the country; and those with high reputation rankings among physicians. They produced similar results.}

There was not enough information in the data to allow for interactions of the latter four variables with both the fixed and the per patient component of the transfers\footnote{When we estimated models allowing these variables 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.}. The results presented below were based on a specification that makes: (i) the fixed component of the contract depend on whether a hospital was in a system, whether another member of that system was not contracted with, and whether it was a star hospital, and (ii) the variable component depend upon whether the hospital was capacity constrained and the costs per admission of the hospital.

**Estimates (without structural errors)**

Table 22 provides the results from the specification without the hospital inequality, and without the Star Hospital determinant of $f_c(\cdot)$. Since the proposed estimator has minimal computational time, it was feasible to explore a number of robustness checks.\footnote{One run which uses two hundred simulated estimators to approximate the limit distribution, takes about ten minutes on our machine which has an Intel Pentium 3 processor, 1.33 GHz hardrive, and 512 MB of RAM.} Table 17 in Section 3.8.2 presents a selection of (the many) robustness tests we ran. Table 16 includes results with the Star Hospital measure.

The estimate of $\Theta_0$ from every specification was a singleton, i.e. there was no parameter vector that satisfied all the inequality constraints. On the other hand, the
test in Corollary 1 was always accepted at a confidence level of .05, so it does not give us a reason to reject our specifications. All specifications have over eighty inequality constraints, so neither of these results should be surprising. We have not yet run the more powerful test in Theorem 2.

The point estimates in Table 22 all have the expected sign, and have magnitudes which are consistent with the (little bit) of information from other sources at our disposal (see below). Three of the five coefficients are significant at the traditional 5% level, the “hospital in a system” coefficient is significant at the 10% level, and the constant term in the markup was significant at the 12% level. On the other hand the confidence intervals are reasonably large.

Figures 4 and 5 in Chapter 3 provide the simulated distributions for four of our coefficients. All of these distributions are somewhat right skewed and have most of the mass between zero and about a third of the upper bound to the confidence intervals. So there may be somewhat more information in the estimates than is suggested by the confidence intervals. Note also that with over eighty inequalities and instruments that are not particularly highly (or even positively) correlated with one another, the conservative confidence intervals for this example “blow up”. We also computed a confidence interval by first dropping all those constraints that are not binding, and then applying the standard GMM formula for obtaining standard errors to the moments from the binding constraints. Table 22 shows that the GMM confidence intervals can be quite misleading.

What is clear from the empirical results is that we have identified variables that are important determinants of the outcomes of the negotiations between hospitals and insurers. Hospitals in systems do in fact both: (i) take a larger fraction of the surplus, and (ii) penalize a plan that does not contract with all its members. Moreover capacity constrained hospitals earn higher markups per patient, and hospitals with higher costs per patient earn lower markups.
To help interpret magnitudes we note that the total average cost of an admission is roughly $11,000 (American Hospital Association, 2001). 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%. We note that the Kaiser Family Foundation Report (2004) estimates the average profit margin for community hospitals to be 4.2%, which is somewhat higher. We do estimate that capacity constrained hospitals receive an extra $1800 per patient which, accounting for their cost figures, translates into an average markup of approximately 15% of their revenues. 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 revenues. Also we are estimating a large 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 determinants of hospital markups available.

We now consider the logit estimates. The differences between these estimates and our inequality estimates are quite striking. Most striking is that the logit estimates indicate that we should be quite certain that hospitals in systems receive lower markups (the t-value for this coefficient is over sixteen). A natural explanation for this finding would be 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 results by estimating that the plan can contract with in-system hospitals at a lower cost. The logit model also estimates markups for capacity constrained and for cost per admission which are implausibly large in absolute value (they are a factor of about three higher than our
estimates and have relatively small standard errors). 193

<table>
<thead>
<tr>
<th>Characteristics of Hospitals</th>
<th>Fixed Component (Units = $/million, per month)</th>
<th>Per Patient Component (Units = $/thousand, per patient)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>θ</td>
<td>95% CI</td>
</tr>
<tr>
<td>Hosp. in system</td>
<td>0.18</td>
<td>0.62</td>
</tr>
<tr>
<td>Drop Same System Hospital</td>
<td>0.60</td>
<td>1.81</td>
</tr>
<tr>
<td>Constant</td>
<td>2.61</td>
<td>11.21</td>
</tr>
<tr>
<td>Cap. constrained</td>
<td>1.87</td>
<td>11.39</td>
</tr>
<tr>
<td>Cost per admission</td>
<td>-0.26</td>
<td>-0.20</td>
</tr>
</tbody>
</table>

Notes: "Hosp. in system" refers to whether the hospital is in a system, “Drop Same System Hosp” refers to a situation in which a same system hospitals dropped, “Cap-constrained.” is the measure of capacity constrained developed in text. The “GMM SE” is the standard error computed from the GMM formula for the constraints that bind. “Conservative 95% CI” refers to the confidence interval from our Theorem 1, “Simulated 95% CI” refers to the confidence interval from our Theorem 3.

