Semi-parametric instrument-free demand estimation: relaxing optimality and equilibrium assumptions*

Sungjin Cho, *Seoul National University*†
Gong Lee, *Georgetown University*‡
John Rust, *Georgetown University*§
Mengkai Yu, *Georgetown University*¶

September 19, 2019

Abstract

In most markets, consumer demand results from a compound arrival/choice process: consumers arrive to a market stochastically and make independent discrete choices over which item to purchase (or not to purchase, often referred to as the "choice of the outside good"). Market demand results from an aggregation of individual consumer choices, and in general is more accurately modeled as a price-dependent probability distribution (or stochastic process) rather than as a linear demand function. We consider the problem of identifying the underlying structure of demand — consumer preferences and the distribution of arrivals — when market prices are endogeneously determined but the implied distribution of demand (including mean demand) is potentially nonlinear in prices and there are no relevant instrumental variables. In addition, demand data are truncated and censored: we do not observe the number of arriving customers or those choosing the outside good, and we only observe the minimum of demand and each hotel's capacity. Recent studies have shown that the hypotheses of a) optimality, and b) equilibrium constitute powerful identifying restrictions that enable consistent estimation of demand in the presence of a variety of endogeneity and censoring problems. In this paper we consider whether it is possible to identify demand when assumptions a) and b) are relaxed. We introduce a conditional independence assumption that implies that after controlling for a vector of "demand shifters" x other variables affecting firm prices, z, do not also affect demand. This implies that by controlling for x we can exploit the residual variation in firm prices due to the unobserved shocks z as "virtual price experiments" to identify the underlying structure of demand. We illustrate our approach via an empirical analysis of demand for hotels. We show that the distribution of hotel occupancies is a mixture of censored multinomial distributions that depends on prices, consumer preferences, and the distribution of arriving customers. In our empirical application, we show that the structure of demand can be estimated in a situation where there are no relevant instrumental variables without the need to impose optimality or equilibrium restrictions, with limited data. Our estimates imply stochastically shifting but downward sloping expected demand curves. Using the estimated stochastic process for demand we are able to test and reject the hypothesis that hotels are setting their prices optimally (as well as prices predicted under Bertrand-Nash equilibrium) and we strongly reject claims that revenue management systems (RMS) used by hotels to set prices are implementing "algorithmic collusion."

Keywords endogeneity, truncation, censoring, identification, maximum likelihood estimation, semi-parametric estimator, instrument-free, structural estimation, identification of mixture models, conditional independence, Bertrand-Nash equilibrium, revenue management systems, algorithmic collusion

*__Acknowledgements:__ We thank Michael Keane, Dennis Kristensen and two anonymous referees for helpful suggestions that lead to us to develop a new estimator that relaxes a key assumption of optimal dynamic hotel pricing that we relied on in the previous version of this paper. We are also grateful to Yuichi Kitamura for helpful correspondence on the topic of identification of mixture models, and to the revenue manager of hotel 0 (who must remain anonymous due to confidentiality restrictions) for providing the reservation data that made this study possible. We also acknowledge STR for market level hotel occupany data it provided that proved key to the empirical identification of our demand estimation. We thank Georgetown University for research support, and Rust gratefully acknowledges financial support provided by the Gallagher Family Chair in Economics.

†Department of Economics, Seoul National University e-mail: sungcho@snu.ac.kr

‡Department of Economics, Georgetown University, e-mail: gl430@georgetown.edu

§Department of Economics, Georgetown University, e-mail: jr1393@georgetown.edu

¶Department of Economics, Georgetown University, e-mail: my262@georgetown.edu

1 Introduction

In most markets consumer demand results from a compound arrival/choice process: consumers arrive to a market over time stochastically and make independent discrete choices over which item to purchase, including the choice not to purchase (or to purchase an item outside of this market which economists refer to as the "choice of the outside good"). Market demand results from an aggregation of individual consumer choices, and is more appropriately modeled as a nonlinear price and state-dependent stochastic process rather than a linear demand curve that is traditionally used for demand estimation. We consider the problem of identification of consumer preferences and arrivals when market prices are endogeneously determined but the implied stochastic process for demand is nonlinear in prices and there are no relevant instrumental variables to deal with price endogeneity. In addition, most data are truncated: we typically do not observe the number of arriving customers or the subset choosing the outside good.

The motivation for our paper is an empirical analysis of hotel pricing in a specific luxury hotel market in a major US city, see Cho, Lee, Rust, and Yu (2018). Demand shocks in the face of the fixed room capacity for the 7 hotels in this local market lead to strong positive correlation between hotel prices and room occupancy: the hotels raise their prices to ration their available capacity on days where the demand for rooms in this market is high, but lower their prices significantly to compete for market share on days when demand is low and occupancy rates are well below 100%. This leads to strong co-movement in prices and occupancy rates in the hotels in this market. OLS estimation of room occupancy on hotel prices results in positively sloped "demand curves" for hotel rooms due to the endogeneity in hotel pricing in response to fluctuating demand.

Price endogeneity also arises as a result of unobserved characteristics of the hotels in this market: although all seven hotels are classified as "luxury hotels" and we can control for their star rating, there are unobserved characteristics that make customers willing to pay more to stay in the top tier hotels in this local market, and their higher willingness to pay enables these hotels to charge more.

Though it has long been recognized that it is possible to deal with the latter form of endogeneity in certain types of nonlinear models using market share data only without the benefit of observing the discete choices of individual customers (see e.g. Berry, Levinsohn, and Pakes (1995)), these approaches depend on the ability to invert market shares to obtain transformed equations that are linear in prices, to which instrumental variable approaches can be applied. However it is unclear how to apply this approach in our hotel market example, where we typically observe only observe the occupancy of a single hotel in the market, but not occupancies at competing hotels. As a result, we do not observe market shares that the BLP estimator inverts to form the regression equations to which instrumental variables can be applied.

Even in situations where we can observe the joint occupancies of all of the hotels, we still almost never observe the total population of who "arrived" and considered whether to book a room in this market. Thus, we do not observe the share of consumers who chose the "outside good' (i.e. to not stay in any of the hotels in this market). Without information on the outside good, we cannot construct the market shares necessary to apply BLP. We argue that this is a typical situation in many if not most markets.

On top of this, there are many situations where there may not be any relevant instrumental variables. Our hotel data set is one such example. A good instrument is an observed variable that causes "exogenous shifts" in a hotel's price but does not enter the hotel's pricing strategy. It is hard to think of such variables in the hotel market example. A typically used instrumental variable, the so-called "Hausman instrument", is the price of competing hotels. However the validity of instrumental variables depends on an "exclusion restriction" that the prices of competing hotels are not a determinant of any given hotel's pricing decision in this market. For the hotel we study, the price of competing hotels is probably the most important determinant of its own pricing decisions, and thus can hardly be plausibly assumed to be an excluded variable from the hotel's pricing equation (which is a reduced-form approximation to its pricing decision).

These problems motivate a search for new approaches to demand estimation that can also handle econometric problems such as truncation and censoring. A recent paper by MacKay and Miller (2018) introduces a novel "instrument-free" approach to demand estimation: "Our main result is that price endogeneity can be resolved by interpreting an OLS estimate through the lens of a theoretical model. With a covariance restriction, the demand system is point identified, and weaker assumptions generate bounds on the structural parameters. Thus, causal demand parameters can be recovered without the availability of exogenous price variation." (p. 32). However their approach depends on the assumption that demand or market share data, after a suitable transformation, is a linear (or semi-linear) function of price and it requires a prior restriction on the covariance between cost shocks and an additively separable unobserved demand shock or unobserved product characteristic. In our hotel example, there is no transformation of occupancy rates or market shares that results in the linear equations needed to implement the MacKay and Miller (2018) estimator. Further, since the supply of rooms is inelastic in the short run, we are not aware of covariance restrictions on cost and demand shocks that could help identify the stochastic process for demand for hotel rooms and enable us to predict how consumers react to hotel prices.

Structural models of demand and firm price setting provide an attractive alternative to traditional linear instrumental variables approaches to demand estimation because they enable us to more directly model the dynamics of demand in real world markets, such as the hotel market we study in this paper. Identification of arrival probabilities and heterogeneous consumer preferences can be obtained without

the use of instrumental variables or the use of covariance restrictions as in MacKay and Miller (2018). However the structural approach is heavily dependent on three key assumptions: 1) parametric functional forms for consumer preferences and the stochastic process governing arrival probabilities, 2) firms set prices optimally, and 3) prices are in equilibrium, i.e. their prices mutually satisfy conditions for a Nash (or Markov Perfect equilibrium or some related solution concept in dynamic models).

Together these three assumptions are often powerful enough to secure identification of the unknown parameters of consumer preferences and the stochastic process for arrival of customers to the market. The structural approach requires modeling the entire market and incorporating observed and unobserved variables that capture the demand shocks and unobserved product characteristics that result in the endogeneity in the prices we observe. By explicitly modeling the endogenous determination of prices under the assumptions of optimality and equilibrium, dynamic structural models are able to bring to bear "cross equation" identifying restrictions that imply that equilibrium prices are an implicit function of firms' beliefs about the stochastic process of customer arrivals and preferences, as well as each others' price-setting and strategic behavior. Demand is downward sloping in a well formulated structural model, since upward sloping demand would result in price dynamics that are at odds with what we actually observe.

However a big drawback of structural models is the computational demands of solving and simulating a dynamic Markov Perfect Equilibrium for an entire market. This is a daunting task even for a relatively small local hotel market consisting of 7 luxury hotels that we analyze in this paper. Fortunately, recent work has shown that it is often possible to identify demand in the presence of endogeneity in a framework that relaxes the equilibrium assumption so long as the optimality assumption is still imposed. The idea is that the behavior of competing firms or agents can be flexibly modeled using semi-parametric estimators by treating the pricing strategies of competing firms as infinite-dimensional "nuisance parameters."

For example, Merlo, Ortalo-Magne, and Rust (2015) studied optimal dynamic strategies of home sellers in the London housing market. Endogeneity arises in this market due to the presence of unobservable characteristics of houses that make some more attractive to most buyers. Homes that are superior in unobservable dimensions (even after controlling for a large set of observed hedonic neighborhood and home characteristics) experience a high rate of arrival of offers and sell for more. Thus observed housing demand appears to be "upward sloping" in the list price of a home if we fail to control for price endogeneity. Yet there were few instrumental variables the authors could find to do this. The alternative of estimating a dynamic structural model that imposes the assumption of a full dynamic equilibrium in the London housing with thousands of competing buyers and sellers is computationally infeasible.[1] Merlo et al. (2015) were

[1]If there is twos-sided incomplete information, then even the bargaining "subgame" between a seller and an buyer can have have a huge multiplicity of equilibria. To our knowledge nobody has succeeded in solving the overall twos-sided search, matching

able to flexibly model the arrival of buyers and the dynamic bargaining process they employed. These can be regarded as the infinite-dimensional nuisance parameters in their estimation problem. However by assuming home sellers follow an optimal dynamic pricing and bargaining strategy in response to these beliefs, the authors were able to structurally estimate their model and obtain a plausible downward sloping "demand curve" for housing. The hypothesis of optimality restricts demand to be downward sloping in price, since if it were upward sloping, then it would be optimal for sellers to set far higher list prices than we actually observe.

We wish to take this approach one step further, to see if it is possible to identify demand by relaxing the hypothesis of optimality as well as equilibrium in the hotel market. Though the assumption of optimality is a powerful identifying assumption, it is also a potentially dubious one that could distort our estimates of demand if firms do not behave optimally. Herbert Simon introduced the concept of *bounded rationality* as a key reason why organizations and firms fail to optimize in complex environments.[2] Cho et al. (2018) discuss a multi-billion *revenue management industry* that is experiencing very rapid growth by helping hotels, airlines, and other firms in the hospitality industry set better prices. If all of the hotels, airlines and other firms were already optimizing (the typical default assumption in most economic models), there would be little need and value-added for the revenue management industry. Yet, Phillips (2005) notes that despite the fact that pricing decisions "are usually critical determinants of profitability" "pricing decisions are often badly managed (or even unmanaged)." (p. 38). If this is true, it calls into question the standard structural assumptions that individual firms price optimally, and even more so the stronger assumption that firm prices collectively are determined as the outcome of a Bertrand-Nash equilibrium.

In this paper we adopt a structural semi-parametric approach that explicitly models and attempts to identify the probability distribution governing the arrival of potential customers that constitutes the fundamental "demand shock" that leads to the co-movement of prices and occupancy in this market. We also identify consumer preferences for the different hotels and their willingness to pay to stay in them. Via microaggregation of the individual discrete choices of arriving consumers (which we do not observe) we derive a mixed and censored multinomial distribution for the joint occupancies of the hotels in this market. The censoring arises because of the hotel capacity constraints: when a hotel is fully booked, some consumers who would have liked to stay there are turned away and will either book at a competing hotel or choose the outside good. We can identify both the probability distribution of arriving customers and the fraction of them choosing the outside good, even though we only observe joint occupancies of the hotels and not the total number of arriving customers, nor the number choosing the outside good.

and bargaining game that would be the most realistic way to model real housing markets.

[2] See Rust (2019) for further discussion and evidence in support of Simon's view that many firms *satisfice* rather than optimize.

We are able to do this without the imposition of equilibrium or optimality assumptions, or the use of instrumental variables. Our key identifying assumption is a conditional independence assumption on firm prices that we refer to as *conditional exogeneity.* In essence, we assume there is a *demand shifter* x that hotels in this market observe and use in their price setting decisions that constitutes a sufficient statistic for their beliefs about "demand shocks" that result in the positive correlation between demand (arrivals) and the hotels' prices. In our application, x is the *expected market occupancy rate* i.e. expected market demand for rooms in this market divided by the total room capacity. Though we do not directly observe x, we show how we can proxy for this using *predicted occupancy* where we treat x as a latent variable. Firm prices are a function of this demand shifter x and other variables z which we do not observe. We interpret these other z variables as "pricing shocks" that arise due to potential mistakes or other idiosyncratic factors on the part of the hotels in this market. The conditional independence assumption states that once we condition on x and the hotels' prices, the z variables do not affect the distribution of hotel demand. In essence, once we condition on x any residual variation in hotel prices can be considered to be "virtual random price experiments" that allows us to identify how prices affect demand, by enabling us to identify the deeper structure of demand — consumer preferences and the distribution of arrivals.[3] Following the literature, we will also refer to our assumption as conditional exogeneity: i.e. conditional on x and prices p the idiosyncratic random variables z affecting prices have no effect on the realized demand for hotels.

We discuss the non-parametric identification of our model demand in the presence of endogeneity and censoring when we relax the assumptions of equilibrium and optimality. We treat the price setting strategies of firms as infinite-dimensional nuisance parameters and discuss semi-parametric estimation strategies including the method of sieves that are capable of consistently estimating these infinite-dimensional nuisance parameters while still estimating the parameters defining consumer preferences at the usual $\sqrt{N}$ rates (where N is the sample size). There is a cost to relaxing any maintained assumption, and relaxing the optimality and equilibrium assumptions has the consequence that the estimated pricing strategies are no longer implicit functions of firms' beliefs about consumer preferences and arrival rates. That is, the semi-parametric estimators we propose no longer benefit from the "cross equation restrictions" that link consumer preferences and arrival rates to the pricing strategies of firms. Even though actual pricing strategies undoubtedly do depend on firms' beliefs about their customer preferences and arrival rates, the use of semi-parametric methods to estimate our econometric model comes at the cost of a loss of information

[3] Our conditional independence assumption is similar to the "unconfoundedness" assumption in the treatment effects literature, where conditional on x the distribution of potential outcomes of a binary treatment are independent of the treatment assignment. That is, conditional on x the "treatment assignment" T is akin to a "virtual random experiment" which enables analysts to identify the average treatment effect in situations where treatments are actually "endogenous" and not actually determined by randomized assignment. In our application we can regard the hotels' prices p as a "continuous treatment" and the "average treatment effect" we are attempting to identify is the effect of prices on demand for hotels in this market, i.e. the slope of the demand curve.

from failing to impose the optimality and equilibrium restrictions. At a minimum this loss of information is reflected in larger asymptotic standard errors for the parameters, but in the worst case the demand model parameters may not be identifiable, and thus cannot be consistently estimated.

On the other hand, structural estimators that impose optimality and equilibrium restrictions will generally result in inconsistent demand estimates if actual firm behavior violates these assumptions. Thus, it is desirable to develop estimators that can identify demand without imposing these strong assumptions. We consider identification of the model under two scenarios: 1) we observe only the occupancy at a single hotel, but not its competitors; 2) we observe the occupancy of all hotels in the market. We show that it is not possible to identify the demand for competing hotels and the outside good when we only have occupancy data for a single hotel in the market: this model is only partially identified. Yet data on a single hotel is sufficient to identify its own demand function when we also have data on the prices charged by its competitor, and this is sufficient to conduct a number of interesting counterfactuals such as calculating the optimal pricing strategy for this hotel.

However in order to test more advanced hypotheses, such as whether the overall pricing by the firms in the market is best described as a Bertrand-Nash equilibrium, or to conduct certain counterfactuals, such as predicting the effect of hotel mergers or collusion by the hotels in the market, it is crucial to identify the joint demand function for all of the hotels as well as the demand for the outside good. If this is possible, it opens up a potentially powerful new avenue for econometrics to be useful for policy making, by enabling us to test hypotheses about firm behavior without imposing them *a priori.* For example, Harrington (2017) and Ezrachi and Stucke (2016) raise the specter of "algorithmic collusion" by sophisticated revenue management systems (RMS). The nature of "deep learning" algorithms from the artificial intelligence and reinforcement learning literatures makes it difficult to actually inspect the computer code used by commercial RMS to determine if it been explicitly designed to collude, or whether the algorithms "learn to collude" through repeated interaction. As we noted earlier, there is strong co-movement of prices and occupancy rates in the hotel market we analyzed, and some analysts might interpret such co-movement as a telltale sign of algorithmic collusion. Further, a structural model that imposed the hypothesis of collusion may result in distorted estimates of demand to help "rationalize" the maintained assumption of collusion. If it is possible to estimate demand without imposing strong assumptions about the type of equilibrium in this market, we can use the estimated demand model to solve for equilibrium under different equilibrium concepts, such as collusive pricing or Bertrand (competitive, non-collusive) pricing, and compare the predicted behavior to non-parametric estimates of the pricing strategies firms are actually using in this market. This may allow us to reject the hypothesis of algorithmic collusion in favor of a model

of competitive, Bertrand pricing where the price and occupancy co-movements are a natural response of a competitive market with inelastic supply of rooms that is subject to variable demand shocks.

Revenue management systems are proprietary so we do not know what sort of optimization principles they use and what types of data and econometric methods they employ. McAfee and te Veld (2008) note that "At this point, the mechanism determining airline prices is mysterious and merits continuing investigation because airlines engage in the most computationally intensive pricing of any industry." (p. 437). For reasons that are unclear, the RMS industry seems to have largely ignored econometric modeling and the substantial literature on demand estimation in the industrial organization litearture. Phillips (2005) notes that "The tools that pricers use day to day are far more likely to be drawn from the fields of statistics or operations research than from economics." (p. 68) and he credits marketing (which he regards as a subfield of operations research and management science) from bring "some science to what was previously viewed as a 'black art"' (p. 70). Yet "there remains a gap between marketing science models and their use in practice. The reasons for this gap are numerous. Many marketing models have been build on unrealistically stylized views of consumer behavior. Other models have been build to 'determine if what we see in practice can happen in theory.' Other models seem limited by unrealistically simplistic assumptions." (p . 70).

Our study can be viewed as an attempt to show that econometric literature on demand estimation may have value to the RMS industry and may shed light on the optimality of the price recommendations from these systems. Our empirical analysis of hotel pricing demonstrates that it is possible to identify a realistic stochastic model of hotel demand that relaxes equilibrium and optimality assumptions, and therefore provides a way to test rather than assume them. We focus on a single hotel that uses the IdeaS^{TM} RMS, a subsidiary of the SAS satistical software company. This hotel, which we refer to as "hotel 0" due to a non-disclosure agreement that prevents us from revealing its identity, follows the price recommendations of the IdeaS RMS approximately 60% of the time. On other occasions the revenue manager at hotel 0 deviates and chooses her own price. Though we are unable to observe which prices are the IdeaS recommended prices and which are set by the human revenue manager, we do know the information hotel 0 uses to set its prices, including the information in its own reservation database and real time information on the prices of its competitors from the *Market Vision^{TM}* pricing service. Hotel 0 cannot access the reservation databases of its competitors, and it generally does not know their occupancy rates, either *ex ante* (i.e. the number of rooms booked so far) or *ex post* (the actual or realized occupancy on a day by day basis). Though we show that it is only possible to partially identify the demand model parameters when we do not observe occupancy of competing hotels, our ability to identify overall demand, the distribution of arrivals, and the

fraction of consumers choosing the outside good was greatly assisted by auxiliary data we obtained from the company STR Global which collects price and occupancy data on over 63,000 hotels worldwide. The augmented data set enables us to identify consumer preferences and the distribution of arrivals and thus to fully construct the probablity distribution of demand in this market.