Table 17 in Chapter 3 presents results from a selection of the robustness checks. We have done a large number of other robustness tests on the specification that

193In explaining the biases leading to these two results one should keep in mind that any error in the costs per admission enters our specification in two places (once in the surplus calculation and once directly), and that the hospitals that are capacity constrained in a market have lower observed costs.
only uses the plan constraints (they include additional interactions as well as using inequalities induced by reversing the decision made on any two of the six largest hospitals). Though none of these robustness tests change the qualitative nature of the results, some of the coefficients do change in magnitude.

4.4.2.3 Allowing for Structural Errors

We begin with the general case where \( \nu_{2,i,h} \) just satisfies the conditions in Assumption 2, and then come back to the familiar special case of hospital and plan effects. As noted we have not yet had time to do the empirical analysis that corresponds to the models presented in this section.

Let \( H_i \setminus h \) be the set consisting of all the hospitals in \( H_i \) but hospital \( h \), and \( H_i \cup h \) be the set obtained when we add hospital \( h \) to the set \( H_i \). Then we require the following differences

\[
\Delta \pi_i^M (+h, \beta) = \pi_i^M (H_i, H_{-i}, \beta) - \pi_i^M (H_i \setminus h, H_{-i}, \beta)
\]

\[
\Delta \pi_i^M (-h, \beta) = \pi_i^M (H_i, H_{-i}, \beta) - \pi_i^M (H_i \cup h, H_{-i}, \beta)
\]

\[
\Delta \pi_h^H (+i, \beta) = \pi_h^H (M_h, M_{-h}, \beta) - \pi_h^H (M_h \setminus i, M_{-h}, \beta).
\]

and define \( \Delta r_i^M (-h, \beta), \Delta r_i^M (+h, \beta), \) and \( \Delta r_h^H (+i, \beta) \) analogously. Finally let \( \chi(i, h) \) be the indicator function which takes the value of one if plan \( i \) and hospital \( h \) contract, and zero elsewhere. We are focusing on the cases where the differences are based on addition or subtraction of one hospital to a plan, but certainly multiple hospital differences would also be possible (and perhaps quite informative).
The General Case.

Then Assumption 1 implies that the expectation of

\[ U^\pi(i, h; \beta) \equiv \chi(i, h)\Delta \pi^H_i (+i, \beta) + (1 - \chi(i, h))\Delta \pi^M_i (-h, \beta), \]

is positive. The intuition is that hospitals expect to increase their profit from signed contracts, and the plans only reject offers when their expected profits are higher without them. If \( x \in J_i \cap J_h \), then the conditional expectation of \( U^\pi(\cdot) \) is non-negative as required in Condition (1) (equation (36)).

Suppose \( U^r(\cdot) \) is defined analogously to \( U^\pi(\cdot) \). From the definitions in equations (51) and (52),

\[
U^\pi(i, h; \beta) = \chi(i, h)[\Delta r^H_i (+i, \beta) + \nu_{1,i,h}] + (1 - \chi(i, h))[\Delta r^M_i (-h, \beta) + \nu_{1,i,h}] + \nu_{2,i,h} = U^r(i, h; \beta) + \nu_{2,i,h},
\]

which is linear in \( \nu_{2,i,h} \), as required by Condition 1 (equation (35)).

Since \( E[U^r(i, h, ; \beta) | x] \geq 0 \), and \( U^r(i, h, ; \beta) \) is observable up to \( \beta \), \( U^r(\cdot) \) can be one component of the \( \Delta r(\cdot) \) used to construct the moment inequalities in our objective function (equation (38)).