Using the estimated demand model parameters, we calculate counterfactual optimal and equilibrium dynamic hotel pricing strategies. In essence, our econometric demand model and optimization algorithm constitute our own "RMS" and prediction of optimal recommended prices. We find that the optimal prices from our model deviate significantly from the prices that hotel 0 actually charged. As a result, we are also able to strongly reject the hypothesis that pricing of the hotels in this market is consistent with Bertrand-Nash equilibrium. Given our relatively inelastic demand estimates and relatively low substitution to the outside good, we predict that *all* hotels in this market should be pricing significantly higher. Overall, we conclude that far from engaging in "algorithmic collusion" we can reject the hypothesis that hotel 0 is even using an optimal pricing strategy (i.e. a best response to its competitors). Thus we can also reject the hypothesis that the firms in this market are setting prices in accordance with a dynamic Bertrand-Nash equilibrium. This conclusion is broadly consistent with Herbert Simon's work on satisficing behavior by firms. Since we do not observe which prices charged by hotel 0 are those recommended by the IdeaS RMS and which were chosen by its revenue manager, we cannot determine whether the source of hotel 0's suboptimality is due to its RMS or decisions by its human revenue manager.

There have been a number of claims that commercial RMS (known as "yield management systems" in the airline industry) lead to significant improvements in profitability. Gallego and van Ryzin (1994) claim that "The benefits of yield management are often staggering; American Airlines reports a five-percent increase in revenue worth approximately $1.4 billion dollars over a three-year period, attributable to effective yield management." (p. 1000). However, we are unaware of studies that provide scientific validation (say via controlled experiments or other means) of the claims that commercial RMS have resulted in significant increases in hotel revenues and profits. Our study can be regarded as providing one of the first independent evaluations of the pricing strategy of a particular hotel, though further testing including field experiments would be required to confirm that the gains we calcluated from stochastic simulations of our econometric model could be realized in practice.

The general approach developed in this paper, i.e. of using a semi-parametric estimator to estimate demand parameters while relaxing the assumption of optimality, has been used previously. Hall and Rust (2019) studied the pricing and inventory investment decisions by a firm that trades ("speculates") in the steel market. In their application, they had to confront the econometric problem of "endogenous

sampling" due to censoring in the wholesale prices of steel that the the firm buys steel at. That is, the firm only records the price of steel on days it purchases steel. They compared a dynamic structural model of optimal steel price speculation with an "unrestricted" model that relaxes the assumption of optimality. Using the Method of Simulated Moments (MSM, McFadden (1998)), they showed that it is possible to consistently estimate the parameters of the wholesale price process by censoring simulated data for the firm in the same way that actual data are censored. Using a Hausman test, they were able to test and reject the assumption that the firm's steel purchases were governed by an optimal (S,s) inventory investment strategy. Using stochastic simulations, they also showed that if the firm had adopted an optimal (S,s) strategy it would have earned significantly higher trading profits over the sample period.

Section 2 summarizes the key features of our data set on hotel pricing, providing a concrete illustration of the price endogeneity, truncation and censoring problems we face. Section 3 discusses the identification of demand in a simplified static setting where the key ideas underlying our approach to identification can be explained more clearly. Section 4 applies our model to the hotel market and provides estimates of demand both for the mixed censored multinomial model introduced in section 3 but also for a linear demand model that does not attempt to identity the probability distribution of arrivals, the fraction or consumers choosing the outside good, or derive demand from a microaggregation of individual discrete choices. While we show that in-sample, the estimated expected demand functions from this simpler linear model and our structural mixed censored multinomial demand model are quite similar to each other, the out-of-sample predictions of the model, particularly under collusive scenarios are quite different. We believe this is due to the inability of the linear model to adequately capture substitution to the outside good as firms collectively raise their prices. Section 5 provides some conclusions and suggested directions for future research.

2 Hotel Data

This section describes the hotel market that motivated the questions about demand estimation and identification that we attempt to answer in this paper, and in particular the simple static model of hotel demand that we introduce in section 3. As we noted in the introduction, due to a non-disclosure agreement with the hotel that provided the data for our study, we are unable to provide too much detail about the local market in which hotel 0 operates to guarantee the anonymity of the hotel and the owner. We can say that it is a luxury hotel located in a highly desirable downtown location of a major US city.

Hotel 0 is one of seven luxury hotels operating in a well defined local market that is recognized by online travel agencies (OTAs) and other travel agents. Though customers can book at other luxury hotels

Table 1: Hotels in the local market in our study

Property	Avg. BAR	Star	Class	Chained Brand	Rate	Relative Capacity	Distance to mass transit	Cancel Policy
hotel 0	$ 293.26	4	Luxury	No	4.4	79%	3 min	1 day before
hotel 1	$ 282.64	4.5	Upper Up	No	4.4	81%	5 min	3 day before
hotel 2	$ 285.16	4	Upper Up	No	4.4	63%	3 min	1 day before
hotel 3	$ 338.29	4	Upper Up	Yes	4.2	99%	8 min	2 day before
hotel 4	$ 397.09	4	Luxury	No	4.6	100%	10 min	Strict
hotel 5	$ 253.51	4	Upper Up	No	4.2	47%	8 min	3 day before
hotel 6	$ 454.30	5	Luxury	Yes	4.7	52%	10 min	1 day before

in other parts of this city, the locations of these other luxury hotels are sufficiently far from this particular desirable area that they are not regarded as relevant substitutes for customers who wish to stay in this specific area of the city. We consider any choice of another hotel outside the seven hotels in this local hotel market, including the decision not to stay in any hotel, as a choice of the outside good.[4]

Table 1 lists some summary information about the seven hotels: all are 4-star or higher rated luxury hotels. To avoid identifying the hotels we show only the relative capacity, where we normalize the capacity of the largest hotel to 1. However our model uses all relevant information including the actual capacity, which we will show is an important factor in hotel pricing. The customers of the hotel are both business/government customers who mainly stay in the hotel on weekdays and tourists who typically stay on weekends. Since business customers and government customers are reimbursed for their travel expenses, we can expect them to be more price inelastic than tourists. On the other hand, many government agencies and large corporations that do frequent business in this city have negotiated government and corporate discounted rates with this hotel. These discounted rates are typically a fixed percentage, often 15 to 20%, off the currently quoted price that is called the *best available rate* (BAR). The revenue manager of hotel 0 is in charge of updating an array of BARs for different room classes and different future arrival dates and posting these prices to the web via the Global Distribution System (GDS, a network of computer connections that give travel agents access to a hotel's reservation database to check availability and reserve rooms) and via its own website.

Customers generally book hotel rooms in advance, and generally only a small fraction of customers (approximately 8%) book their rooms on the same day that they intend to occupy them. Hotel 0's reservation database provides information on when each hotel room was booked and through which *channel.* A

[4]There is also limited capacity of private residences in this area, so alternatives such as AirBnB is a minor factor in this market, and so we also lump this option into the catch-all category, "outside good."

hotel room can be booked over the phone, via hotel 0's website, via a traditional travel agency, or via an *online travel agency* (OTA) such as Priceline or Expedia. The decentralized nature by which consumers search for hotels and book rooms implies that there is no single site or source that observes all consumers who "arrive" to book a room in a particular market during a particular point in time. A hotel will know how many consumers have booked one of its own rooms at every possible future arrival date, but there is no entity that observes all consumers searching for rooms or the number of bookings made in all of the competing hotels in a given market for arrival at any given future date. So this is the sense in which *arrivals are unobserved.*

Similarly, no single entity will know how many of the customers who arrive to book a room in a market will choose the outside good, i.e. to either choose to stay at a hotel outside this market or other form of accommodation such as AirBnB. Thus, from hotel 0's perspective, suppose that it received 10 bookings on a particular day. Hotel 0 cannot distinguish between situations a) and b):

a) 100 customers arrived and 10 of these customers chose hotel 0, 50 of them chose one of its competitors, and the remaining 40 chose the outside good

b) 70 customers arrived, 10 booked at hotel 0, 50 booked at one of its competitors and only 10 chose the outside good.

In both cases a) and b) hotel 0 observes only the 10 customers who booked one of its rooms, but it has no information on the number of customers arriving in total, the number choosing the outside good, or even the number of customers who book a room at one of its competitors. In view of this, it is very difficult for a hotel to determine its market share on any particular day. Hotels do, of course, observe each others capacities and they can obtain (at a cost) historical data on occupancies of their competitors from firms such as STR, but hotel 0 does not have real time information on the bookings and occupancies of its competitors.

Hotel 0's revenue manager uses a uniform price strategy and does not sell blocks of rooms to wholesalers under contracts that give wholesalers discretion to set their own prices for the blocks of rooms they purchase. Thus, there is no ability to "arbitrage" prices of rooms for hotel 0 by searching different OTAs. However hotel 0 does pay a significant commission, ranging from 15 to 25%, for reservations that are made via OTAs such as Expedia. The GDS that hotel 0 uses allows the revenue manager to change prices as frequently as she desires, though there is a short lag before the prices are propagated everywhere on the Internet including the leading OTAs. However for hotel 0's own website and reservation system, price changes take place instantaneously, and hotel 0 has its own loyalty program that provides discounts to

customers who are members of the program. There are other groups that include weddings that involve a larger group of guests that are typically individually negotiated with the hotel revenue manager, but the discounts to these groups are typically quoted as a percentage discount off the BAR similar to corporate and government contract rates.

As we noted in the Introduction, hotel 0 subscribes to the IdeaS RMS that provides recommended prices. The hotel revenue manager uses her own discretion to select a relatively small number of different possible BARs (effectively, she discretizes the pricing space) which are treated as a predefined choice set that is entered into the RMS. Based on a proprietary algorithm that considers remaining availability, seasonal effects, cancellation rates and competitors' prices, the RMS communicates a recommended BAR to the revenue manager at the start of each business day. Even though the revenue manager has some control over the prices the RMS can recommend via her choice of a predefined finite set of possible BARs, she often ignores the recommended price from the RMS and instead sets her own BARs based on her own experience, judgement and intuition. Unfortunately, our data do not specify which prices were ones recommended by IdeaS and which are ones she set herself, but she told us that she estimated she used the recommended prices approximately 60% of the time.

Thus, we do not know to what extent the IdeaS RMS is able to observe and adapt to the knowledge that the revenue manager is disregarding their recommended prices. This would seem to be important information that any RMS would want to collect, including the revenue manager's feedback about the overall quality of the recommended prices from the system. We can imagine that manual "price overrides" are common for newly launched hotels where the RMS may initially not have enough data to form good predictions about demand, or when there are unexpected changes to demand or entry/exit of other hotels in the local market. In these cases we might expect that the recommended prices from the RMS would be less trustworthy until sufficient data are accumulated to enable the RMS to provide an updated model of customer demand that provides accurate predictions for the local market in question.

Hotel 0 provided us information from its reservation database that enabled us to track all bookings, cancellations, and prices for a 37 month period between September 2010 and October 2013. In addition, we were provided aggregate daily reports and their competitive daily rates of hotel 0's six competitors from a service called *Market Vision* provides quotes from hotel 0's six competitors for several room rate categories several times per day. While Market Vision provides excellent data on prices, as we noted above, it provides no information on the number of bookings or occupancies at hotel 0's competitors. This information does not seem to be readily available, but we were able to obtain data on the occupancy of hotel 0's competitors on a daily basis thanks to data provided by STR. Table 2 summarizes the data

Table 2: Data sources used in this study

Data	The first day of occupancy	The last day of occupancy	Observations	Description
market vision	2010-09-21	2014-08-13	609,181	competitors' price
reservation raw	2009-09-01	2013-10-31	201,176	reservations detail information
cancellation raw	2009-09-01	2013-10-31	29,241	cancel detail information
daily pick-up report	2010-09-16	2014-05-21	475,187	daily revenue report
STR market data	2010-01-01	2014-12-31	1,731	competitors' occupancy
Data range	**2010-10-01**	**2013-10-31**		**37 months**

sources we used for our study.

Our data are unique in the level of detail we have on reservations and cancellations. Our reservation database contains the full history of each individual booking, including the channel through which the booking was made. Each booking is identified with a unique reservation identification number that is created when the reservation is initiated and becomes the permanent identifier for each reservation along with time stamps and dates of arrival and departure and amounts actually paid including incidental charges.

Hotel 0 has two basic categories of rooms: regular rooms and higher tier rooms such as luxury suites, but 95% of the rooms in the hotel are regular rooms. Thus we focus on the demand and prices of regular rooms. We rarely observe overbooking of the regular rooms, though on the few occasions where this happens the overflow customers are automatically upgraded to a room in one of the the higher tiers.

There are around 200 rate codes which can be broken into 14 categories. To simplify our analysis, we divided the codes into two; transient and group bookings. Transients are individual travelers who pay the BAR or discounted BAR. Although the net of commission price that hotel 0 receives differs depending on which channel was used to do the booking (i.e. an OTA versus hotel 0's own website), transient customers themselves pay the same price regardless of channel, namely the BAR in effect at the time they booked. Group bookings are also generally based on the BAR in effect when they booked, however it will vary by pre-negotiated contract discount rate that differs from different groups (rate codes).

Another commonly analyzed price in the hotel industry is the *average daily rate* ADR: the ratio of the total gross revenue collected each day divided by the total number of rooms booked (including no-shows who are generally charged for their rooms). Although the data for hotel 0 provides an incredible level of detail, we need data on room reservations at hotel 0's competitors that are not provided in the Market Vision data, which provide only competitors' prices. As we will see, information on the total number of consumers who "arrive" and book rooms at one of the seven hotels in this local market is critical for

our inferences about customer demand, and especially how customers respond to daily fluctuations in the relative BARs of the seven competing hotels. Unfortunately we do not have access to the reservation databases of hotel 0's competitors, so we cannot observe the total number reservations that are made at all of the hotels and at which prices (including group, corporate discounts, etc).

Fortunately we were able to obtain this information from STR via an academic research contract it has with Georgetown University. In addition to total occupancy at each competing hotel on a daily basis, the STR data provide information on the competitors' ADRs and total revenue. The STR data turn out to be crucial for our ability to estimate a credible demand model. Unfortunately, the STR data do not identify the individual occupancy and ADRs of hotel 0's competitors: it only provides the aggregate occupancy at the 6 competing hotels on a daily basis. Therefore, we will simplify our analysis by treating these 6 competing hotels as a single aggregate competitor, which we will refer to as "hotel c."

2.1 Data summary

As we noted in the introduction, there is strong co-movement in prices and occupancy rates in this market. Figure 1 illustrates the cyclical pattern of occupancy and prices, both over a given week and over the year, reflecting seasonal variations in the demand for hotels. The bars in the left hand panel of figure 1 show a typical weekly cycle of occupancy for hotel 0 where the lowest occupancy is on Sunday, but a peak occupancy on Saturday, and a midweek peak occupancy on Tuesdays and Wednesdays. The ADR peaks on Tuesday, and the higher rates during the weekdays reflects price discrimination for less price elastic business guests, whereas the lower rates on Fridays and Saturdays are designed to attract more price elastic tourists. Occupancy is lowest on Sundays when tourists are checking out to return home for work on Monday, whereas a typical business guest checks in during the middle of the week and departs before the weekend. The right hand panel of figure 1 shows the price and occupancy dynamics over the year. Occupancy rates are the highest in the spring and early fall, and are lowest around holidays such as Thanksgiving, Christmas and New Year's. The black line in the figure plots hotel 0's ADR and total revenues, and we see that both of these move in sync with the ups and downs in occupancy rates. This suggests that prices and revenues at hotel 0 are highly demand driven.

Figure 2 compares the price dynamics for hotel 0 to those of its six competitors over the year. It plots the weekly average BAR from October 2010 to October 2013 for same-day reservations using the Market vision data, though we would obtain similar results if we plot a time series of ADRs using the STR data.[5] The bold line plots the average BAR of hotel 0 while the other lines are the BARs of its six

[5]Since we have BARs for each of the 6 competiting hotels, we prefer to plot these more detailed data: as we noted above, the STR data does not allow us to determine hotel-specific ADRs, only an average ADR for all of hotel 0's 6 competitors.

Figure 1: Booking and price dynamics over the week and year

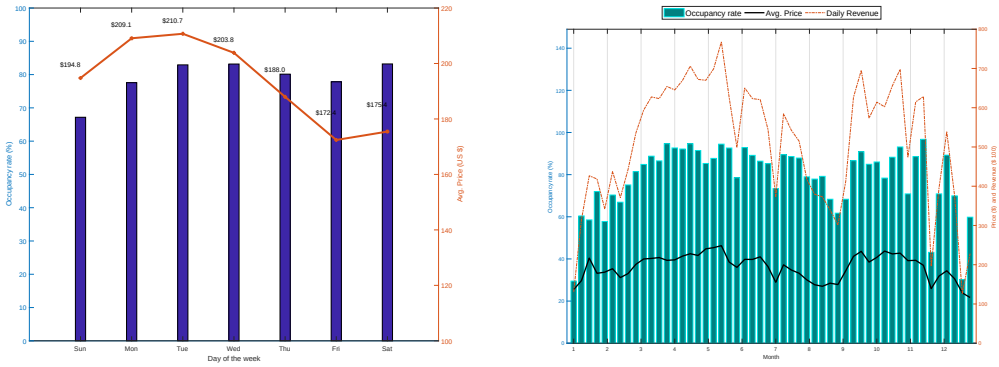
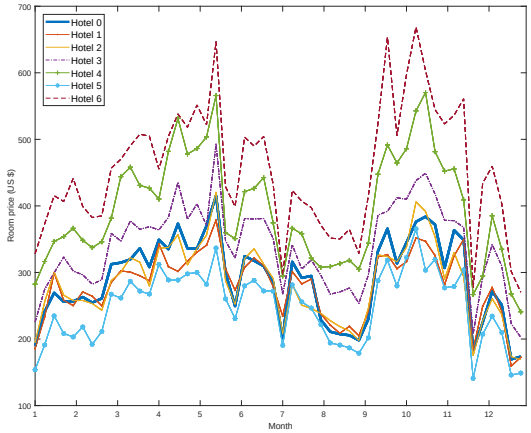


Figure 2: Annual price dynamics for all seven hotels



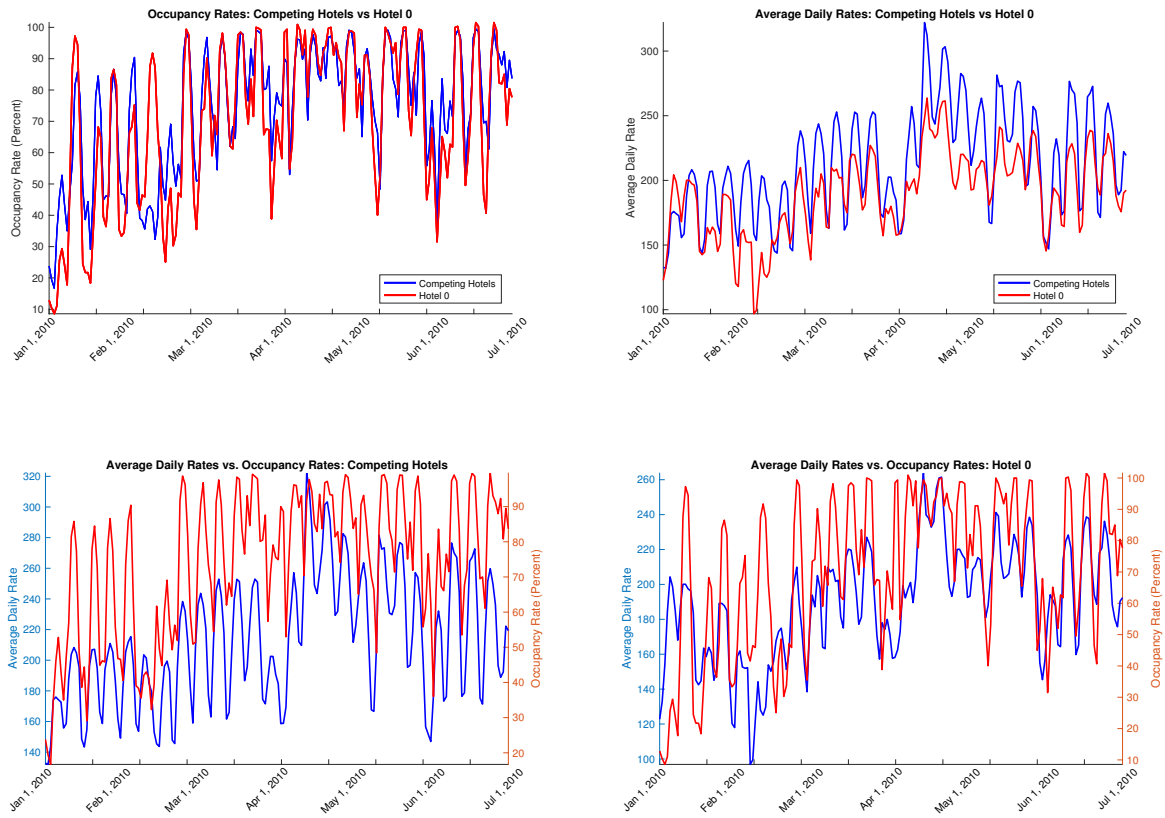
competitors. We see strong co-movement in the prices of the seven hotels, and that they follow similar cyclical fluctuations, and hotel 0 generally underprices its competitors with the exception of hotel 5. There is strong seasonality in prices which are highest in the spring and the fall with peaks in early May and mid-September and October. Prices are lowest at the key holidays: Thanksgiving, Christmas, New Year's, as well as early July and August. During peak periods the average BAR of hotel 0 can be over $350 per night, whereas in the lowest periods it averages about $200.

The pattern of co-movement in the prices in this market might be described as "price following" and given the fact that most hotels use RMSs and have extensive knowledge of their competitors' prices from services such as Market Vision, it could raise concerns about the possibility that RMSs enable these hotels to engage in algorithmic collusion. The price troughs following price peaks might be interpreted as "price wars" that are designed to punish hotels that deviate from the recommended prices that are highest when

prices are peaking. However, we do not think this is the correct way to interpret these price patterns.

Figure 3 plots the time series of ADRs and occupancy rates for all seven hotels in this market for the first half of 2010 using the STR data. The top left panel plots the occupancy rate for hotel 0 versus the occupancy rate of its competitors, where the competitor occupancy rate is defined as the total occupancy at the six competing hotels divided by the total room capacity of those hotels. With few exceptions, we see that occupancy follows the same weekly cycle at all of the hotels that we illustrated in the left panel of figure 1 for hotel 0, as well as the seasonal fluctuations (i.e. higher in the spring but lower at end of June) that we observed in the right panel of figure 1. The top right panel of figure 3 shows that all seven hotels also have strong weekly cycles in their ADRs and the reasons are likely to be much the same as we conjecture for hotel 0: higher mid-week prices to discriminate against less price elastic business guests and lower weekend rates to try to attract the more price elastic tourists.

Figure 3: Co-movement in ADR and cccupancy rates for all seven hotels



The lower two panels of Figure 3 plot the cycles in occupancy rates (red lines) and ADRs (blue lines) for hotel 0 (right hand panel) versus its competitors (left hand panel). The data suggests that the weekly

price cycles are driven not only by different compositions of guests (business versus tourists) but also to ration scarce capacity, since these hotels tend to be fully booked midweek but not on weekends. Both hotel 0 and its competitors follow similar weekly occupancy and price cycles, as well as similar seasonal price/occupancy cycles. For example we see that ADRs for both hotel 0 and its competitors peaked in mid April 2010, during a period where occupancy was close to 100% both mid-week and on the weekends.

It is natural to ask the question: which motive is more important for hotel 0? That is, does the revenue manager increase prices mainly to ration scarce capacity, or to try to exploit the more inelastic demand of business travelers who are more likely to be staying in the hotel midweek? Or, is hotel 0 simply following the prices of its competitors? If so, is this price following behavior a sign that all of the hotels are following the recommended prices from their RMS, and could this be evidence of tacit collusion mediated by the RMS?

Table 3 provides some insight into this question by presenting the results of a simple OLS regression of the logarithm of hotel 0's ADR on the average ADR of its six competitors and on its own and competitors' occupancy rates. This simple model results in an R^2 of 86% when we also add dummies for different days of the week and months of the year to capture the weekly and seasonal price cycles.

Note that the occupancy also affects hotel 0's pricing but in a counterintuitive fashion: hotel 0's occupancy rate has a negative coefficient, but the occupancy rate of its competitors has a much larger positive coefficient. We may suspect that the co-movement in occupancy rates leads to a collinearity issue but hotel 0's own occupancy has a negative coefficient even after we remove the occupancy of the competing hotels from the regression. The coefficient estimate for Hotel 0's own occupancy rate only turns positive when we remove the ADR of the competing hotels, but then the fit of the model drops precipitously, to an R^2 of 0.17.

Table 3: Ordinary least squares regression with dependent variable ADR_0

Variable	Estimate	Standard Error
constant	27.93	2.24
ADR_c	0.73	0.01
OCC_0	-0.09	0.027
OCC_c	0.273	0.044
$N = 1277, R^2 = 0.83$		

The regression findings suggest that the effect of occupancy on hotel 0's pricing decisions are second order relative to the dominant effect of the prices set by its competitors. To a first approximation, hotel 0 sets its prices at 70% of the average of its competitors' prices. The R^2 drops to 0.69 when we remove

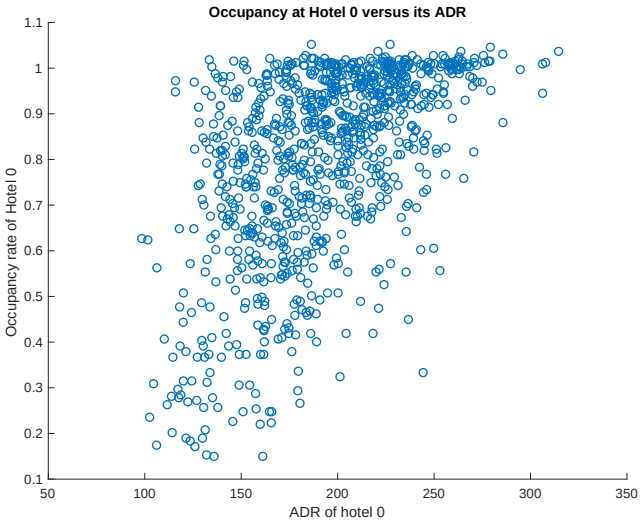
ADR_c from the regression but retain occupancy variables and daily and seasonal dummies. Overall, the regression results suggest that the revenue manager is setting prices in accordance with a "price following" strategy, and that knowledge of her competitors' prices is the most important piece of information besides the day of the week and the season of the year that she uses to set her own prices. The fact that hotel 0's own occupancy appears to have only a second order effect on its price setting once we condition on the prices of competitors suggests that raising prices to ration scarce capacity is not an important motive for hotel 0.

On the other hand it is not clear whether the fact that hotel 0's behavior is well approximated by "price following strategy" is evidence in favor of "algorithmic collusion" that Ezrachi and Stucke (2016) and Harrington (2017) discuss. Even if demand for rooms cycles in a systematic way during the week versus weekends, it is not clear that collusive prices would necessarily follow the same cyclical pattern that we observe in this market. In particular, we would expect that if the hotels in this market operated as a cartel, their prices would rise sufficiently high that there would be excess capacity even during the peak weekday periods, and the excess capacity would serve in part as a credible threat to engage in a price war that would deter any of the hotels that contemplated deviating from the collusive recommended prices, see Benoit and Krishna (1987) and Davidson and Deneckere (1990).

An alternative hypothesis is that this market is best approximated by a dynamic competitive equilibrium in a market characterized by strong Bertrand price competition subject to fixed capacity constraints. Stochastic shocks to demand lead to the price cycles we observe, with prices peaking to ration the available capacity in periods where demand exceeds available supply, but prices falling significantly as predicted by Bertrand price competition in periods of low demand where there is excess capacity. In this paper we will argue that the latter explanation is more likely to be closer to the truth, especially given what we have already reported about hotel 0's disinclination to follow the recommended prices of its RMS, combined with the fact that the revenue manager believes that the recommended prices are too low.

Regardless of the interpretation, the strong co-movement of hotel 0's prices with the prices of its competitors creates real difficulties for demand estimation. We can see the problem in figure 4, which is a scatterplot of hotel 0's own occupancy against its ADR over the period of our sample. This figure shows an *upward sloping* relationship between price and occupancy, which encapsulates the endogeneity problem we discussed in the introduction. We believe the endogeneity is emblematic of the classic Cowles Commission simultaneous equations type of endogeneity between prices and quantities. If the market is well approximated as a Bertrand equilibrium but subject to large stochastic demand shocks, then we would expect to see high prices that ration demand given the finite hotel capacity when demand is high but we

Figure 4: Price scatterplots for hotel 0 and its competitors



observe low prices when the hotels compete for the available demand in periods where there is excess supply of rooms. This will generate a positively sloped scatterplot of prices similar to what we observe in figure 4 and generally a positively sloped relationship between ADRs and occupancy for each hotel individually. Thus, simple OLS regressions will result in *positively sloped demand curves* in this market.

There are no obvious instrumental variables that can solve this endogeneity problem. One possible instrumental variable is a decrease in capacity of the hotel. If we regard the hotel as setting prices to ration demand, then in periods where there is a reduction in available rooms for exogenous reasons (such as a bursted pipe or other problems that remove rooms from service temporarily, or planned upgrades to rooms that take rooms out of service for a period of time or permanently, such as when the hotel converted 23 of its standard rooms to deluxe rooms), then the decrease in supply of rooms may serve as exogenous supply shifter that might allow us to estimate a negatively sloped demand curve. Unfortunately when we tried to use available capacity as an instrument we find highly unreliable and generally insignificant results. There is not enough exogenous variation in hotel 0's available capacity to make this a good instrument for estimating the effect of hotel 0's price on demand.

3 Identification of a Static Model of Hotel Demand

We analyze the identification problem in the context of a simple static model of hotel demand, i.e. a model where there are no advance bookings of hotel rooms. The static setting allows us to illustrate our approach and key issues in a simpler model with less notational complexity. We focus on a particular local hotel

market and assume that the L hotels in this market do not take advance bookings but rather on each day t the hotels set prices (simultaneously), then a random number of customers arrive and choose which hotel to stay at, after observing the vector of prices p_t that the hotels set that day. One of the options available to all consumers is the "outside good" i.e. to not stay in any of the hotels. We develop a model that is rich enough to reflect the truncation, censoring and endogeneity problems that we observe in our hotel data set, and thus provides a simple initial framework to convey the basic ideas underlying our approach.

The fundamental challenge is one of identification: the econometrician does not observe $\tilde{A}$, the number of customers arriving to book rooms, nor does the econometrician observe the number of arriving customers who choose the outside good. In the statistics literature this is known as a problem of *truncation.* We also face a related problem of *censoring:* at each hotel we only observe the minimum of its actual demand for rooms and the hotel's room capacity. A further data problem is that the econometrician may not observe the occupancy rates at competing hotels. We also consider the possibility of *unobserved heterogeneity* i.e. the econometrician does not observe an individual's preferences for the various hotels or their degree of price sensitivity, which is crucial information necessary to determine the customer's willingness to pay to stay in various hotels in this market.

The econometrician can observe the vector of prices p chosen each day by the hotels, and the realized occupancies, d, (an $L \times 1$ vector) though as we noted, these occupancies are censored at the capacities C (also an $L \times 1$ vector) at each of the hotels. When a hotel reaches capacity it turns customers away. In reality when a hotel reaches capacity and has to turn customers away, they may choose the outside good or a competing hotel in this market that has availability. However to keep our analysis simpler, we will assume all rationed customers choose the outside good, and with probability 1 the hotel capacity constraint $d \leq C$ is enforced. Let d_t be the total number of consumers who are booked in one of the hotels on day t where e is an $L \times 1$ vectors of 1's. We have with probability 1, $\tilde{A}_t \geq d_t' e$, and the difference, $\tilde{A}_t - d_t' e$ is the number of customers who end up with the outside good, either via a voluntary choice, or because the hotel they chose was full.

We assume the hotels (and the econometrician) observe a vector of demand shifters x_t that helps them predict the number of arrivals A. There may also be other idiosyncratic variables z_t that are specific to each hotel that each hotel observes, that affect their pricing decision. We assume the econometrician does not observe the vector z_t, which we assume for simplicity is an $L \times 1$ vector where the l^{th} component z_{tl} represents a scalar idiosyncratic shock that is observed only by hotel l and reflects factors specific to hotel l how sets its price that the other hotels (and the econometrician) do not observe. Thus x_t is common knowledge among the hotels, but each shock z_{tl} is private information observed only by hotel l, though

we place no restriction on the correlation between z_{tl} and $z_{tl'}$ for $l \neq l'$.

The timing of decisions in each market day t are as follows:

1. At the start of the day, all L hotels observe the demand shifter x_t that help them predict the number of customers who will arrive to book rooms later that day. Each hotel l also observes idiosyncratic factors that affect its own pricing decision, z_{tl}, but not the idiosyncratic shocks observed by each of its competitors $\{z_{tl'}\}$ for $l' \neq l$. Customers may observe x_t but no customer nor the econometrician, observes the idiosyncratic shocks z_{tl}, $l \in \{0,\ldots,L-1\}$.

2. Based on the information (x_t, z_t) the firms set their prices, so we can write $p_t = p(x_t, z_t)$ is the price set at the start of day t prior to the arrival of customers. The price set by hotel l depends only on its own idiosyncratic realization z_{tl}, so the pricing rule for hotel l is independent of $\{z_{tl'}\}$, $l' \neq l$ and so can be written as

$$p_{tl} = p_l(x_t, z_{tl}), \quad l \in \{0,\ldots,L-1\}. \tag{1}$$

   Hotel prices also reflect common knowledge of an $L \times 1$ vector ξ of characteristics of each of the hotels that constitute attributes of each of the hotels that affect customer preferences that the hotels and customers observe but the econometrician does not observe. We assume that customers observe ξ and these characteristics affect their preferences for the hotels. Similarly the hotels' perception of customer preferences in turn affects their pricing decisions. However we do not include the unobservable variables ξ as an explicit argument of the pricing function (1), though our model does allow prices to be an implicit function of their perceptions of consumer preferences which may in turn depend on the time-invariant unobserved attribute vector ξ.

3. Customers observe the demand shifter x_t and the characteristics of the competing hotels (and the outside good) ξ, but the number of customers arriving to book a room in hotel market is a random process that does not depend on prices p_t given x_t. We will let $H(A|x)$ be the distribution of consumers who arrive when market conditions are summarized by the observed demand shifter x but assume that the distribution of arrivals does not depend on ξ or prices p since customers only learn about ξ and p once they arrive at the market and consider the available alternatives.

4. We assume that each customer who arrives on day t to book a room observes ξ and the price vector p_t and chooses to stay in one of the L hotels based on a simple static utility maximization decision, where the hotel chosen by customer $a \in \{1,\ldots,\tilde{A}\}$ can be any of the L hotels or $l = \emptyset$ denotes the choice of the outside good (to not stay in any of the L hotels and go to some other hotel outside of this local market). We assume there are a finite number of possible types of consumers indexed by

τ and there are *IID* random utility shocks that affect each consumer's choice of which hotel to stay at. We assume that the choice of hotel by consumer j on day t, l_{tj}, is given by

$$l_{tj} = \underset{l \in \{\emptyset,0,1,\ldots,L-1\}}{argmax} [u_\tau(l, p_{tl}, x) + \varepsilon_{tj}(l)] \tag{2}$$

The utility each consumer obtains from the different hotels is an implicit function of the hotels' attributes ξ, so in this sense, the set of consumer utility functions $\{u_\tau\}$ is a sufficient statistic for the set of hotel attributes ξ and in order to set prices hotels should have a good knowledge of consumer preferences. Knowledge of how hotel attributes ξ affect customer preferences is only relevant for longer term decisions by hotels, such as investment in hotel upgrades, etc.

5. We assume the $(L+1)\times 1$ vectors of the random utility components $\varepsilon_{tj} \equiv \{\varepsilon_{tj}(l) | l \in \{\emptyset, 0, 1, \ldots, L-1\}$ are continuously distributed with unbounded support over R^{L+1} and are independently distributed across different customers who arrive in this market, the behavior of a consumer of type τ can be represented by a conditional choice probability $P_\tau(l|p,x)$ which provides the probability that the consumer will choose hotel l (or the outside good if $l = \emptyset$) given the price vector p when the observed demand shifter is x. Since choices of different customers are made independently of each other, the total *potential demand* by the A customers who arrive on t given by a multinomial distribution with parameters $(A_t, \pi_\emptyset(p,x), \pi_0(p,x), \ldots, \pi_{L-1}(p,x))$ where

$$\pi_l(p,x) = \sum_{\tau=1}^{\mathcal{T}} P_\tau(l|p,x) g(\tau|x), \quad l \in \{\emptyset, 0, 1, \ldots, L-1\} \tag{3}$$

where $g(\tau|x)$ is the fraction of consumers of are of type τ on a day where the observed demand shifters equal x.

6. Let $f(d|A,p,x)$ be the conditional distribution of *realized occupancy* in the L hotels, where $d = (d_0, d_1, \ldots, d_{L-1})'$ is an $L \times 1$ vector of occupancies in the L hotels that satisfies each hotel's capacity constraint, $d_l \leq C_l$, where $l \in \{0, \ldots, L-1\}$. This distribution is a *censored multinomial distribution* given by

$$f(d|A,p,x) = \begin{cases} M(d|A,p,x) & \text{if } d_l < C_l,\ l \in \{0,\ldots,L-1\} \\ \sum_{d' \in D^c(d)} M(d'|A,p,x) & \text{otherwise} \end{cases} \tag{4}$$

where

$$D^c(d) = \left\{ d' | d'_l \geq d_l \text{ if } d_l = C_l,\ d'_l = d_l \text{ if } d_l < C_l, \text{ where} \sum_l d'_l \leq A \right\} \tag{5}$$

where C_l is the room capacity of hotel l and $M(d|A,p,x)$ is the uncensored multinomial density of potential demand

$$M(d|A,p,x) = \frac{A!}{(A - \sum_{l=0}^{L-1} d_l)! d_0! d_1! \cdots d_L!} \pi_\emptyset(p,x)^{(A - \sum_{l=0}^{L-1} d_l)} \pi_0(p,x)^{d_0} \pi_1(p,x)^{d_1} \cdots \pi_{L-1}(p,x)^{d_{L-1}} \tag{6}$$

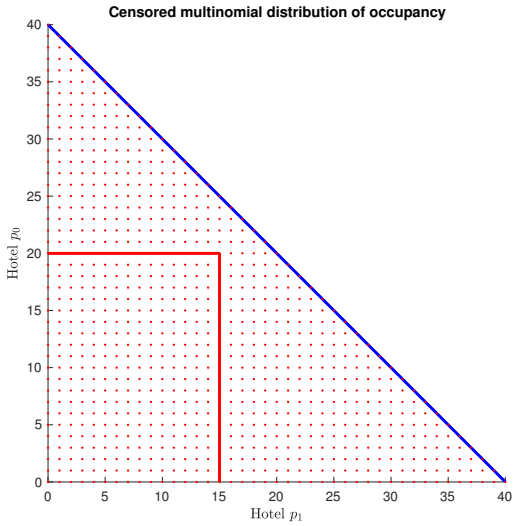


Figure 5: Censored multinomial distribution of occupancy

where $d_l \in \{0,1,\ldots,A\}$, $l \in \{0,\ldots,L-1\}$ and $\sum_l d_l \leq A$.

Figure 5 illustrates the rectangular support of the censored multinomial distribtion in the case of $L = 2$ hotels, one with a capacity of $C_0 = 20$ rooms and the other with a capacity of $C_1 = 15$ rooms. The triangular region illustrates the support of the uncensored trinomial distribution when there are $A = 40$ arriving customers. Thus, each pair of potential demands (d_0, d_1) satisfying $d_0 + d_1 \leq 40$ is in the support of the trinomial distribution, with the residual customers $d_\emptyset = 40 - d_0 - d_1$ taking the outside good. Realized occupancies (d_0, d_1) must statisfy the capacity constraints $d_0 \leq C_0$ and $o_1 \leq C_1$ with probability 1. For example the demand $(d_0, d_1) = (20, 20)$ is not feasible since the demand for hotel 1 is $d_1 = 20$ which exceeds its capacity, $C_1 = 15$. In such a case we assume the hotels serve customers on a first come, first served basis until their capacity is reached and all "excess customers" choose the outside good. Thus the realized occupancies equal $(20, 15)$ in this case, and the excess 5 customers who could not be accommodated at hotel 2 are assumed to choose the outside good. When we calculate the probability of any realized occupancy where one or more hotels is sold out, we sum the multinomial probabilities over all possible potential demands d' that are consistent with the specified hotels being at capacity, i.e. over all indices $d'_l \geq C_l$ for for those hotels l which are at capacity, $d_l = C_l$. For example the probability of the occupancy pair $d = (d_0, d_1) = (0, 15)$ is the sum over all potential demands $d' = (d'_0, d'_1)$ in the set $D^c(d)$ which is the set of all indices d' where $d'_0 = 0$ and $d'_1 \geq 15 = C_1$ and $d_\emptyset = 40 - d'_1$, i.e. the sum of the probabilities of all integer coordinates on the x axis of figure 5 from hotel 1's capacity of $C_1 = 15$ to the total number of arrivals, $A = 40$.

Below are the key assumptions on the timing of decisions, price setting, and demand that underlie our identification result.