To obtain a second component, define

\[
S^\pi(i, h; \beta) = \chi(i, h) \left[ \pi^M_i (+h, \beta) + \Delta \pi^H_i (+i, \beta) \right],
\]

i.e. \( S^\pi(\cdot, \beta) \) is the sum of plan and hospital profits if they contract and zero otherwise, and as a result has a non-negative (conditional on \( x \)) expectation. Moreover since this sum does not depend on the transfers between these two agents (though it does depend on transfers between both of them and other agents), it does not depend on \( \nu_{2,i,h} \), or more formally \( S^\pi(i, h; \beta) = S^r(i, h; \beta) \), where \( S^r(i, h; \beta) \) is defined analogous to \( S^\pi(\cdot) \).
This linear combination of profit differences eliminates the structural error completely \((g(\cdot) = 0\) in Condition 1), and selection issues are eliminated. Consequently any element of \(J_i \cap J_h\), say \(z_{i,h}\), can serve as an “instrument” to interact with \(S^r\).

Thus for this model of buyer-seller networks equation (38) is given by averages across markets of terms

\[
[U^r(i, h; \beta) \otimes h^U(x_{i,h})], S^r(i, h; \beta) \otimes h^S(z_{i,h})]
\]

for nonnegative functions \(h^U\) and \(h^S\).

**Special Case: “Effects” Models.**

We now consider the special case in which

\[
\nu_{2,i,h} = \nu_{2,i} + \nu_{2,h}, \quad \mathcal{E}[H^{-1} \sum_h \nu_{2,h} | x] = 0, \quad \text{and} \quad \mathcal{E}[M^{-1} \sum_i \nu_{2,i} | x] = 0,
\]

where \(H\) is the total number of hospitals and \(M\) is the total number of plans. The conditional mean zero assumptions are analogous to the traditional “random” (in contrast to the “fixed”) effects assumptions. In deriving the additional inequality restrictions that can be obtained under equation (54), we will point out which of them survive the weakening of our conditions to those of a fixed effect model (i.e. to a model that allows the conditional means of the unobservables not to be mean zero).

We begin with some inequalities we will need. Assumption 1 implies that for each \(i\) and \(x \in J_i \cap J_h\)

\[
\sum_h \mathcal{E} [\chi(i, h)\Delta \pi^H_{i,h}(+i, \beta) + [1 - \chi(i, h)]\Delta \pi^M_{i,(-h)}(\beta)|x] \geq 0.
\]

Writing the \(\Delta \pi(\cdot)\) in terms of the \(\Delta r(\cdot)\) the \(\nu_{2,h}\) and the \(\nu_{2,i}\) we obtain
\[ 0 \leq \sum_{h} E \left[ \chi(i,h) \Delta r^H_h (+i, \beta) + [1 - \chi(i,h)] \Delta r^M_i (-h, \beta) \right] + \sum_{h} E[\nu_{2,h} | x] + HE[\nu_{2,i} | x]. \]

Since \( H^{-1}E \sum_{h} [\nu_{2,h} | x] = 0 \), this can be rewritten as

\[ \mathcal{E}[\nu_{2,i} | x] \geq \mathcal{E}[l(i, \beta) | x] \quad (55) \]

where

\[ l(i, \beta) \equiv -H^{-1} \left[ \sum_{h \in H} \Delta r^H_h (+i, \beta) + \sum_{h \notin H} \Delta r^M_i (+h, \beta) \right]. \]

Similarly for any hospitals that contracts with all plans

\[ \mathcal{E}\left[ \sum_{i \in M_h} \Delta r^H_h (+i, \beta) | x \right] + M_h \mathcal{E}[\nu_{2,h} | x] + \mathcal{E}\left[ \sum_{i \in M_h} \nu_{2,i} | x \right] \geq 0, \]

which implies that

\[ \mathcal{E}[\nu_{2,h} | x] \geq \mathcal{E}[l(h, \beta) | x], \quad \text{where} \quad l(h, \beta) \equiv -M^{-1} \left[ \sum_{i \in M_h} \Delta r^H_i (+i, \beta) \right]. \quad (56) \]

First consider a hospital that contracts with every plan. Then for those hospitals

equations (52), (55), and (56) imply that for every \( i \)

\[ \mathcal{E}[\Delta r^M_i (+h, \beta) - l(i, \beta) - l(h, \beta) | x] \geq 0, \quad (57) \]

which produces \( M \) inequalities for every such \( h \). Next consider a hospital which contracts with all plans but \( i^* \). Then for any \( i \neq i^* \) equation (51) implies

\[ \mathcal{E}[\Delta r^M_i (+h, \beta) + \Delta r^M_{i^*} (-h, \beta) - \nu_i + \nu_{i^*} | x] \geq 0. \]
Now take any other hospital $h'$ that $i^*$ does contract with. If $i$ does not contract with $h'$, equation (51) implies

$$\mathcal{E}[(\Delta r_i^M (+h, \beta) + \Delta r_i^M (-h, \beta) + \Delta r_i^M (-h', \beta) + \Delta r_i^M (+h', \beta)|x] \geq 0 \quad (58)$$

which is a difference in difference inequality. If $i$ also contracts with $h'$ we have from equation (52)

$$\mathcal{E}[(\Delta r_i^M (+h, \beta) + \Delta r_i^M (-h, \beta) + \Delta r_i^H (+i, \beta) + \Delta r_i^M (+h', \beta)|x] \geq 0. \quad (59)$$

Thus for every hospital which omits one plan we have $M - 1$ such inequalities.