Assumption 0 (Endogeneity) *Hotels set their prices prior to knowing the number of customers A who arrive to book rooms in the market, however due to the common dependence on x, prices p and arrivals A will generally be positively correlated, and thus also positively correlated with occupancy d. Hotel prices will also generally depend on the unobserved characteristics of hotels, ξ, which affect consumer preferences and willingness to pay for different hotels.*

Assumption 1 (Stationarity and Independence) *The pricing rule that hotels use to determine prices in equation (1) is time-invariant. The demand shifters $\{x_t\}$ that enter the pricing rule may be serially correlated, but the process $\{x_t, z_t\}$ is strictly stationary. There may be correlation between x_t and z_t and contemporaneous correlation between the components of z_t, i.e. if $l \neq l'$, then z_{tl} and $z_{tl'}$ may be dependent random variables for each t.*

Assumption 2 (Conditional Independence) *The censored multinomial conditional distribution of occupancy given in equation (4) is time-invariant and independent of z given (A, p, x). That is, we have:*

$$f(d|A, p, x, z) = f(d|A, p, x), \tag{7}$$

where $f(d|A, p, x)$ is the censored multinomial conditional distribution of hotel occupancy given in equation (4). The distribution of arrivals is also independent of p and z given x

$$H(A|x, p, z) = H(A|x). \tag{8}$$

Assumption 2 is the key to our semi-parametric identification strategy. The unobserved "price shocks" z create random variability in prices that enables us to identify the effect of price on demand for hotel rooms after controlling for x, which is the observable demand shifter that is the fundmental source of endogeneity in this model. We might also refer to Assumption 2 as *conditional exogeneity of prices* since conditional on x, the remaining variation in prices is random, so we have price variation that is akin to a randomized experiment that helps us to identify the effect of prices on consumer demand for hotels.

We also think of Assumption 2 as analogous to the conditional independence assumption in the treatment effects literature, where the assignment of a "treatment" is assumed to be independent of the potential outcomes, conditional on a vector of covariates x. Thus, assignment of a treatment can be treated as a virtual random assignment, given x in the sense that prices can be regarded as "randomly assigned" given x due to the effect of unobserved idiosyncratic factors z affecting how the hotels set their prices. The conditional distribution $G(p|x)$ captures the variability in prices due to the effect of the z shocks and the conditional distribution $H(A|x)$ reflects the uncertainty about the number of arrivals given only knowledge

of the observed demand shifter x. We believe it is plausible that $H(A|x)$ does not depend on p because customers are not able to learn about prices until they actually arrive at the market (e.g. visit a website to check prices). However this does not imply that arrivals are independent of p: there will be positive correlation between p and A and thus between p and realized occupancies d via the common dependence on the demand shifter x.

We make a final assumption about the variables the econometrician observes in this market.

Assumption 3 (Observables and Unobservables) *The econometrician is able to observe joint occupancy, d, the demand shifter x, and the vector of prices charged by the hotels, p. The econometrician does not observe arrivals, A, or the vector of pricing shocks, z.*

Note that Assumptions 0 to 3 are completely agnostic about how hotels set prices in the market, and in particular there is no assumption about hotels setting prices optimally or in a manner consistent with a price equilibrium, e.g. Bertrand Nash equilibrium. Let $G(p|x)$ be the conditional distribution over the prices set by the hotels given the observed demand shifter x that are induced by the idiosyncratic price shocks z. In effect, this distribution is an infinite dimensional "nuisance parameter" since our primary interest is to infer consumer preferences and the arrival probability $H(A|x)$.

With enough data on a given market, it is possible to non-parametrically estimate the distribution $G(p|x)$ and the conditional distribution of occupancy given (x,p), $f(d|p,x)$. For the purposes of a theoretical analysis of the identification of the model we will treat G and f as known conditional distributions. By Assumptions 0 to 3 above, we can write f and G as follows

$$\begin{aligned} f(d|p,x) &= \sum_A f(d|A,p,x)H(A|x) \\ G(p|x) &= \int I\{p(x,z) \leq p\}\psi(dz|x) \end{aligned} \tag{9}$$

where $I\{\cdot\}$ is the indicator function, $p(x,z)$ is the joint pricing strategy, i.e. the function hotels use to set their prices as a function of their information (x,z), $\psi(z|x)$ is the conditional distribution of z given x and $f(d|A,p,x)$ is the censored multinomial distribution given in equation (4). From our standpoint the pricing function $p(x,z)$ and the conditional distribution $\psi(z|x)$ are nuisance functions that are not of direct interest for estimation. Instead out interest is to identify and estimate the censored multinomial distribution $f(d|A,p,x)$ and the distribution of arrivals $H(A|x)$ which plays the role of a *mixing distribution* in our context. Still deeper, we are also interested in identifying the conditional choice probabilities $\pi_l(p,x)$ and potentially from these, the distribution of consumer types and the type-specific choice probabilities given in equation (3).

Note that prices may reflect the effect of unobserved characteristics of hotels ξ, and we will show

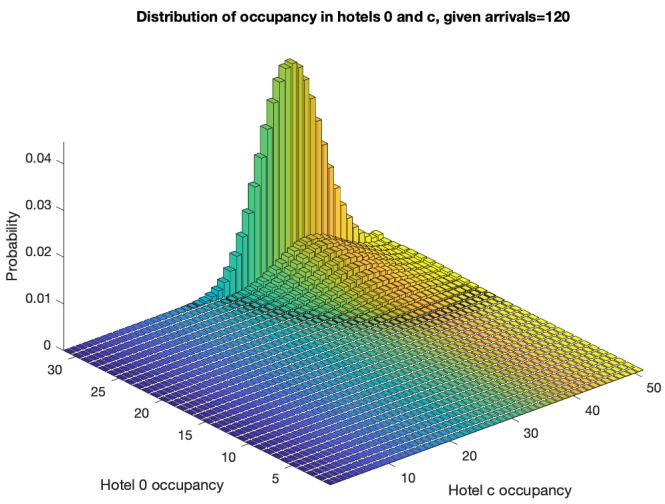


Figure 6: Occupancy distribution, $A = 120$

that prices can reflect endogeneity due to classical simultaneous equations bias. That is, variations in the number of arriving customers across different market days with different *ex ante* values of x that constitute observed demand shifters is enough in the face of the limited capacity of the hotels in this market to result a) a strong co-movement in prices among the various hotels in the market, and b) an upward sloping relationship between price and occupancy at individual hotels. Note, however, that since we can non-parametrically estimate $G(p|x)$ we do not have to take a stand on how individual hotels set their prices. We will now show that Assumptions 0-3 are sufficient to identify the effect of price on the demand for hotel rooms without requiring us to develop a detailed model of equilibrium in the hotel market, or even to assume anything about how individual hotels set their prices as a function of x and z, such as the assumption that individual hotels set their prices optimally as a best response to their beliefs of the prices set by their competitors.

Figure 6 illustrates the censored multinomial occupancy distribution $f(d|A,p,x)$ for two hotels ("hotel 0" and "hotel c") whose capacities are $C_0 = 30$ and $C_1 = 50$ respectively. We assume that the number of arrivals is $A = 120$ and the two hotels have accurate signals of the number of arrivals and set their prices accordingly. Since hotel 0 has a smaller capacity of $C_0 = 30$, it sets a price of $p_0 = 169$ and hotel c with the larger capacity of $C_1 = 50$ sets a price of $p_1 = 181$. We see that there is significant probability that hotel 0 sells out, while the chance that hotel c sells out is close to zero due to its higher capacity.

To illustrate how our framework is consistent with a fully rational, Bertrand-Nash equilibrium model of price setting, consider a duopoly market with two hotels, 0 and c. Suppose the unobserved private information that these firms use and that affects that their pricing decisions is partitioned as $z = (z_0, z_c)$

and the hotels have knowledge of the joint distribution of (x,z) and thus have well defined conditional probabitlity measures $\Psi_0(z_c|x,z_0)$ (i.e. hotel 0's belief about the distribution of private information of hotel c) and conversely $\Psi_c(z_0|x,z_c)$ is hotel c's belief about the private signal of hotel 0. A Bertrand-Nash equilibrium is a pair of functions $\{p_0(x,z_0),p_c(x,z_c)\}$ defined by

$$\begin{aligned} p_0(x,z_0) &= \underset{p_0}{argmax}\left[\int_{z_c}\sum_A\sum_{d_0}\sum_{d_c}\min(C_0,d_0)(p_0-c_0)f(d_0,d_c|A,p_0,p_c(x,z_c),x)H(A|x)\Psi_0(dz_c|x,z_0)\right] \\ p_c(x,z_c) &= \underset{p_c}{argmax}\left[\int_{z_0}\sum_A\sum_{d_0}\sum_{d_c}\min(C_c,d_c)(p_0-c_c)f(d_0,d_c|A,p_0(x,z_0),p_c,x)H(A|x)\Psi_c(dz_0|x,z_c)\right], \end{aligned} \tag{10}$$

where (C_0,C_c) are the capacities of the two hotels, (c_0,c_c) are the marginal costs of servicing rooms (which could be negative, if hotel guests consume other hotel services such as minibar, restuarant and other hotel amenities). The main point here is that identification of the "structural objects" $\{f,H\}$ enables us to calculate counterfactuals and test whether the actual conditional distribution of prices, $G(p|x)$ is consistent with the distribution induced by the assumption that firm behavior is given by Bertrand-Nash equilibrium in prices.

However the observed distribution of prices $G(p|x)$ may be consistent with any number of other theories of firm price-setting, including theories based on bounded rationality and that incorporate the effects of "pricing mistakes" that are the source of the random shocks z affecting firm prices. Given assumptions 0-3, any one of these theories can generate data that reveals the endogeneity we illustrated for hotel 0 in figure 4 of section 2, i.e. where the best fitting regression line results in an upward sloping demand curve for each hotel as a function of its own price. We will now show that by controlling for the demand shift variable x, Assumptions 0-3 can enable us to identify a downward sloping expected demand curve, where the expected demand (more precisely, expected occupancy) for hotel 0 is given by

$$E\{d_0|p_0,p_c,x\} = \sum_A\sum_{d_0}\sum_{d_c} min(C_0,d_0)f(d_0,d_c|A,p_0,p_c,x)H(A|x). \tag{11}$$

Lemma 1 *Assume Assumptions 0-3 hold. If the choice probability for hotel 0, $\pi_0(p_0,p_c,x)$ is downward sloping in p_0 and upward sloping in p_c, then the expected demand function $E\{d_0|p_0,p_c,x\}$ is also downward sloping in p_0 and upward sloping in p_c.*

Proof Accounting for the outside good, for each A we can express the expectation in equation (11) as the expectation with respect to the marginal binomial distribution for d_0 with parameters $(A,\pi_0(p_0,p_c,x))$, since given the number of arrivals A the marginal distribution of the trinomial distribution of outcomes for hotel 0 occupancy, hotel c occupancy and the outside good, respectively, $(d_0,d_c,A-d_0-d_c)$, is a

binomial distribution. It is well known that a binomial distribution satisfies the property of first order stochastic dominance with respect to the probability of occurrence, i.e. a binomial random variable d_0 with parameters (A,π) stochastically dominates a binomial random variable d_0' with parameters (A,π') if and only if $\pi \geq \pi'$. This in turn implies that the expectation of any monontone increasing function $h(d_0)$ is a monotone function of π_0, and since $h(d_0) = \min(C_0, d_0)$ is monotone increasing, we conclude that if π_0 is monotone decreasing in p_0 and increasing in p_c then $E\{d_0|A,p_0,p_c,x\}$ is also monotone decreasing in p_0 and monotone increasing in p_c. Since $E\{d_0|p_0,p_c,x\}$ is a mixture of monotone functions, it is also montone decreasing in p_0 and increasing in p_c. □

The Lemma shows that if we were to estimate the demand model by nonlinear regression, then by controlling for the demand shifter x we obtain a regression function that has the right slopes with respect to p_0 and p_c under Assumptions 0-3, provided the individual choice probabilities also slope in the right way as a function of (p_0,p_c). The latter will be true for a wide range of choice models such as where $\pi(p,x)$ are mixtures of logits, probits, etc. Further, the conditional exogeneity assumption 2 implies that once we control for x, the prices (p_0,p_c) are econometrically exogenous and thus, we can recover the expected demand for hotel 0 by non-parametric regression using only occupancy data for hotel 0, the joint prices for all hotels (and the outside good) p, and the observed demand shifter x.

3.1 Non-parametric identification

We now consider the problem of identifcation, to understand what objects can be identified under Assumptions 0-3. We are particularly interested in whether it is possible to identify consumer preferences and arrivals from potentially endogenously generated makret price and occupancy data.

Definition 1 *The structure of the hotel pricing problem consists of the objects*

$$\Gamma = \{\{g(\tau|x)|\tau \in \{1,\ldots,\mathcal{T}\}\},\{P_\tau(l|p,x)|\tau \in \{1,\ldots,\mathcal{T}\},\, l \in \{0,1,\ldots,L\}\},H(A|x)\},\, x \in X. \qquad (12)$$

We exclude the conditional distribution $G(p|x)$ from the elements of the structure from the problem since we wish to relax the assumptions of 1) equilibrium (the firms' set prices in accordance with a Bertrand-Nash equilibrium), and 2) optimality (firms set optimal prices, given possibly non-equilibrium beliefs about the prices charged by their competitors). However we consider the choice probabilities $P_\tau(l|p,x)$, the distribution of consumer types $g(x|\tau)$ and the arrival probability $H(A|x)$ as structural objects that would not change under different assumptions about how the hotels set their prices, such as if they set prices optimally and in accordance with a Bertrand Nash equilibrium. Thus, if we can identify the structural objects, we can in principle solve for a Bertrand-Nash equilibrium and compare the distribution of prices $G^*(p|x)$ arising under a Bertrand-Nash equilibrum to the distribution $G(p|x)$ that could potentially

be identified under the *status quo.* Thus, G is not invariant, and depends on what assumptions we make about how the hotels set their prices.

Definition 2 *The identified objects for the hotel pricing problem are given by*

$$\Lambda = \{f(d|x,p), G(p|x)\}, \quad x \in X. \tag{13}$$

We assume that we can observe the hotels over a sufficiently long period of time where the stationarity assumption holds to consistently estimate the conditional distribution of prices given x, $G(p|x)$ so we can take this conditional distribution as "known" for purposes of the analysis of identification. Since we can consistently estimate $G(p|x)$ without imposing the assumption of equilibrium or optimality, we can be agnostic about firm behavior. Similarly, given sufficient data, we assume we can estimate the conditional distribution of occupancy given (x,p) via non-parametric methods, $f(d|x,p)$.

We start by making some easy observations about the model: Lemma 2 below shows that the expected demand curves are also identified, and the identification of these objects are already sufficient to enable to do to interesting counterfactual calculations and hypothesis tests. For example, we can use the demand functions to test whether the hotels are profit-maximizing, or whether the prices in the market are consistent with the Betrand-Nash equilibrium outcomes

Lemma 2 *Under Assumptions 0-3, the expected occupancy function, $E\{d|p,x\}$ is non-parametrically identified.*

Proof This follows immediately from the non-parametric identification of the joint occupancy density, $f(d|p,x)$, since $E\{d|p,x\}$ is simply the expectation with respect to this joint density. □

Although elementary, there are a number of immediate implications of Lemma 2. First, it implies that we can make interesting counterfactual calculations under much weaker assumptions than are usually imposed in the structural estimation literature. For example, we can test the hypothesis that hotel 0 is using an optimal pricing strategy given the strategy of its competitors. To do this, we use the expected demand function to calculate the optimal pricing strategy $p_0^*(x)$ for hotel 0 as follows

$$p_0^*(x) = \underset{p_0}{argmax} \int_{p_c} (p_0 - c_0) E\{d_0|p_0,p_c,x\} G_c(dp_c|x). \tag{14}$$

where $G_c(p_c|x)$ is the marginal distribution of p_c calculated from the joint conditional distribution $G(p|x)$. The pricing strategy in equation (14) does not depend on any private information shock z_0 that only hotel 0 can observe. This strategy will be optimal for hotel 0 provided that any private information it observes is independently distributed from any private information that hotel c observes, z_c, but it will be suboptimal if hotel 0 does have private information z_0 (not observed by the econometrician) that is correlated with z_c and helps it to predict the prices charged by hotel c. If z_0 and z_c are independently distributed and

Assumptions 0-3 hold, hotel 0's optimal pricing strategy should only be a function of the demand shifter variable x and not on any other extraneous information z_0. This implies that $G_0(p_0|x)$ is a degenerate distribution that puts probability 1 on the price $p_0^*(x)$ which is a testable restriction on the identified joint conditional distribution $G(p|x)$.

Similarly, we can also compute a full-information Bertrand-Nash equilibrium for hotels 0 and c under the assumption that any private information they observe, z_0 and z_c respectively, are independently distributed. Then hotel c will also have a optimal pricing strategy $p_c^*(x)$ that does not depend on its private information shock z_c, and in a Bertrand-Nash equilibrium, the pricing strategies $(p_0^*(x), p_c^*(x))$ will satisfy the fixed-point condition

$$\begin{aligned} p_0^*(x) &= \underset{p_0}{argmax}(p_0 - c_0)E\{d_0|p_0, p_c^*(x), x\} \\ p_c^*(x) &= \underset{p_c}{argmax}(p_c - c_c)E\{d_c|p_0^*(x), p_c, x\} \end{aligned} \tag{15}$$

and this hypothesis has the testable implication that $G(p|x)$ is degenerate distribution that puts probability 1 on the point $(p_0^*(x), p_c^*(x))$.

However there are other counterfactual calculations and hypotheses that require knowledge of the deeper structure of the problem, particularly to separately identify consumer preferences (both their preferences for different hotels and price sensitivity as well as the distribution of different types of consumers in the population), and the arrival probability distribution $H(A|x)$. For example, we might be interested in predicting how the market would be affected if there was a shift in the probability distribution of arrivals, or what would happen to prices in the market if one of the hotels expanded its capacity, or one of the hotels exited the market, or there was a hotel merger. These counterfactuals cannot be calcuated given knowledge of the distribution of occupancy $f(d|p,x)$ under the *status quo* which posits a fixed number L of competing hotels, with fixed capacities C and a fixed arrival distribution $H(A|x)$. As we noted in the introduction, a counterfactual of particular interest is to calculate what hotel prices would be in this market if the hotels were to collude, possibly in response to recommended prices if they were to all use a common RMS that was capable of calculating and recommending a collusive price vector. We would like to calculate a collusive price strategy and compare the implied distribution of prices to the distribution $G(p|x)$ that holds under the *status quo* $G(p|x)$ to provide a test for "algorithmic collusion." However in order to do this, we need to know more about consumer preferences and particularly the rate of substitution to the outside good in the event collusive pricing would significantly raise prices in this hotel market relative to nearby markets. Thus, in the remainder of this section we provide some results on the non-parametric identification of the structure given in Definition 1.

The identification problem concerns the question as to whether there is an invertible mapping from the

identified objects Λ to the structure Γ. Define a mapping $\Psi(\Gamma)$ from the structure to the first component of the identified objects Λ by

$$f(d|x,p) = \sum_A f(d|A,x,p)H(A|x) \equiv \Psi(\Gamma), \tag{16}$$

where $f(d|A,x,p)$ is the censored multinomial distribution of hotel occupancy given the number of arrivals A that we introduced in equation (4). Note that equations (4) and (6) imply that $f(d|A,x,p)$ is itself a function of the other components for the structure Γ, i.e. the distributions over consumer types $g(\tau|x)$ and the consumer choice probabilities $P_\tau(l|p,x)$, for $\tau \in \{1,\ldots,\mathcal{T}\}$ and $x \in X$.

Definition 3 *Two structures $\Gamma \neq \Gamma'$ are observationally equivalent if $\Psi(\Gamma) = \Psi(\Gamma')$.*

Definition 4 *The hotel model with identified objects Λ is identified if there is a structure Γ satisfying $\Psi(\Gamma) = \Lambda$ and there is no other structure $\Gamma' \neq \Gamma$ that is observationally equivalent to Γ.*

The identification problem reduces to a question on the identification of a mixture models, as can be seen in equation (16), where we are interested in "inverting" the distribution of occupancy given (x,p) given by $f(d|x,p)$ to uniquely determine the "component distributions" $f(d|A,x,p)$ and the conditional distribution of arrivals $H(A|x)$. Actually we have a problem of identification of a *nested mixture model*, since in addition to identifying the component distributions $\{f(d|A,x,p\}$ and the mixing distribution $H(A|x)$, we are also interested in identifying the choice probabilities $\pi_l(p,x)$ and from them, the mixing distribution representation of unobserved heterogeneity in consumers given in equation (3). This is also clearly a problem of identification of mixtures, but in this case we presume knowledge of the type-unconditional choice probabilities $\pi_l(p,x)$, $l \in \{\emptyset, 0, 1, \ldots, L-1\}$ and from these identify the component probabilities $\{P_\tau(l|p,x)\}$ and the mixing distribution $g(\tau|x)$ for all possible values of (p,x).

We will assume that maximum number of consumers arriving to the market is uniformly bounded with a known upper bound (even though the actual support of $H(A|x)$ may be unknown):

Assumption 4 (Finite support) *Let $|A|(x)$ be the size of the support of the integer valued distribution $H(A|x)$. We have $|A| \equiv \max_{x\in X} |A|(x) < \infty$ where the upper bound $|A|$ is known a priori.*

Assume for the moment that we can identify the choice probabilities $\pi_l(p,x)$ of the censored multinomial representation of $f(d|A,p,x)$ given in equation (4) in the "upper level" mixture identification problem. To identify the distribution of types $g(\tau|x)$ and type-specific choice probabilities $\{P_\tau(l|p,x)\}$ in the "lower level" mixture identification problem in (3) we need to impose some additional structure and assumptions.

Assumption 5 (Mixed logit) *The utility functions for consumers given in equation (2) are linear in parameters*

$$u_\tau(l,p_l,x) = \alpha(l,x) - \beta_\tau(x)p_l, \tag{17}$$

where for the outside good, $l = 0$ we impose the normalization $u_\tau(0, p_0, x) = 0$, and we assume the error terms $\varepsilon(l)$ are standardized Type 1 extreme value (i.e. have mean zero and scale parameter 1). This implies that the type-specific choice probabilities $P_\tau(l|p,x)$ take the standard multinomial logit form

$$P_\tau(l|p,x) = \frac{\exp\{\alpha(l,x) - \beta_\tau(x) p_l\}}{1 + \sum_{l'=1}^{L} \exp\{\alpha(l',x) - \beta_\tau p_{l'}\}}. \tag{18}$$

Given this we can apply the non-parametric identification result of Fox, Kim, Ryan, and Bajari (2012) to establish that both the number of unobserved types $\mathcal{T}$ and the conditional distribution of types $g(\tau|x)$ as well as the "random coefficients" $\{\beta_\tau(x)\}$ for τ in the support of the discrete distribution $g(\tau|x)$, as well as the (non-type-specific) intercept terms $\{\alpha(l,x)\}$, $l \in \{0, 1, \ldots, L-1\}$ for each $x \in X$.

Thus, the non-parametric identification of the model hinges on whether it is possible to separately identify the component distributions $f(d|A,x,p)$ and the conditional distribution of arrivals (mixing distribution) $H(A|x)$ in equation (16). Promising recent progress on the identification of mixture models by Kitamura and Laage (2018) suggests that this may be possible. However we cannot directly apply their key result, Proposition 6.1, since the structure of our problem is not nested within the class of mixture models that they consider. Specifically, they consider the identification of mixture models that can be written as a regression equation for an observed dependent variable y given covariates x

$$y = f(x) + \varepsilon \tag{19}$$

where the actual observations are drawn from a mixture of regression models

$$y_j = f_j(x) + \varepsilon_j, \quad j \in \{1, \ldots, J\} \tag{20}$$

with probability $\lambda_j \geq 0$ with $\sum_{j=1}^{J} \lambda_j = 1$. In this case the $\{\lambda_j\}$ are the mixing distributions and the $\{f_j(x)\}$ are the component distributions. Proposition 6.1 of Kitamura and Laage (2018) establishes the non-parametric identification of this mixture model, i.e, given knowledge of the regression $f(x)$ and the distribution of ε, they establish that the number of mixture components J and the mixing probabilities $\{\lambda_j\}$ and the component regression functions $\{f_j(x)\}$ are identified. However their result requires the error terms $\{\varepsilon_j\}$ are univariate random variables that are assumed to be continuously distributed and independent of the regressor x, and *IID* across the different types j. In addition their result relies on additional assumptions that guarantee that the regression functions $f_j(x)$ are "non-parallel" as well as other technical assumptions about the moment generating functions of the $\{\varepsilon_j\}$.

In our case we can write occupancies as a multivariate system of regressions

$$d = E\{d|p,x\} + \varepsilon \tag{21}$$

where

$$E\{d|p,x\} = \sum_A E\{d|A,p,x\}H(A|x) \tag{22}$$

so it is tempting to try to apply Proposition 6.1 to our setting. However the component regressions in our case are

$$d_A = E\{d|A,p,x\} + \varepsilon_A \tag{23}$$

(i.e. where the number of arrivals A index the mixture components) but the error terms in our case, $\varepsilon_A = d_A - E\{d|A,p,x\}$ are multivariate random variables that are not continuously distributed or *IID* when considered as indexed over different values of arrivals A.[6] Thus, we cannot directly apply Proposition 6.1 of Kitamura and Laage (2018) to establish the non-parametric identification of our hotel model. Further, they provide counter-examples showing the mixture model (19) is non-identified when the error terms $\{\varepsilon_A\}$ are heteroscedastic, as they are in our case.

Instead we establish the identification of mixtures of censored multinomials via a direct argument. First, observe that if we fix the continuous price regressor p and the demand shifter x, we can consider the key equation (16) as nonlinear system of equations. It is actually a polynomial system of equations in $(\pi_\emptyset, \pi_0, \ldots, \pi_{L-1})$ and a linear system in $H(A|x)$ given π and (p,x). Let $|D| = \prod_{l=1}^{L}(C_i+1)$ be the size of the support of $f(d|x,p)$. Then $|D|$ indexes the number of left hand side "known values" in equation (16), whereas $\{f(d|A,x,p), H(A|x)\}$ on the right hand side are the "unknowns." Let $|A|(x)$ be an *a priori* known upper bound on the support of the number of arrivals. Without imposing any further special structure on the system of equations (16) identification would seem to be hopeless since it constitutes a system of $|D|$ equations in at most $|A|(1+|D|)$ unknowns, and thus in principle there could be far more unknowns than equations. However there is substantial *special structure* to the hotel problem in view of the fact that $f(d|A,x,p)$ takes the form of a censored multinomial distribution in (4). For fixed (p,x), this special structure reduces the problem to a nonlinear system of equations with $|D|$ equations in $L+|A|(x)-1$ unknowns. If $|D| > L+|A|(x)-1$, then equation (16) will be a an over-determined system, i.e. it will have more equations than unknowns. We can consider this to be a basic "rank condition" for identification.

Note that for fixed (p,x), if we treat the component distributions $f(d|A,p,x)$ as known, equation (16) can be viewed as a system of linear equations $f = f_A \times H$ where f_A is a matrix of dimension $|D| \times |A|$ formed with the densities $f(d|A,p,x)$ arrays as its columns. If the matrix f_A has full rank, then there is a unique mixing distribution $H(A|x)$ that solves (16) when the component distributions $f(d|A,p,x)$ are fixed at their true values. However we note that (16) is a *nonlinear* system of equations when we consider $\{H(A|x), \{\pi_l(p,x)\}, l \in \{\emptyset, 0, 1, \ldots, L-1\}\}$ as the full set of unknowns. Therefore a different argument

[6] If the capacities of the hotels are not symmetric, e.g. if $C_l \neq C_{l'}$ for $l \neq l'$, then the different components of ε_A will have different distributions, and the overall vector ε_A will have different distributions for different values of the arrivals A.

is required to establish identification of these functions. We analyze identification under two possible information structures:

- **Full information** The econometrician can observe the number of customers choosing the outside good, or the number of arrivals, or both.
- **Limited information** The econometrician cannot observe the number of customers choosing the outside good, or the number of arrivals.

Our first result is a lemma that establishes that the model is partially identified under either information structure.

Lemma *Under Assumptions* $0,\ldots,4$ *if* $C_l \geq 1$, $l \in \{0,\ldots,L\}$ *then the ratios of the choice probabilities,* $r_l(p,x) = \pi_l(p,x)/\pi_0(p,x)$ *are identified for* $l = 1,\ldots,L-1$ *for any* (p,x) *such that* $\pi_0(p,x) > 0$.

Proof Note first that if $C_l \geq 1$ for $l = 0,\ldots,L-1$, then $|D| \geq 2^L > L$ so the rank condition for identification is satisfied. Let e_l be an $L \times 1$ vector whose elements equal 0 except for element l which equals 1. Then $f(e_l|p,x)$ is the probability that hotel l has exactly 1 customer occupying one of its rooms. By assumption this probability is known and positive for each l. We have

$$f(e_l|p,x) = \pi_l(p,x)\sum_A A\pi_\emptyset^{A-1}H(A|x), \quad l \in \{0,1,\ldots,L-1\} \tag{24}$$

From equation (24) it immediately follows that the ratios $r_l(p,x)$ given by

$$r_l(p,x) = \frac{f(e_l|p,x)}{f(e_0|p,x)} = \frac{\pi_l(p,x)}{\pi_0(p,x)}, \quad l \in \{1,\ldots,L-1\} \tag{25}$$

are identified. □

We can write the choice probabilities $\pi_l(p,x)$ in terms of the identified ratios of choice probabilities, $r_l(p,x)$ as $\pi_l(p,x) = \pi_0(p,x)r_l(p,x)$ and use the fact that the choice probabilities sum to 1 to write the probability of the outside good, $\pi_\emptyset(p,x)$ in terms of $(\pi_0(p,x),\ldots,\pi_{L-1}(p,x))$, to reduce the identification problem to the solution of a system of $|D|-L$ equations in $|A|$ unknowns, $(\pi_0(p,x),\{H(A|x)\})$.

Theorem 0 *Suppose we have full information on arrivals. Then if Assumptions 0-5 hold and* $C_l \geq 1$, $l = 0,\ldots,L-1$ *the hotel model is non-parametrically identified.*

Proof Under full information, the hotels (and the econometrician) can observe all arrivals and all consumers who choose the outside good (though observing one enables us to deduce the other via the identity

$$A = d_\emptyset + \sum_{l=0}^{L-1} d_l. \tag{26}$$

Since arrivals are observed, it follows that $H(A|x)$ is identified for each x (since we presume for the purposes of the analysis of identification we have infinitely many observations and thus can consistently

estimate the discrete distribution $H(A|x)$ from the empirical distribution). So the question of identification reduces to the identification of the probability $\pi_0(p,x)$. Define $r_0(p,x) = 1$. Since the choice probabilities sum to 1 for all (p,x), we have

$$\pi_\emptyset(p,x) = 1 - \sum_{l=0}^{L-1} \pi_l(p,x) = 1 - \pi_0(p,x) \sum_{l=0}^{L-1} r_l(p,x) \tag{27}$$

where the $r_l(p,x)$ are known functions of (p,x) by the partial identification lemma above. Let 0 be an $L \times 1$ vector of 0's, so $f(0|p,x)$ is the probability of zero occupancy in all L hotels given (p,x), which is also a known function given our assumption that $f(d|p,x)$ is identified. We have

$$f(0|p,x) = \sum_A \pi_\emptyset(p,x)^A H(A|x) = \sum_A \left[1 - \pi_0(p,x) \sum_{l=0}^{L-1} r_l(p,x)\right]^A H(A|x) \tag{28}$$

by equation (27). Note that equation (28) is a polynomial equation in $\pi_0(p,x)$ and we know it has at least 1 solution in the unit interval, where at least one root is the true value $\pi_0(p,x)$ that customers choose hotel 0. Define the polynomial $P(y) : R \to R$ by

$$P(y) = \sum_A \left[1 - y \sum_{l=0}^{L-1} r_l(p,x)\right]^A H(A|x). \tag{29}$$

Notice that $P(0) = 1$ and furthermore, we have

$$P'(y) = -\left(\sum_{l=0}^{L-1} r_l(p,x)\right) \left[\sum_A A \left(1 - y \sum_{l=0}^{L-1} r_l(p,x)\right)^{A-1} H(A|x)\right] < 0 \quad y \in [0,1]. \tag{30}$$

Since $f(0|p,x) \in (0,1)$ and we know there is one solution of the equation $P(y) = f(0|p,x)$ in the unit interval (i.e. the true probability $\pi_0(p,x)$), equations (29) and (30) imply that there is only one solution in the unit interval, i.e. $\pi_0(p,x)$ is identified, and thus the entire model $\{(\pi_\emptyset(p,x), \ldots, p_{L-1}(p,x)), H(A|x)\}$ is identified. □

In the limited information case, we do not observe the number of consumers who arrive in the hotel market, nor the consumers who choose the outside good. We can only observe the occupancy in each of the hotels, and with sufficient data, this enables us to consistently estimate $f(d|p,x)$, the joint distribution of occupancy given (p,x). Identification is more difficult in this case since we cannot directly recover the distribution of arrivals, $H(A|x)$, which was the first key step to the proof of Theorem 0 for the case where we have full information (e.g. we observe arrivals). However the intuition that when $|D| > L + |A|(x) - 1$ we have more equations than unknowns and so the rank order for identification is satisfied does not automatically lead to a proof of identification. Though we conjecture that the model is identified when this rank condition holds, at this point we require additional conditions to prove identification, given in Theorem 1 below.

Theorem 1 *Suppose Assumptions* $0,\ldots,5$ *hold, and* $C_l \geq 1$, $l = 0,1,\ldots,L-1$. *Further, suppose that for each* $x \in X$ *that* $|C| > |A|(x)$ *where where* $|C| = \sum_{l=1}^{L} C_l$ *is the total room capacity in the market. Also suppose for each* x *there exists a* p *that satisfies* $\pi_{\emptyset}(p,x) = \pi_0(p,x)$. *Then the hotel demand model is identified.*

Proof: The largest number of arrivals when the demand shifter is x is identified from $f(d|p,x)$ as the largest occupancy vector in the support of $f(d|p,x)$:

$$|A|(x) = \sup_{d}\{|d||f(d|p,x) > 0\} \tag{31}$$

where $|d| = \sum_{l=0}^{L-1} d_l$. Next, by the assumption that for each x there exists a p satisfying $\pi_{\emptyset}(p,x) = \pi_0(p,x)$, we can solve for $\pi_0(p,x)$ via the equation

$$1 - \pi_{\emptyset}(p,x) = 1 - \pi_0(p,x) = \pi_0(p,x)\left[\sum_{l=0}^{L-1} r_l(p,x)\right] \tag{32}$$

or

$$\pi_0(p,x) = \pi_{\emptyset}(p,x) = \frac{1}{1+\sum_{l=0}^{L-1} r_l(p,x)}, \tag{33}$$

and hence the choice probabilities $(\pi_{\emptyset}(p,x),\ldots,\pi_L(p,x))$ are identified, *for this particular* p. Now we show how to identify $H(|A|(x)|x)$, i.e. the probability of the maximum number of arrivals $|A|(x)$ when the the demand shifter is x. First, the identification of the choice probabilities $(\pi_{\emptyset}(p,x),\ldots,\pi_L(p,x))$ implies that for any $A \geq 0$, the censored multinomial distribution $f(d|A,p,x)$ given in equation (4) is identified. Since $|A|(x)$ is the maximal number of arrivals in state x, then for any d satisfying $f(d|p,x) > 0$ and $|d| = |A|(x)$ we have

$$f(d|p,x) = H(|A|(x)|x)f(d|A,p,x) \tag{34}$$

so $H(|A|(x)|x)$, the probability of $|A|(x)$ arrivals in state x, is identified.

Next we show by induction that $H(A|x)$ is identified for all $A < |A|(x)$. Suppose the arrival probabilities $\{H(|A|(x)|x), H(|A|(x)-1|x),\ldots,H(A|x)\}$ are identified. We show that $H(A-1|x)$ is identified as follows. Let d_A be any joint occupancy in the support of $f(d|p,x)$ satisfying: a) $|d_A| = A$ and b) $f(d_A|p,x) > 0$. Let d_{A-1} be an occupancy vector satisfying $d_{A-1,l} = d_{A,l}$ for $l = 1,\ldots,L-1$ and $d_{A-1,0} = d_{A,0} - 1$. Then we have $|d_{A-1}| = A-1$ and we have

$$f(d_{A-1}|p,x) = f(d_{A-1}|A-1,p,x)H(A-1|x) + \sum_{A'=A}^{|A|(x)} f(d_{A-1}|A',p,x)H(A'|x). \tag{35}$$

By our inductive hypothesis, the sum on the right hand side of equation (35) is identified. Since $f(d_{A-1}|A-1,p,x) > 0$, it follows that we can solve this equation for $H(A-1|x)$ and so it is identified as well. Thus we conclude for each $x \in X$ that $H(A|x)$ is identified.

To complete the proof, we need to show that the choice probabilities $(\pi_\emptyset(p',x),\pi_0(p',x),\ldots,\pi_{L-1}(p'x))$ are identified not only for the particular p for which $\pi_\emptyset(p,x)=\pi_0(p,x)$, but also for any p' in the support of $G(p|x)$ that may not satisfy the restriction that $\pi_\emptyset(p',x)=\pi_0(p',x)$. However by repeating the proof of Theorem 0, we see once $H(A|x)$ is identified, it follows that the choice probabilities $(\pi_\emptyset(p',x),\pi_0(p',x),\ldots,\pi_{L-1}(p'x))$ are identified for all p' in the support of $G(p|x)$. □

We believe the hotel model is identified under weaker assumptions than those assumed in Theorem 1, but we have not yet succeeded in providing a proof of this. Theorem 2 shows that once we are able to identify the choice probabilities, $\pi(p,x)$, we can also identify the distribution of unobserved heterogeneity $g(\tau)$ and the random coefficients $\{\beta_\tau(x)\}$ in the multinomial logit type-specific choice probabilities in equation (18).

Theorem 2 *Under the assumptions of Theorem 0 with full information, or Theorem 1 with limited information, the distribution of types $g(\tau)$ and associated random coefficients $\{\beta_\tau(x)\}$ are identified.*

Proof This result follows from the identification result of Fox et al. (2012).

We conclude this section with a negative result that is quite intuitive: if we only observe occupancy for a single hotel in the market, say hotel 0, then we can only identify the expected demand for hotel 0, but we cannot separately identify the choice probabilites for the other hotels, $\pi_l(p,x)$, $l=1,\ldots,L$ and the probability of choosing the outside good, $\pi_\emptyset(p,x)$, nor can we generally fully identify the distribution of arrivals, $H(A|x)$. To see why, note that when we only observe occupancy at hotel 0, we can only identify the marginal distribution of occupancy at hotel 0, $f(d_0|p,x)$, which is a mixture of binomials

$$f(d_0|p,x)=\sum_A \binom{A}{d}\pi_0(p,x)^d[1-\pi_0(p,x)]^{A-d}H(A|x). \tag{36}$$

Note that since the right hand side only depends on $H(A|x)$ and $\pi_0(p,x)$, it will not be possible to identify the probabilities $\{\pi_l(p,x)\}$, $l=1,\ldots,L$ and $\pi_\emptyset(p,x)$. Secondly, since the capacity of hotel 0, C_0, will generally be far smaller than the total capacity of the market as a whole, it is not plausible to assume that the maximal number of arrivals, $|A|(x)<C_0$, so in general we will not be able to infer $|A|(x)$ from knowledge of $f(d_0|p,x)$. In general, the upper tail of the arrival distribution will only be partially identified. Note if we knew the entire distribution $H(A|x)$ we could adapt the proof of Theorem 0 to establish identification of $\pi_0(p,x)$ from knowledge of $f(d_0|p,x)$ and $H(A|x)$. However if H is only partially identified, it is no longer even clear that it is possible to identify $\pi_0(p,x)$. We record this as

Theorem 3 *If we only observe occupancy at a single hotel, the choice probabilities and arrival distribution are only partially identified. In general the only fully identified objects in this case are the conditional distribution $f(d_0|p,x)$ and its expectation $E\{d_0|p,x\}$.*

4 Empirical Application to the Hotel Market

The identification analysis in the previous section shows that in principle the endogeneity problem can be solved, enabling us to identify the deeper underlying structure of demand — consumer preferences and the distribution of arrivals — even in situations where there are no instrumental variables and there is censoring and truncation. The latter problems arise due to our inability to observe arrivals, the choice of the outside good, and the number of consumers who are rationed when the hotels rearch their capacity constraints. Our analysis shows that demand for hotels can be consistently estimated under weak assumptions that relax the traditional optimality and equilibrium assumptions imposed in empirical work. However we emphasize the words *in principle* since the theoretical analysis of identification presumes we have access to an infinite amount of data and thus can non-parametrically estimate the conditional distribution of joint occupancy $f(d|p,x)$ and the conditional distribution of prices, $G(p|x)$.

However even though we have a unique set of data on both prices and occupancies of the hotels in this market, our data has limitations. As we noted in section 2, we only have 1737 daily observations on occupancies and ADRs for the hotels in our market. Even conditional on a given value of the covariates, (p,x), the total number of elements in the support of $f(d|p,x)$ (i.e. the total number of possible elements of the joint distribution of occupancy) is 2×10^{17}, so it is clearly hopeless to estimate the entire conditional distribution $f(d|p,x)$ non-parametrically from the limited data we have at hand. Further, as we noted in section 2, we only have the sum of the occupancies of hotel 0's competitors, not the individual occupancies of the competing hotels on a daily basis. Therefore we have chosen to treat hotel 0's competitors as a single "aggregate competitor" and model the market as if it were a duopoly with only two competing hotels: hotel 0 and hotel c (where the latter is the aggregate of the 6 competitors to hotel 0). However even when we do this, so $f(d|p,x)$ is reduced to a two-dimensional distribution over the joint occupancy of hotel 0 and hotel c, there are still nearly 600,000 possible (d_0, d_c) values in the support of this conditional distribution. Thus, in order to proceed with an empirical analysis, additional parametric functional form assumptions are necessary.