Note that the inequalities in equations (58) and (59) did not require the assumption that $H^{-1} \sum_h \nu_{2,h} = 0$ or the assumption that $M^{-1} \sum_i \nu_{2,i} = 0$ (the special case where Condition 1 equation (35) is satisfied with $g = 0$). That is, these inequalities survive the strengthening of our assumptions to those of a fixed effect model. As a result they do not require the hospitals and the plans in the sample to be randomly drawn from a population which abides by our conditional mean zero assumption.

When a hospital does not contract with two or more plans we repeat this process, once for each plan it does not contract with, and this will generate $(M-2)^2$ inequalities for each such $h$, and so on.

When the assumptions in equation (54) are relevant the inequalities (57),(58), and (59) are added to the inequalities in (53) and the entire system is used in the moment inequalities that define our objective function (equation (38)).
4.5 Appendix

Let $Q(Z, W)$ denote the following linear program:

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

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

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

The value of the program $Q(Z, W)$ is $f_k(Z, W)$ as defined previously, and similarly for $\overline{Q}(Z, W)$. Under Assumption 4, $Q(Z, W)$ has a unique optimal basis, which we will denote $\beta \subset \{1, \ldots, h\}$. Also let $\theta$ denote the solution to $Q(Z, W)$.

**Lemma** Under Assumption 4, there exists $\eta > 0$ such that for all $(Z, W)$ with $\| (Z, W) - (Z, W) \| < \eta$,

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

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

**Proof:**

If the conclusion does not hold, then there exists a sequence $(Z_n, W_n) \longrightarrow (Z, W)$ such that $\beta$ is not the unique optimal basis for $Q(Z_n, W_n)$. Without loss of generality, we can assume $(Z_n, W_n)$ is close enough to $(Z, W)$ to ensure that $Q(Z_n, W_n)$ is feasible (i.e. that there is a value of $\theta$ which solves this problem). The constraints $\theta \in \Theta$ and Assumption 4 imply that if $Q(Z_n, W_n)$ is feasible then a bounded solution exists. Since $\beta$ is not the unique optimal basis for $Q(Z_n, W_n)$, there exists a different basis $\alpha_n$ that is optimal. Also let $\theta_n$ denote the solution corresponding to the basis $\alpha_n$. Note that $\alpha_n$ is a sequence that can only take on a finite number of possible values. The sequence $\alpha_n$ must then have at least one limit point in this finite set of values. For any such limit point, say $\alpha$, there exists a subsequence $n'$ such that $\alpha_{n'} = \alpha$. Now
note that $\theta_{n'}$ is a sequence in the compact space $\Theta$ and so must have a convergent subsequence, $\theta_{n''} \to \theta^*$. 

The unique solution to $Q(Z, W)$ is bounded, then by Goldfarb and Todd (1989) Theorem 4.2(b), the dual program to $Q(Z_n, W_n)$ 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, $f_k$, is continuous at $(Z, W)$. Hence, $\theta_k^* = \theta_{k,0}$.

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

$$Z^\beta \theta = W^\beta \text{ and } Z^{-\beta} \theta > W^{-\beta} \quad (60)$$

and

$$Z^\alpha_{n''} \theta_{n''} = W^\alpha_{n''} \text{ and } Z^{-\alpha}_{n''} \theta_{n''} \geq W^{-\alpha}_{n''}.$$

Taking limits,

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

Thus, $\theta^*$ is a feasible solution to $Q(Z, W)$. Since $\theta_k^* = \theta_{k,0}$, $\theta^*$ is also optimal. By uniqueness, we must have $\theta^* = \theta$. Since $\alpha \neq \beta$, there exists an element $l$ such that $l \in \alpha$ and $l \notin \beta$. But then $Z^l \theta^* = Z^l \theta = W^l$, which contradicts the strict inequality in (60). The conclusion (a) follows for $Q(Z, W)$ and the argument for $Q(Z, W)$ is symmetric.

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

205
References


[34] Imbens, G. and C. Manski (2003), “Confidence Intervals for Partially Identified Parameters," manuscript.