We employed two different estimation strategies:

1. **Regression,** to uncover the demand curve only, $E\{d|p,x\}$

2. **Maximum likelihood,** to estimate the full structure of the model (i.e. preferences and the distribution of arrivals) using the likelihood function

$$f(d|p,x,\theta,\gamma) = \sum_A f(d|A,p,\theta) H_N(A|x,\gamma) \tag{37}$$

where $f(d|A,p,\theta)$ is a censored trinomial distribution with parameters (A,π), $\pi(p)$ is a 3×1 vector of trinomial logit probabilities specified below and $H(A|x,\gamma)$ is a multinomial logit probability disribution over a fixed set $\{A_1,\ldots,A_N\}$ of N arrival support points discussed below.

We used parametric and semi-parametric approaches for both estimation strategies. For the regression analysis, we estimated standard linear regression models for demand as well as non-parametric regressions (local linear regressions). Of course semi-parametric regression approaches can be employed as well such as a partially linear specification

$$d_0 = g(x) + \sum_{k=1}^{K_0} \theta_k^0 p_0^k + \sum_{k=1}^{K_c} \theta_k^c p_c^k + \varepsilon_0 \tag{38}$$

where the dependence on x is captured by the non-parametric component $g(x)$ but the dependence of demand on prices is captured by a flexible polynomial specification. Semi-parametric specifications can also allow for interaction effects between x and prices p, since as we show below, occupancy is necessarily a more inelastic function of prices on days where the hotel is close to capacity compared to days where there is substantial excess capacity.

For maximum likelihood estimation, we adopt a semi-parametric estimation strategy where the parametric part of the model consists of four preference parameters $\theta = (\alpha_0, \alpha_c, \beta, \beta_\emptyset)$ for a trinomial logit model of hotel choice, where the three choices are 1) hotel 0, 2) hotel c, and 3) the outside good. The probability of choosing hotel 0 is given by $\pi(p) = (\pi_0(p), \pi_c(p), \pi_\emptyset(p))$ where

$$\pi_0(p) = \frac{\exp\{\alpha_0 - \beta p_0\}}{\exp\{-\beta_\emptyset p_c\} + \exp\{\alpha_0 - \beta p_0\} + \exp\{\alpha_c - \beta p_c\}} \tag{39}$$

and the probabilities $\pi_c(p)$ and $\pi_\emptyset(p)$ are defined similarly. In equation (39) we have made a standard identification normalization, fixing the intercept term for the outside good to zero, $\alpha_\emptyset = 0$. We do not have data on the price of the outside good (e.g. prices on hotels outside this market) but we assume that $p_\emptyset$ is a fixed multiple of p_c, and we embed this into the price coefficient $\beta_\emptyset$. Thus the key parameters determining the slope of the implied demand curve are $(\beta, \beta_\emptyset)$ and if $p_\emptyset = p_c$, then we expect that $\beta = \beta_\emptyset$, and that $\beta_\emptyset$ will be higher or lower than β to the extent that the unobserved price of the outside good is higher or lower than p_c. Notice under this specification, consumer preferences are assumed to be independent of x, so the only way that x affects demand is by shifting the distribution of arrivals, $H(A|x)$. We can allow the distribution of consumers who arrive to this market to depend on x. In this case the mixing distribution will depend on price, resulting in mixed choice probabilities $\pi(p,x)$ that depend on x. In our analysis below we consider specifications where x can affect preference parameters. This captures selection effects such as the possibility that consumers who arrive on the busiest days have stronger preference for the hotels,

or are less price sensitive compared to consumers who book rooms on less busy days (lower x). Both of these possibilities are allowed via our general specification of choice probabilities $\pi(p,x)$ in equation (3). However due to the limited number of observations, our empirical analysis starts by imposing the exclusion restriction that x does not enter consumer choice probabilities and thus only enters the model as a demand shifter affecting the the arrival distribution $H(A|x)$.

We avoid making strong assumptions on the conditional distribution of arrivals, $H(A|x)$, and instead treat it as an unknown infinite dimensional parameter that we estimate via semi-parametric maximum likelihood. Since we cannot directly estimate $H(A|x)$ as an infinite dimensional parameter, we approximate it via the method of sieves and rely on the consistency and asymptotic normality results of Wong and Severini (1991) to establish the asymptotic distribution of the parameters of interest, θ. We consider a sieve consisting of a sequence of families of conditional distributions $H_N(A|x)$ with a finite support over N integers $\{A_1,\ldots,A_N\}$ and we index N to the sample size T which we denote by $N(T)$. We allow $N(T)\to\infty$ at the right rate as a function of the sample size T to ensure that $H_N(A|x)$ can consistently estimate any conditional density $H(A|x)$. In our empirical work we use a flexible family of multinomial logit models that depend on a vector of parameters γ of dimension $2N-1$ that provides a fully flexible distribution over an increasingly fine grid of N support points $\{A_1,\ldots,A_N\}$ given by

$$H_N(A_i|x,\gamma)=\frac{\exp\{\gamma_{1,i}+\gamma_{2,i}x\}}{\sum_{j=1}^{N}\exp\{\gamma_{1,j}+\gamma_{2,j}x\}} \tag{40}$$

where we impose the identifying normalizations that $\gamma_{1,1}=\gamma_{2,N}=0$. In the empirical results we report below, we used a total of $N=12$ support points given by $\{A_1,\ldots,A_{12}\}=\{500,1000,\ldots,6500\}$, where the upper bound on the number of arrivals, $|A|=6500$, is over 3 times larger than the total hotel capacity in this market and the lower bound, 500, is less than one fourth of total capacity.

A key part of the model is constructing a good variable for x the demand shifter. As noted in the introduction, we used the *expected market occupancy rate* (i.e. the expected total occupancy for all 7 hotels in this market divided by their total capacity) as our proxy for the demand shifter x. Hotels have a good expectation of what the market level occupancy will be on different days, as well as the likely occupancy for their own hotel. As we noted in section 2, hotels consider predictable seasonal factors, weekly variations in occupancy, and holidays, as well as more "locally predictable" events such as whether large conventions or events will be taking place in the city or unusual weather or other shocks to demand that are likely to affect occupancy rates.

We used the STR data to regress daily level market occupancy rates on market occupancy rates on the same day one year in the future (i.e. by adding 364 days to the current date). The R^2 of this regression is 66% so this "predicted occupancy" $\hat{x}$ provides a proxy for whatever demand information x the hotels'

are actually using when it comes to their price setting decisions. Note that the actual x that hotels use is likely to reflect more information than is contained in our crude proxy for it, $\hat{x}$. If $\hat{x}$ is a poor proxy for the actual x, it could violate our key conditional independence assumption 2. That is, if actual demand and hotel prices depend on a value of x that we do not actually observe, then conditioning on a coarse proxy for $\hat{x}$ may not satisfy the conditional independence assumption 2 because there is information contained in the true latent x that leads to a correlation between hotel demand and the prices the hotels set that we cannot control for using only the crude proxy $\hat{x}$.

We deal with this problem using latent variable methods. We assume there is a relationship between our proxy for the demand shifter $\hat{x}$ and the true demand shifter x that can be captured by a conditional density $\phi(x|\hat{x},\delta)$ where δ is a vector of additional parameters to be estimated. We used the following simple linear Gaussian specification for $\phi(x|\hat{x},\delta)$:

$$x = \delta_0 + \delta_1 \hat{x} + \varepsilon \quad \varepsilon \sim N(0,\delta_2^2). \tag{41}$$

Since x is unobserved, it is clear that scale and location normalizations are required. In our analysis, we normalize $\delta_0 = 0$ and $\delta_1 = 1$ and treat $\hat{x}$ as an unbiased but noisy proxy for the actual demand shifter x that hotels use. Under this interpretation, $x = \hat{x} + \varepsilon$, and ε represents the additional information hotels receive about the likely market occupancy on a particular day, above and beyond the *ex ante* regression prediction $\hat{x}$. Though hotels have an initial expectation of market occupancy rates $\hat{x}_t$ based on past experience, just prior to setting their prices (in a simultaneous move game) at the start of each day t they collectively observe other information ε_t that affects their expectation of what the realized market occupancy will be that day. Then based on their total information consisting of $x_t = \hat{x}_t + \varepsilon_t$ and the idiosyncratic pricing shocks z_t, the hotels set their prices $p(x_t, z_t)$. Then A_t customers arrive, modeled as a draw from $H(A|x_t)$. Given the prices p_t the arriving customers independently choose their preferred hotel, or the outside good.

Using the implied normal distribution $\phi(x|\hat{x},\delta)$ we can "integrate out" the unobserved latent x and obtain likelihoods in terms of our observable proxy $\hat{x}$ while still continuing to posit that the original model with the true but unobserved demand shifter x satisfies the conditional independence assumption. In the case of the regression estimation, we posit a "reduced-form" relationship for the pricing strategies of the two hotels $(p_0(x,z_0), p_c(x,z_c))$ given by simple linear models with normally distributed error terms

$$\begin{aligned} p_0(x,z_0) &= \eta_{0,0} + \eta_{0,1}x + \eta_{0,2}x^2 + z_0 \quad z_0 \sim N(0,\eta_{0,3}^2) \\ p_c(x,z_c) &= \eta_{c,0} + \eta_{c,1}x + \eta_{c,2}x^2 + z_c \quad z_c \sim N(0,\eta_{c,3}^2). \end{aligned} \tag{42}$$

and we assume independence between z_0 and z_c: $E\{z_0 z_c\} = 0$. Thus, hotels 0 and c set their prices each day after observing the demand shifter x but only observing their own respective pricing shocks, z_0 and z_c,

resulting in the realized prices given in equation (42). Given these prices, realized demand is given by

$$\begin{aligned} d_0 &= \theta_{0,0} + \theta_{0,1}x + \theta_{0,2}x^2 + \theta_{0,3}p_0 + \theta_{0,4}p_c + \varepsilon_0 \quad \varepsilon_0 \sim N(0, \theta_{0,5}^2) \\ d_c &= \theta_{c,0} + \theta_{c,1}x + \theta_{c,2}x^2 + \theta_{c,3}p_c + \theta_{c,4}p_0 + \varepsilon_c \quad \varepsilon_c \sim N(0, \theta_{c,5}^2) \end{aligned} \tag{43}$$

where we also assume the demand residuals are independently distributed, $E\{\varepsilon_0\varepsilon_c\} = 0$, though this restriction can easily be relaxed.

To form the likelihood for the latent x case, we first condition on the true, unobserved x and apply the conditional independence assumption 2, and then integrate over x using the conditional density $\phi(x|\hat{x},\delta)$ to get the following likelihood for a single observation $(d_0, d_c, p_0, p_c, \hat{x})$

$$L(d_0, d_c, p_0, p_c, \hat{x}, \theta_0, \theta_c, \eta_0, \eta_c, \delta) = \int_x \phi(d_0|x, p_0, p_c, \theta_0)\phi(d_c|x, p_0, p_c, \theta_c)\phi(p_0|x, \eta_0)\phi(p_c|x, \eta_c)\phi(x|\hat{x}, \delta)dx \tag{44}$$

We also used the latent variable approach to estimate the full mixed, censored trinomial model $f(d|p,x)$. We used the same linear relationship between x and $\hat{x}$ given in equation (41) above, but also incorporate the reduced-form pricing relations in equation (42) to obtain the following likelihood $f(d|p,\hat{x},\theta,\gamma,\delta,\eta_0,\eta_c)$ given by

$$f(d|p,\hat{x},\theta,\gamma,\delta,\eta_0,\eta_c) = \int_x f(d|p_0,p_c,x,\theta,\gamma)\phi(p_0|x,\eta_0)\phi(p_c|x,\eta_c)\phi(x|\hat{x},\delta)dx. \tag{45}$$

Our empirical analysis compares estimated demand curves under the scenario where we assume that $x = \hat{x}$ (observed demand shifter) with the specification with a latent demand shifter given in equation (41). If $\hat{x}$ is not a good proxy for the true x, we expect that the estimated demand curve from the latent x specification will be more price-elastic.

4.1 Imposing optimality and equilibrium restrictions

Note that the estimation strategy discussed above is both instrument-free and it is also free of any assumptions about optimizing or equilibrium behavior of the firms. It is possible to perform estimation subject to these additional restrictions and use likelihood ratio tests to assess the validity of the assumptions of optimality and equilibrium. For example to impose the restriction of optimal pricing of hotel 0, we replace the reduced-form equation for $p_0(x,z_0)$ in (42) with the equation

$$p_0(x,z_0) = p_0^*(x) + z_0 \quad z_0 \sim N(0,\theta_{0,5}) \tag{46}$$

where $p_0^*(x)$ is the optimal price for hotel 0 given in (14) except we use $\phi(p_c|x,\theta_c)$, the normal density implied by the reduced-form pricing equation for hotel c in (42) as the conditional distribution $G_c(p_c|x)$

when calculating $p_0^*(x)$ in equation (14). In the case where we estimate the linear specification for demand in equation (43) we simply use the estimated linear demand curve in place of the conditional expectation $E\{d_0|p_0, p_c, x\}$ given in equation (11).

To impose equilibrium constraints in addition to optimality constraints, we need to solve for the Bertrand-Nash equilibrium functions $(p_0^*(x), p_c^*(x))$ as per equation (15) and use these instead of the unrestricted reduced-form equations for (p_0, p_c) in equation (42). In the case where demand is linear, (43), we can derive analytic expressions for $(p_0^*(x), p_c^*(x))$. However in the case where expected demand is calculated from the mixed trinomial model (37) there are no analytic closed-form expressions for $(p_0^*(x), p_c^*(x))$ and they must be calculated numerically. In this case full structural estimation requires the use of a nested fixed point algorithm (e.g. Rust (1987)) or the MPEC algorithm that estimates the model subject to the equilibrium constraints (e.g Su and Judd (2012)).

Notice that imposing optimality or equilibrium constraints, assuming these assumptions are correct, leads to more efficient efficient estimators. This is because of *cross equation restrictions* on the parameters. For example, when optimality is imposed, we no longer need to estimate the additional coefficients $(\eta_{0,0}, \ldots, \eta_{0,2})$ in the reduced-form equation for $p_0(x, z_0)$ (42). Instead, the conditional mean of $p_0(x, z_0)$ is equal to $p_0^*(x)$ which depends only on the structural coefficients (θ, γ) and also on δ if we assume the demand shifter x is latent. However if these assumptions are incorrect, they will generally lead to biased, inconsistent estimates of demand. Essentially, these assumptions coerce the estimated demand curves to be sufficiently price-elastic to "rationalize" the observed pricing of the hotels in the market.

4.2 Results – linear regression specification

In section 2 we showed there are systematic and predictable differences in hotel prices depending on the day of the week. During weekends the hotels have a much greater share of more price elastic leisure customers, whereas on weekdays there are relatively more price inelastic business and group customers. For example over the period of our sample (January 1, 2010 to October 31, 2013) the average share of business and group customers at hotel 0 is 47%. However on the prime weekdays (Monday, Tuesday and Wednesday) the share is 58%. This difference in composition is reflected in the ADRs: the average over all days is $192, but for the prime weekdays the average ADR is $206. Therefore, we opted to estimate our model for a subsample of the prime weekdays only, giving us a total of 575 daily observations on occupancy and ADR between January 1, 2010 and October 31, 2013.

We begin with the regression analysis of demand for hotel 0, illustrating how controlling for x enables us to estimate downward sloping demand curves, without instrumental variables or imposing any optimality or equilibrium restrictions. Table 4 shows the coefficient estimates for the linear model of demand

and reduced-form pricing strategies given in equations (42) and (43) above. The first column shows what happens when we do not include the demand shifter x: here the failure to control for endogeneity leads to positive and significant coefficient estimates of both p_0 and p_c. The next column shows that when we control for x (see middle column where we assume $x = \hat{x}$, the observed demand shifter case), we now obtain negative and significant coefficient estimates for $\theta_{0,3}$, the coefficient of p_0 in the demand curve for d_0. The next column shows the estimated parameters in the case where x is treated as a latent variable. We see that as we expected, the estimated value of $\theta_{0,3}$ is more negative and the estimated standard deviation of the unobserved "information shocks" that firms receive about market occupancy, ε, the δ_2 parameter in equation (41), is large and highly significant. The total share of the variance of the latent demand shifter x accounted for by unobserved shocks ε is 32%. However there is still significant uncertainty about *ex post* occupancy rates: the total variance of the latent demand shifter x relative to the variance of *ex post* total market occupancy rates is estimated to be 72%. We can think of this as akin to the "R^2" for this predictor of market occupancy, and thus, the latext demand shifter x can predict 72% of the variance in *ex post* market occupancy rates, whereas the *ex ante* regression predictor $\hat{x}$ only explains 54% of this variation.

We conclude that it is possible to obtain downward sloping demand curves by controlling for the effect of demand shifters x, but to do so it is important to have a good proxy for x. If we have a poor proxy, there will still be residual information ε that affects arrivals and firm prices, in violation of our conditional independence assumption 2. Notice that the fit of the model, whether measured by the log-likelihood value or the R^2 values for the linear model's prediction of occupancy in hotels 0 and c, d_0 and d_c, is significantly improved by controlling for x in both the observed x and unobserved x specifications. Thus we are able to easily reject the hypothesis that x does not affect hotel occupancy, and a likelihood ratio test also rejects the hypothesis that the demand shifter is observed and is fully captured by predicted occupancy $\hat{x}$.

However the linear specification is problematic in that even after controlling for the demand shifter x (whether observed or unobserved) the linear model still predicts that the expected occupancy at hotel c is an upward sloping function if its own price, p_c (see the coefficient estimates of $\theta_{c,3}$ in table 4). Table 5 presents estimates of the model when we impose optimality and equilbrium. Now we see that demand for both hotels 0 and c is downward sloping in their own prices, and demand is significantly more price elastic than in the specifications that do not impose optimality or equilibrium constraints. At the same time, we can also see from comparing the log-likelihood and R^2 values in tables 4 and 5 that we can decisively reject the optimality and equilibrium restrictions. We can also test the hypothesis that the hotels are colluding, perhaps "algorithmically" via the use of revenue management systems. We calculated what the jointly optimal prices would be for hotels 0 and c would be if they were operated by a single owner.

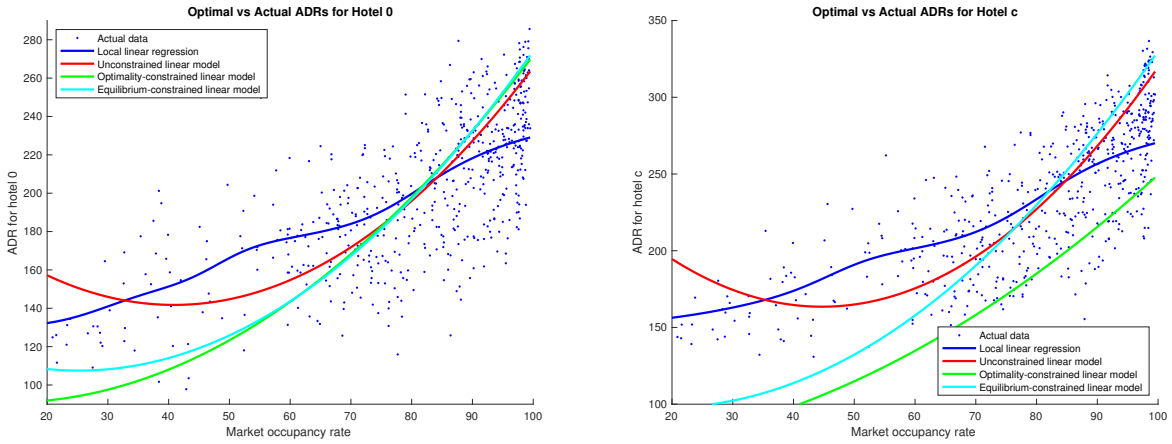


Figure 7: Predicted versus actual ADRs for hotels 0 and c

Even if we take the more elastic demand estimates that impose optimality or equilibrium restrictions, the optimal monopoly prices are *unbounded,* since the parameter estimates imply that the monopoly profit objective function (a quadratic form in the vector of prices (p_0, p_c)) is not negative definite.

Though it is tempting to conclude that the data strong reject the hypothesis of collusion, an alternative conclusion is that the data reject the linear specification of hotel occupancy and pricing given in equations (42) and (43). For example, figure 7 plots the actual prices set by the hotels as a function of x and compares a local linear regression fit to the linear regression predictions of prices for the prices set by hotels 0 and c under the unrestricted, optimality-restricted and equilibrium-restricted specifications. We see that none of the linear model predictions provide a good approximation to actual pricing over the full range of market occupancy rates x. Figure 8 plots predicted versus actual occupancy in hotels 0 and c as a function of the difference in price of the two hotels (i.e. $p_0 - p_c$ and $p_c - p_0$, respectively) for a subset of observations where the predicted market occupancy rate exceeds 90%. Though the unconstrained linear specification (red lines) provide a reasonable approximation to the local linear regression result (blue lines), the unconstrained demand is only downward sloping for hotel 0, but still upward sloping for hotel c, whereas the equilbrium and optimality constrained curves (green and cyan curves) are significantly more downward sloping comparing to the non-parametric result. Further, at sufficiently low relative prices, the linear model's predictions violate the capacity constraints for hotels 0 ($C_0 = 327$ and $C_c = 1824$, respectively). Evidently, the cross-equation constraints cause the estimation algorithm to trade off the model's ability to fit the hotel occupancy data (d_0, d_c) (by making demand more price elastic) in order to enable it to do a better job of fitting prices (p_0, p_c). Yet despite these restrictions, the linear model fails to

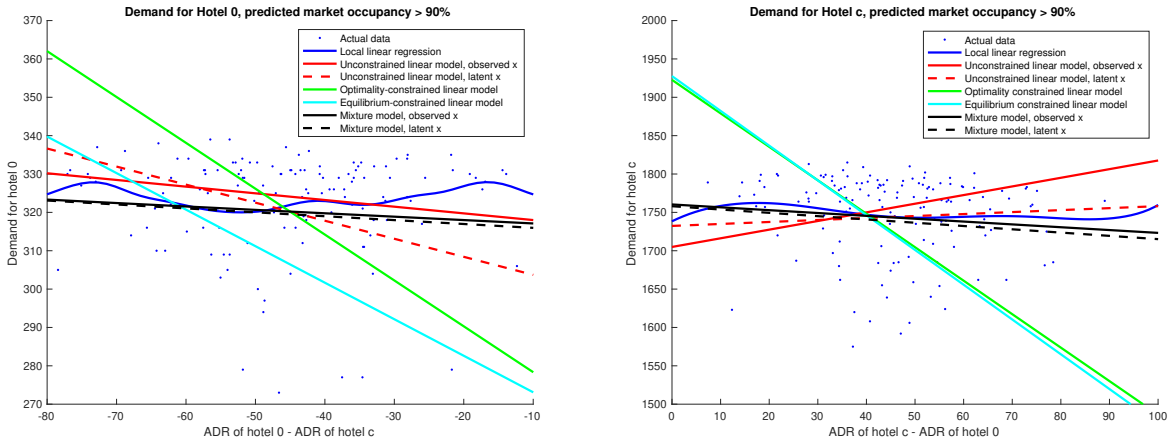


Figure 8: Predicted versus actual ADRs for hotels 0 and c

produce a credible counterfactual prediction for price collusion in this market.

4.3 Results – nonlinear mixture model specification

The problematic empirical findings of the previous section may be caused by our use of a simple linear model that misspecifies the expected demand function. Our structural model of consumer arrivals and discrete choices over hotels and the outside good results in a fundamentally probabilistic, nonlinear model of demand. Further the regression model is unable to identify the demand for the outside good, and as we show below, it predicts *hotel occupancy* rather than *hotel demand.* That is, the regression model was estimated from censored data that reflect the hotels' capacity constraints. The regression specification does not allow us to infer the number of customers who are rationed (i.e. the number who wanted to book rooms and were turned away due to capacity constraints), or the number of customers who chose the outside good.

Specifically, expected *ex ante* demand for hotel 0 is given by $E\{d_0|p_0,p_c,x\} = E\{A|x\}\pi_0(p_0,p_c)$, and it is simply the product of the expected number of arrivals $E\{A|x\}$ times the probability that an individual consumer will choose to book at hotel 0, $\pi_0(p_0,p_c)$. However *ex ante* demand ignores hotel 0's capacity constraint C_0, whereas expected occupancy $E\{d_0|p_0,p_c,x\}$ given in equation (11) of section reflects the fact that the capacity constraint must hold with probability 1. As we show below, when the capacity constraint is likely to be binding, expected occupancy will be far more price inelastic than expected *ex ante* demand. We now present estimates of a richer structural model of demand that can enable us to make predictions and inferences about these unobserved quantities. The structural mixture model enables

us to distinguish *ex ante demand for hotels* from *ex post occupancy.* We will show that it is crucial for hotels to know the former when making their *ex ante* pricing decisions, and that regression models that ignore this important distinction and confuse demand and occupancy can result in spuriously inelastic estimates of the underlying demand for hotels.

Table 6 presents the maximum likelihood estimates of our mixed trinomial model of hotel demand and occupancy. The first column presents the estimates under the assumption that the demand shifter x is observed and equals the predicted market occupancy rate $\hat{x}$. The second column presents the results for the version where we assume that the hotels base their pricing on a latent x which equals the predicted market occupancy rate $\hat{x}$ plus a normally distributed deviation ε. Based either on a Wald test or a likelihood ratio test, we see that we can strongly reject the hypothesis that $x = \hat{x}$: the estimated standard deviation of ε is about 4.3%, which amounts to nearly 25% of the standard deviation in the overall demand shifter x.

However the key finding is that the price coefficients $(\beta, \beta_\emptyset)$ (where β is the coefficient for the disuility of the price of hotel 0 or c, and $\beta_\emptyset$ is the price coefficient on the unobserved price of the outside good alternative) are both significantly negative, implying that expected demand for both hotel 0 and c is downward sloping in their own price but upward sloping in the price of their competitor, by Lemma 1 of section 3. We also see that the estimate of β is higher in the unobserved x specification, consistent with our hypothesis that failing to fully control for the true value of the demand shifter can result in an omitted variables problem that results in a downward biased estimate of β.

The coefficient estimates of the γ parameters for the flexible MNL specification for $H_N(A|x,\gamma)$ imply that the expected number of arrivals increase monotonically in x, and on average (i.e. unconditional on x) our estimates imply that about 5300 potential customers arrive on any given day wishing to book a room. However collinearity in the dummy variables for different numbers of arrivals results in large estimated standard errors for many of the γ parameters, but this is consistent with what we would expect from a sieve estimator of the large number of "nuisance parameters" in the semi-parametric component of the model. However notice that though the standard errors can be large, the point estimates are remarkably similar for the two specifications of the model so our overall inferences about the distribution $H(A|x)$ are remarkably robust to changes in the model specification.

Aside from the semi-parametric parameters γ, the other parameters of interest in the model θ are estimated fairly precisely, consistent with the asymptotic normality results of Wong and Severini (1991) for for the finite-dimensional components of the parameters of semi-parametric maximum likelihood estimators. Even though the information matrix is not block-diagonal in the parameter vector (θ, γ), we find that it is approximately block diagonal, so the high estimated variances of the γ parameters do not contaminate

and lead to high estimated variances of θ.

The preference coefficients, α_0 and α_c for both hotels 0 and c are estimated to have significantly negative values, which implies that about two thirds of the arriving customers choose the outside good. Of the one third who book a room at hotel 0 or c, about 15% book at hotel 0 and the remaining 85% book at hotel c. Overall the mixed trinomial model fits the occupancy data quite well, as indicated by the solid and dashed black curves in figure 8, which are the expected occupancies for hotels 0 and c, respectively, implied by the model. We see these estimates are quite close to the blue lines, which are the local linear regression estimates of how occupancy varies with prices.

However this is also a problematic aspect of these estimates, since the estimates of $(\beta, \beta_\emptyset)$ in table 6 result in fairly inelastic demand which in turn, implies that hotels 0 and c are *underpricing* — i.e. both hotels could substantially increase their profitability by raising their prices significantly. However the predicted optimal prices from our estimates of the demand model are implausibly large: for example if predicted market occupancy is $x = 0.8$, the predicted optimal price for hotel 0 is $p_0^* = 3007$, which increases the hotel's expected profits from $42,000 per night to nearly $251,000.

Another problematic aspect of this specification is it is unable to capture the variation in hotel 0's market share. On a daily basis, the market share of hotel 0 is 15% with a standard deviation of 2.24%. Our model implies variation in hotel 0's market share from two sources: a) the variation in aggregate occupancy rates x which leads to relative price variation in the ADRs of hotel 0 and c, which in turn causes variation in the choice probabilities $\pi(p)$, and thus variation in the *expected market share* of hotel 0, $\pi_0(p_0, p_c)$ driven by the daily variability in prices, and b) "random sampling variation" caused by the fact that the arriving customers make independent choices of which hotel to stay at (or the outside good). However given the inelastic estimates of demand, even fairly large relative price variation results in miniscule variability in the choice probabilities and the implied expected market share for hotel 0: the standard deviation in $\pi_0(p_0, p_c)$ driven by the daily variability in (p_0, p_c) is only 0.09. The additional variability in market shares due to random sampling variation only increases the standard deviation in hotel 0's market share to rise to 0.84, or less than half its actual value in the data.

4.4 Results – nonlinear mixture model specification with random coefficients

These findings motivate our final specification of the demand model: to allow *random coefficients.* The mixed trinomial model involves the implicit assumption that the preferences of customers arriving to the hotel market on different days are identical. However there is no reason to make this restriction: one can imagine that the set of customers arriving on different days are draws from different populations with different preferences for the hotels. For example, the types of customers who book hotel rooms in

order to attend a large professional meeting or convention may be very different from customers who are booking rooms in order to attend a football or baseball playoff game. Thus, we assume that in addition to observing the demand shifter x the hotels also observe information on the relative preferences of the customers arriving that day, and we assume this preference information is also common knowledge. In our example, let (ξ_0, ξ_c) be *preference shifters* that the hotels observe about variations in customers' preferences for hotel 0 and hotel c about their expected values (α_0, α_c), respectively. Thus, conditional on having observed (ξ_0, ξ_c) the hotels' beliefs about the probability that customers will choose to book at hotel 0 is given by

$$\pi_0(p, \xi_0, \xi_c) = \frac{\exp\{\alpha_0 + \xi_0 - \beta p_0\}}{\exp\{-\beta_\emptyset p_c\} + \exp\{\alpha_0 + \xi_0 - \beta p_0\} + \exp\{\alpha_c + \xi_c - \beta p_c\}}. \tag{47}$$

For simplicity we assume that (ξ_0, ξ_c) are normally and independently distributed with standard deviations given by ω_0 and ω_c, respectively. This is the simplest possible specification of a random coefficents model, however it is possible to estimate versions where ξ_0 and ξ_c depend on x or other variables, where the price coefficients $(\beta, \beta_\emptyset)$ might also have random variation, and it is possible to allow for serial correlation in the random coefficients across successive days (for example to capture correlations in customer types who attend conventions that last multiple days and thus book rooms over multiple days as well, etc).

Table 7 presents our maximum likelihood estimates of the random coefficients version of the mixed trinomial model of hotel occupancy. Similar to the model estimates in table 6 we present two columns the first under the restriction that $\hat{x} = x$ and the second for the latent demand shifter where $x = \hat{x} + \varepsilon$. The results in table 7 present a purely linear reduced form model of hotel pricing, where $p_0 = \eta_{0,0} + \eta_{0,1} x + \eta_{0,2}\xi_0 + z_0$, and similarly for p_c. That is, we omitted the quadratic terms in x^2 that we included in table 6.

The two key conclusions from table 7 are 1) the inclusion of random preference shifters (ξ_0, ξ_c) result in substantially more elastic demand, as reflected by the much more negative $(\beta, \beta_\emptyset)$ coefficients, and 2) the random coefficient specification fits the data significantly better, as reflected by the significantly higher values of the log-likelihood. Using a likelihood ratio test, we can easily reject the hypothesis of population homogeneity (i.e. no random coefficients, where we restrict the variance parameters $(\omega_{0,1}, \omega_{c,1}) = (0,0)$ so that $(\xi_0, \xi_c) = (0,0)$ with probability 1). Also comparing columns 1 and 2 of table 7 we see that the latent x specification fits the data significantly better than the observed x specification, so the best fitting specification is one where the demand shifter is the three-dimensional vector (x, ξ_0, ξ_c). Evidently, it is crucial to condition not only on the aggregate shock x but also on the preference shocks (ξ_0, ξ_c) to avoid an omitted variable problem that leads to spuriously inelastic demand estimates. For the latter specification we see that $(\hat{\beta}, \hat{\beta}_\emptyset) = (-3.37, -4.31)$ which are far more negative than in any of the other specifications we estimated. However aside from these key changes, broadly speaking the other parameter estimates do

not change a great deal in the random coefficients specification.

We now provide some evidence of the improved fit of the random coefficients model (using the best fitting specification in the second column of table 7). The random coefficients specification is able to fit the intertemporal variability in the market share for hotel 0. In the previous section we noted that the standard deviation in hotel 0's market share over the 575 days in our sample was 2.24% but the specification without random coefficients resulted in an implied standard deviation of only 0.09%. Once we include random coefficients, the implied standard deviation in hotel 0's market share increases to 2.69%. This variability is not driven exclusively by the standard deviation of the preference shocks (ξ_0, ξ_c): the increased price elasticity of the random coefficients specification implies that some of the variability in $\pi(p_0, p_c, \xi_0, \xi_0)$ is driven by the intertemporal variability in (p_0, p_c). Overall, the standard deviation of $\pi_0(p_{0,c}, \xi_0, \xi_c)$ increases to 2.26% in the random coefficients specification. Thus the additional variability due to "random sampling noise" (i.e. the fact that customers make indpenedent choices of which hotel to stay at) accounts for only a small share of the overall variability in hotel 0's market share.

The left panel of figure 9 shows expected market occupancy rates as a function of the *ex ante* predicted rate $\hat{x}$. The results are close to the 45 degree line, indicating that the estimated model implies that $\hat{x}$ is an approximately unbiased predictor of market occupancy rates, just as we found in the actual data. The figure plots x against the model's prediction of the *ex post* expected occupancy rate given x, $E\{(d_0 + d_c)|x\}/(C_0 + c_c)$, where expected occupancy given the *ex ante* predicted market occupancy rate x is given by

$$E\{(d_0 + d_c)|x\} = \int_{p_0} \int_{p_c} \sum_A \sum_{d_0} \sum_{d_c} (d_0 + d_c) f(d_0, d_c|A, p_0, p_c) H(A|x) \phi_0(p_0|x) \phi_c(p_c|x) dp_0 dp_c \qquad (48)$$

where $\phi_0(p_0|x)$ and $\phi_c(p_c|x)$ are the conditional distributions over prices of hotels implied by the Gaussian specification (42).

The right hand panel of figure 9 plots the *ex ante* expected unconstrained estimates of the nunber of arrivals, A, and the expected number of these consumers choosing the outside good, hotel c and hotel 0, respectively. The expected total number of customers arriving in this market increase from about 1500 when $x = 0.3$ to slightly over 6000 when $x = 1.0$. Thus, our estimation procedure is able to successfully identify the full distribution of arrivals $H(A|x)$ which constitutes the fundamental "demand shock" that drives the strong positive correlation in prices that we noted in section 2.

Figure 10 plots the implied cumulative distributions of occupancy at hotels 0 and c, expressed as CDFs over the occupancy rates. The blue lines in the graphs plot the empirical CDFs of the occupancy rate for different subsets of the 575 business days in our sample, depending on whether the *ex ante* expected market occupancy rate $\hat{x}$ was less the 60%, between 60% and 80%, between 80% and 90%, and greater

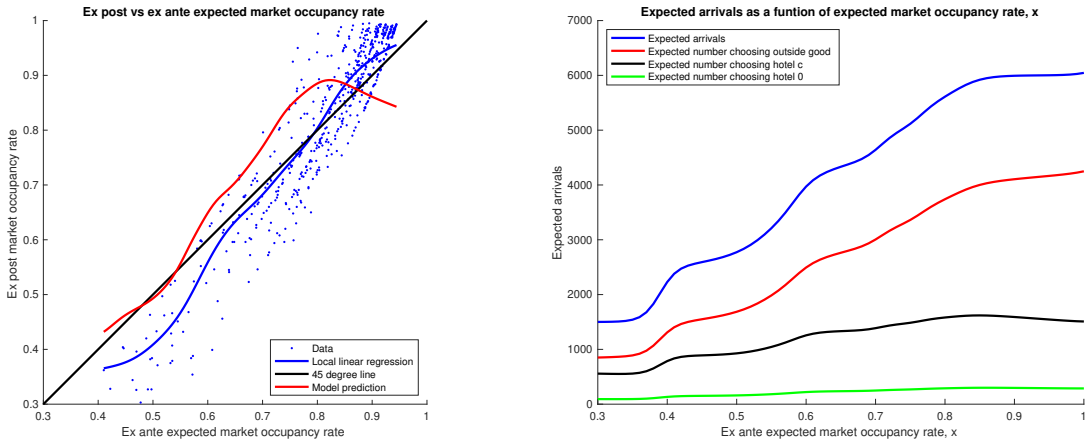


Figure 9: *Ex ante* vs *ex post* market occupancy rates, x, and expected arrivals $E\{A|x\}$

than 90% respectively. We conclude that our estimated model does a good job in matching the variability in occupancy both across different days with different values of x, and also it succeeds in matching the uncertainty the hotels face in what their occupancy will be even among subsets of days with similar values of x. The inclusion of the preference shocks (ξ_0, ξ_c) is a key to capturing this latter uncertainty.

Figure 11 provides deeper insight into how the random coefficients specification is able to capture the significantly negative demand elasticity for hotels. The key insight is that we should consider the data to be generated by a *probability distribution of demand curves* rather than by a fixed demand curve with an additive stochastic error term, such as in the linear specification (43). In figure 11 we plot the observed occupancy at hotel 0 as a function of the price difference, $p_0 - p_c$ for two subsets of our data set: 1) highest expected occupancy days, where $\hat{x} > 0.9$, and 2) high occupancy data, where $\hat{x} \in (.8, .9]$. The data points are the blue dots and visually, if we think of them as realizations around a single "expected demand curve" it is obvious that this way of conceptualizing and estimating demand will result in a fairly inelastic demand curve. This is captured by the solid blue line, the result of running a local linear regression on these subsets of the data set.

However figure 11 also plots green, black and red curves which are different *ex ante* expected demand curves and expected occupancy curves implied by our structural model for different realized values of the preference shifters, (ξ_0, ξ_c). For example, in the left hand panel of figure 11, the dotted green line is the *ex ante* expected demand for hotel when $(\xi_0, \xi_c) = (.25, -.15)$, i.e. when there is a large positive preference shock for hotel 0 and a negative shock for hotel c. This causes an upward shift in the *ex ante* demand which is significantly higher than hotel 0's available capacity C_0 for most of the range of values of $p_0 - p_c$

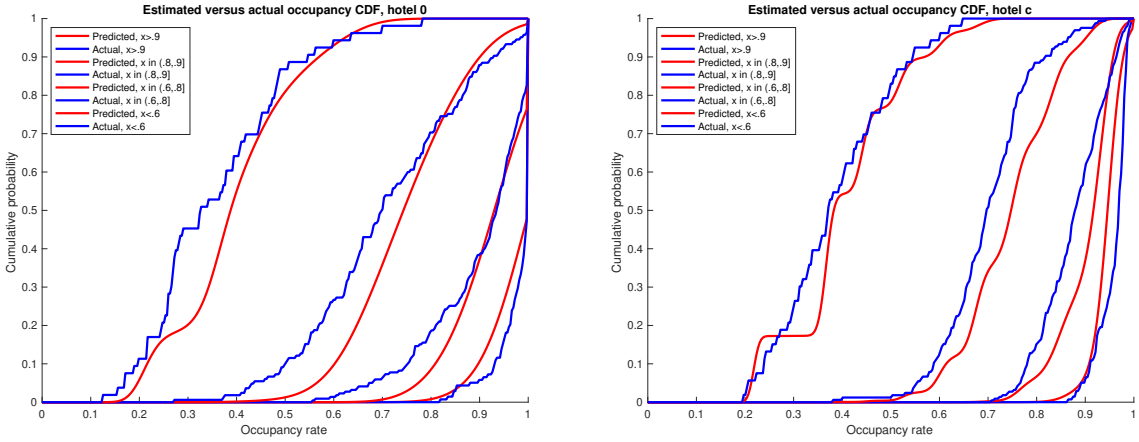


Figure 10: Predicted versus actual cumulative occupancy distributions

plotted in the figure. The solid green line in this same figure shows *expected occupancy* for the same set of shocks, i.e. we have enforced the capacity constraint that occupancy is less than C_0 with probability 1 in computing this expectation. Notice that the solid green curve is much flatter (and thus more inelastic) than the the dotted green curve, and this naturally occurs as a simple consequence of enforcing the capacity constraint. A key observation here, is that studies that fail to distinguish between expected demand and occupancy, and thus treat occupancy and demand as identical are likely to result in spuriously inelastic estimates of hotel demand, at least in the highest demand periods.

The black lines in the left hand panel of figure 11 plot expected demand and occupancy in the same aggregate market conditions (i.e. for the same $\hat{x}$ which equals the mean of $\hat{x}$ given $\hat{x} > .9$, or 0.922) but for a neutral preference shock $(\xi_0, \xi_c) = (0,0)$. In this case both the dashed and dotted black lines are virtually identical since this draw of preference shocks is not sufficiently positive to put hotel hotel 0 in danger of selling out for any of the values of the price differential $p_0 - p_c$ in the figure. Instead, expected demand and occupancy given (ξ_0, ξ_c) are shifted down relative to the green curves where hotel 0 expected a large positive preference shock. In summary, we can consider the hotel occupancy data as realizations of a complicated probability distribution that shifts significantly in response to predictable signals of overall arrivals, x, as well as relative shifts in market share due to different realized preference shocks (ξ_0, ξ_c). Instead of viewing the data as a scatterplot about a single expected demand curve, our model interprets the data as realizations from a family of probability distributions that shift in response to the demand shifters (x, ξ_0, ξ_c) and once we control for these demand shifters, demand is much more price elastic than would be inferred by a model that does not attempt to control for these price shifters (even though they may be

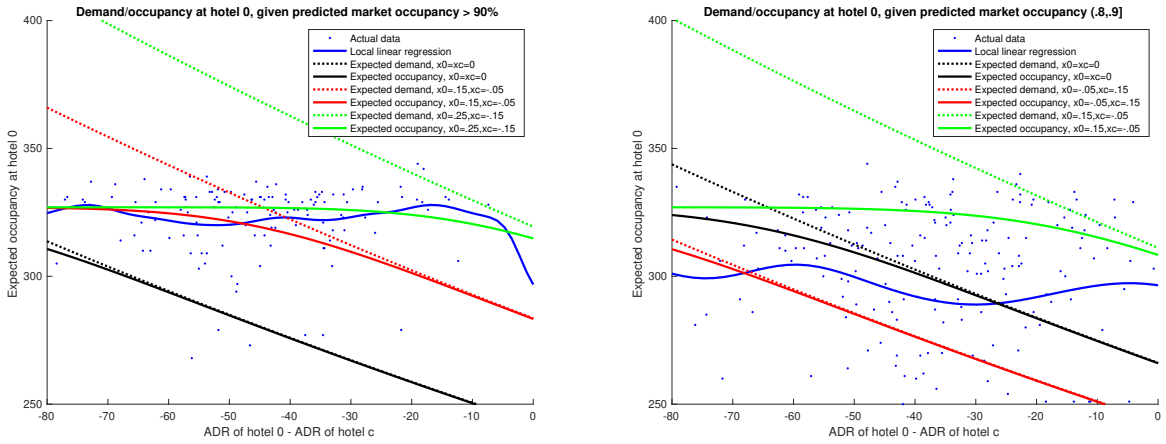


Figure 11: Stochastic *ex ante* demand and *ex post* occupancy implied by the random coefficients model

unobserved to the econometrician). The other key reason why simplistic regression approaches result in spuriously inelastic demand estimates is that they confound *occupancy* with *demand.* We have shown that in situations where capacity constraints are likely to be binding that expected occupancy is a far more inelastic function of hotel prices than expected *ex ante* demand.

We conclude our empirical analysis with several counterfactual calculations designed to address the question raised in the introduction: do we find evidence of "algorithmic collusion" due to the "price following behavior" we observed for hotels 0 and c? A related question is, are the prices we observe in this market consistent with optimizing behavior on the part of hotels 0 and c, and in particular, are the data consistent with the hypothesis of a Bertrand-Nash equilibrium in prices? As we noted in the introduction, our prior belief is that the price following behavior in this hotel market is not due to collusion but rather a natural competitive response by the hotels to ration scarce capacity in response to shifts to overall market demand that naturally result in a positive correlation between the prices different hotels charge and also a positive correlation between occupancy and price that we demonstrated in the figures in section 2. Figure 12 plots actual ADRs of hotels 0 and c versus the predictions from our model, along with two counterfactual curves: 1) an "optimal price" where we calculate the profit maximizing price implied by our estimated model of hotel demand, and 2) a "collusive price" where we assume that hotels 0 and c are owned by a single "multiproduct monopolist" that sets the ADRs to maximize the joint expected profits from both hotels. The dark blue line in the figures is a plot of a local linear regression fit to the occupancy data as a function of the *ex ante* expected market occupancy rate $\hat{x}$. The red lines are the predictions from our estimated structural model of hotel demand (i.e. they reflect the *status quo* pricing policies of the

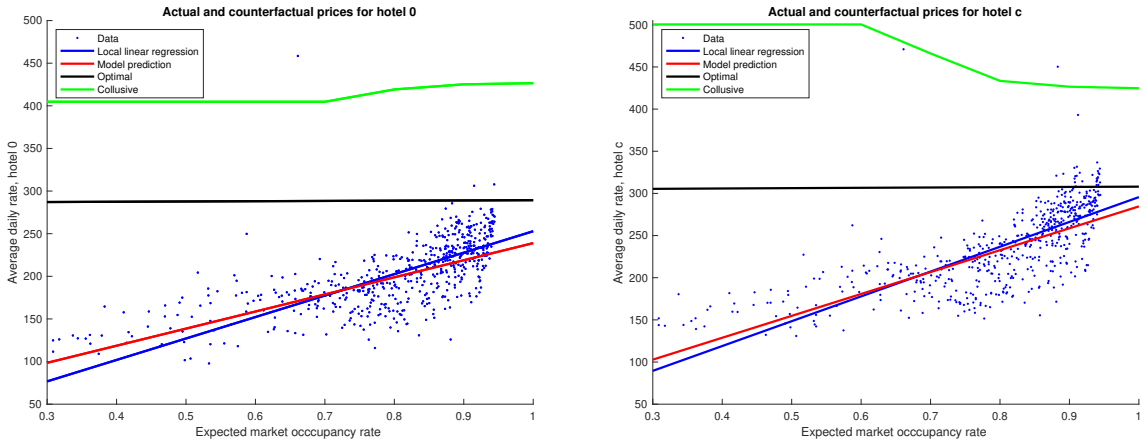


Figure 12: Actual versus counterfactual ADRs

hotels). The fact that the red and blue lines are nearly identical is another indication of the fact that our structural model fits the data well. In particular, our reduced-form model of hotel pricing correctly predicts that both hotels raise their price significantly to ration available capacity when they expect demand to be high (i.e. when x appproaches 1).

The green and black lines in figure 12 are counterfactuals. The black lines plot the optimal ADRs for hotels 0 and c respectively as a function of x. In making this calculation we assume that each hotel believes its competitor will not increase its price in response, but rather will continue to set its prices according to historical practice, i.e. according to the predictions of the red and blue lines in the figure. We conclude that both hotels are significantly *underpricing* and both could increase their profits significantly by raising their prices unilaterally. For example, the expected price hotel 0 charges under the *status quo* when $x = .9$ and $(\xi_0, \xi_c) = (0,0)$ is $E\{p_0|x\} = 218.88$ and this results in an expected profit of \$70788. However given the estimated preferences and demand, we calculate that it would be optimal for hotel 0 to charge \$289.01 and this would result in a expected profit of \$73174, a 3.3% increase.

The relatively high optimal prices we calculate are a direct consequence of our relatively inelastic estimates of hotel demand. Our model predicts that even on days where x is well below 1 and hotels expect to have significant excess capacity, their customers (who are largely business customers) are sufficiently price inelastic that it is better for the hotels to keep high prices than to cut prices and try to gather a higher share of the market. The actual pricing behavior of the hotels, on the other hand, appears focused on price cutting when market demand is low, and this price cutting behavior leads to the upward sloping price curves under the *status quo* that are reflected in figure 12.

Our demand estimates also imply that the substitution effect to hotel 0's competitors (and the outside good) is second order in comparison to the "own price" effect. Thus, the best response values we have calculated in figure 12 do not increase significantly if we were to allow hotel c to respond to hotel 0's price increase. We also calculated the Bertrand-Nash equilibrium prices and they are only slightly higher than the optimal prices plotted in figure 12. Thus, we conclude that we can simultaneously reject the hypotheses that hotel 0 and hotel c are behaving individually optimally and according to the predictions of Bertrand Nash equilibrium. Instead both hotels appear to be *underpricing* for reasons that are not fully clear to us. Hotel 0's revenue manager noted that her own impression is that the IdeaS revenue management system that hotel 0 subscribes to does tend to underprice, and to sucn an extent that she wondered whether the system had been programmed to *maximize occupancy* rather than *maximize profits.*

The green curves in figure 12 plot the collusive (joint profit maximizing) prices for hotels c and 0. Not surprisingly these are even higher than the black curves that plot the optimal prices for hotels 0 and c under the hypothesis that these hotels price independently and competitively. Given the evidence already presented that suggests hotels 0 and c are underpricing relative to the benchmark of competitive (Bertrand Nash) pricing, it seems clear *a fortiori* that there is little evidence of "algorithmic collusion" in this market. Instead, it seems possible that the RMS that these hotels are using actually lead to *lower* rather than higher prices for consumers.

5 Conclusion

We have shown that it is possible to estimate structural models of demand under challenging conditions: price endogeneity due to aggregate demand shocks that create positive correlation in prices and occupancy (upward sloping demand), truncated data on consumers (i.e. we do not observe the number of arrivals or the number of consumers choosing the outside good), and censoring (we only observe the minimum of hotel demand and the hotel's capacity). In the hotel example there are no relevant instrumental variables to deal with the endogeneity problem. Even if there were, it is not clear how instrumental variables can be used in a model where demand is a nonlinear price and state-dependent stochastic process, and not a linear demand curve for which instrumental variable methods can be applied. The main strategy previously used to deal with these challenges is to impose the hypotheses of optimality and equilibrium. These create "cross equation restrictions" that enable structural models to generate downward sloping estimated demand curves. However these assumptions can be problematic and in our empirical analysis we strongly reject them. The restrictions distort the estimated demand curves by forcing them to be more price elastic in order to rationalize firms' pricing decisions.

Perhaps an even more problematic aspect of the strategy of imposing optimality and equilibrium assumptions is that doing so makes it impossible to test whether these assumptions are valid. Our approach is motivated by a huge literature on bounded rationality dating back to the work of Herbert Simon that questions the ability of individuals and firms to perfectly solve difficult optimization problems. In the case of hotels, we discussed a rapidly growing industry on revenue management systems. If firms were the perfect optimizers that economic theories presume they are, there would seem to be little need for the RMS industry. We cited literature in the introduction that pricing decisions are often poorly managed, and showed evidence that RMS systems can help firms to significantly increase their profitability. But RMS systems are essentially "black boxes" and little is known about their internal mechanics and how they are able to produce sensible recommended prices in real time. There have been a number of claims that RMS result in impressive gains in profitability, but few independent evaluation of these claims. How do we know that the firms providing the RMS services are certifiably better at solving difficult pricing problems than expert human revenue managers?

As Rust (2019) noted, in order to behave optimally, firms need to solve two types of learning problems. Ultimately the firm wants to learn the optimal strategy which in the case of hotels means learning the optimal pricing strategy (i.e. how to set its prices under different states of the world). But in order to to solve this problem firms need to solve a deeper learning problem, i.e. to understand the environment in which they operate. In the case of hotels, this means learning the preferences of their customers, the stochastic process by which customers arrive to the market, as well as the pricing strategies used by their competitors. Our analysis suggests that learning about demand is an extremely challenging problem. It is not clear how an RMS that provides pricing advice to hotels operating in many of thousands of different local hotel markets around the world is able to learn about demand in these local markets if it only has access to fragementary information on prices and occupancy of all the competing hotels in these markets.

We analyzed a specific hotel market where we have access to an unsually good data set: one that provides prices and occupancy rates not only on the hotel in question (hotel 0) but also on its competitors. But even our data is rather limited in its detail and we do not have data on the occupancy rates of individual competitors, only the sum of the occupancies of all competing hotels. Our analysis of identification has shown that for many counterfactual calculations, having good data on the joint distribution of occupancy is critical. In particular, we have shown that in order to calculate optimal prices, a hotel needs to have a good knowledge of: 1) consumer preferences, 2) the distribution of arrivals, and 3) the degree of substitution not only to competing hotels, but also of the outside good. We showed that traditional regression approaches that focus only on estimating expected occupancy at a single hotel, and fail to distinguish between demand

and occupancy and do not attempt to identify the distribution of arrivals are severely hampered in making reasonable counterfactual predictions, and one of the most important counterfactuals is predict the optimal price for a hotel given what can be learned about demand.

This paper has introduced a new approach that at its heart, involves conditioning on a vector of *demand shifters* x that capture the endogeneity between prices and demand. We showed that it is possible to identify the underlying structure of demand with sufficiently good data (i.e. data on prices and joint occupancies of the hotels) under a *conditional exogeneity assumption* that posits that there may be other unobserved information z that affects firm pricing, but once we condition on x and prices $p(x,z)$, this other information z does not affect the probability distribution for demand, $f(d|x,p)$. In other words, conditional on x any residual variability in prices due to z can be regarded as "virtual price experiments" that can be exploited to learn about demand. We have shown that is it not necessary for the econometrician to actually observe the demand shifter x in order to make valid inferences: it is possible to treat x as a latent variable provided we have sufficiently good proxies $\hat{x}$ for the underlying latent demand shifter x.

While our approach does enable us to relax optimality and equilibrium assumptions and is correctly described as "instrument-free" it is far from being "assumption-free." The success of our approach depends on whether our conditional exogeneity assumption is testable, and similar to the treatment effects literature, this assumption is not easily testable. Our approach also depends on either being able to observe x or having a sufficiently good proxy for it, $\hat{x}$, combined with additional assumptions to implement a latent variable approach that "integrates out" the stochastic demand curve with respect to a flexible conditional distibution $g(x|\hat{x})$.

In our analysis, we relied heavily on latent variable methods (particularly a random coefficients specification of consumer preferences for the preference shocks (ξ_0, ξ_c)) to obtain estimates of the distribution of demand that were sufficiently price elastic. However critical readers may feel that our estimated model is still rather price-inelastic and that some of our key conclusions — such as the predicted "underpricing" by the hotels — is actually an artifact of a failure of our conditional exogeneity assumption. We also acknowledge that our static model of hotel pricing is an oversimplification of reality: hotels are actually doing *dynamic pricing* and adjusting their BARs many times each day in response to new information that includes new reservations and cancellations of existing reservations. Thus, our use of the ADR as the "the price" that a hotel sets each day is a huge oversimplication: there is actually much greater price variation in the dynamics of the BAR and price discrimination of different types of consumers (e.g. group discounts, etc) that cannot be reflected in our static analysis of hotel pricing.

In a related paper Cho et al. (2018) do solve for optimal dynamic prices (i.e. the BARs) in a fully

dynamic model. Using a similar approach to the one described here, they relax the assumption of optimal pricing and use the greater variation in BAR prices to identify a dynamic model of demand for hotels. The greater price variation in the BARs results in more eleastic demand estimates, and this leads to a more complicated prediction of counterfactual optimal prices. In particular, their model predicts that the optimal BAR should be higher than the BARs that hotel 0 sets for reservations made more than 20 days before the arrival date, but prices should be significantly cut below the BAR that hotel 0 sets for resrvations made within 20 days of arrival. This counterfactual dynamic pricing schedule is able to increase hotel 0's expected profits from 10 to 15%, and it exploits the fact that the relatively more price-inelastic customers (e.g. business customers) tend to book their reservations more than 20 days in advance of arrival, whereas the more price-elastic customers (e.g. leisure customers) tend to book within 20 days of arrival.

Thus it is possible that our static model suffers from a problem of attenuation bias due to errors-in-variables in using the ADR as the price customers pay, when customers actually pay the BAR, which has much greater variability. So a combination of specification error from using a static model to study a fundamentally dynamic pricing problem plus errors in variables may contribute to the relatively inelastic demand estimates that we obtained in our paper. Nevertheless, we believe it is better to communicate the essential ideas underlying our approach in the simplest possible setting: there is substantially greater notational complexity to describe our approach in a dynamic setting and the estimation problem is much more challenging. Instead of being able to use the full information maximum likelihood approach that we used in this paper, the dynamic analysis does not admit a likelihood function. Instead we have to resort to a method of simulated moments (MSM) estimation approach, and our experience with MSM is that the objective function can have many local optimal, making the identification problem and the search for a global minimum of the estimation criterion much more challenging.

Finally, since the goal of this paper is to *learn about demand* it may be useful in future work to explore the relation between our approach and the currently fashionable literature on machine learning. In the taxonomy of Rust (2019) our approach is an example of a traditional econometric modeling approach that depends critically on *human input and intuition.* However our econometric model (which in the traditional econometric parlance is a multi-level finite mixture model) bears some resemblance to a *deep neural network* because the distribution of hotel occupancy generated from our modeling approach is a complicated multi-level mixture of logits, which is similar to a neural network model with multiple "hidden layers." For example, we can add an additional "layer" to our demand model by estimating a lower level mixture model to identify the preferences of different types of consumers, see equation (3) of section 3.

There is a new literature on *deep choice models* (see, e.g. Mottini and Acuna-Agost (2017)) that tries to bridge the econometric literature on discrete choice with the literature on deep neural networks. Deep nets and other methods from the machine learning literature offer the possibility of greater flexbility in fitting the data and superior capability for prediction. However a drawback of deep neural networks is that they are regarded as "black boxes" that are difficult to intrepret. The study by Mottini and Acuna-Agost (2017) proposed a deep choice model to predict consumer choice of different airplane flights using anonymized booking data collected by the Amadeus Global Distribution System (an intermediary that executes bookings made by travel agents on airlines). They report that their deep choice model outperforms a standard MNL model in terms of predictive accuracy, but they do not report whether their deep choice model can be aggregated to produce accurate predictions of demand for air flights, and particularly, whether the deep choice model is capable of accurate *counterfactual* predictions such as how demand for a particular flight depends on the airfare charged by the airline.

As we noted above, commercial RMSs are widely regarded (at least in the academic community) as proprietary black boxes: it is not clear how they generate recommended prices and whether these recommended prices are optimal. It remains to be seen whether machine learning algorithms are capable, with a minimum level of human supervision, of learning about consumer demand and setting optimal prices in response to that knowledge. We believe that the most critical input to an effective RMS is an accurate model of demand. In order for firms to have confidence in these models (particularly for counterfactual predictions such as setting optimal prices), the models of demand should be intuitive, transparent and interpretable. In other words, the "black box" aspect of many machine learning approaches is an important drawback that limits their credibility and level of acceptance. As a result, we believe that human input, intuition, and modeling expertise and the use more economically interpretable structural estimation methods will remain the preferred method for learning about consumer preferences and producing credible models of consumer demand for the forseeable future.

References

Benoit, J., & Krishna, V. (1987). Dynamic duopoloy: Prices and quantities. *Review of Economic Studies*, *54*, 23–36.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *63*(4), 841–890.

Cho, S., Lee, G., Rust, J., & Yu, M. (2018). *Optimal dynamic hotel pricing.*

Davidson, C., & Deneckere, R. (1990). Excess capacity and collusion. *International Economic Review*, *31*(3), 521–541.

Ezrachi, A., & Stucke, M. E. (2016). *Virtual competition: The promise and perils of the algorithm-driven economy*. Harvard University Press.

Fox, J., Kim, K., Ryan, S., & Bajari, P. (2012). The random coefficients logit model is identified. *Journal of Econometrics*, *166*, 204–212.

Gallego, G., & van Ryzin, G. (1994). Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, *40*(8), 999–1020.

Hall, G., & Rust, J. (2019). *Econometric methods for endogenously sampled time series: The case of commodity price speculation in the steel market* (Tech. Rep.). Georgetown University.

Harrington, J. (2017). *Developing competition law for collusion by autonomous price-setting agents* (Tech. Rep.). Wharton School of Business.

Kitamura, Y., & Laage, L. (2018). *Nonparametric analysis of finite mixtures.*

MacKay, A., & Miller, N. (2018). *Demand estimation in models of imperfect competition.*

McAfee, P., & te Veld, V. (2008). Dynamic pricing with constant demand elasticity. *Production and Operations Management*, *17*(4), 432–438.

McFadden, D. L. (1998). A method of simulated moments for estimation of discrete choice models without numerical integration. *Econometrica*, *57*(5), 995–1026.

Merlo, A., Ortalo-Magne, F., & Rust, J. (2015). The home selling problem: Theory and evidence. *International Economic Review*, *56*(2), 457–484.

Mottini, A., & Acuna-Agost, R. (2017). *Deep choice model using pointer networks for airline itinerary prediction.*

Phillips, R. L. (2005). *Pricing and revenue optimization*. Stanford University Press.

Rust, J. (1987). Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica*, *55*(5), 993–1033.

Rust, J. (2019). Has dynamic programming improved decision making? *Annual Review of Economics*, *11*, 833–858.

Su, C.-L., & Judd, K. L. (2012). Constrained optimization approaches to estimation of structural models. *Econometrica*, *80*(5), 2213–2230.

Wong, W., & Severini, T. (1991). On maximum likelihood in infinite-dimensional parameter spaces. *Annals of Statistics*, *19*(2), 603–632.

Table 4: Maximum likelihood estimates of hotel regression model

Parameter	No x in d_0 or d_c	Observed x ($x = \hat{x}$)	Latent x
$\theta_{0,0}$ (constant)	41.41 (19.94)	-91.80 (45.80)	-39.94 (47.41)
$\theta_{0,1}$ (x)		4.14 (1.27)	3.57 (1.31)
$\theta_{0,2}$ (x^2)		0.001 (0.009)	0.008 (0.009)
$\theta_{0,3}$ (ADR_0)	0.058 (0.223)	-0.174 (0.103)	-0.27 (0.11)
$\theta_{0,4}$ (ADR_c)	0.905 (0.207)	0.234 (0.096)	0.078 (0.108)
$\theta_{0,5}$ ($\text{std}(\varepsilon_0)$)	53.22 (2.11)	34.98 (0.88)	32.99 (0.95)
$\theta_{c,0}$ (constant)	280.95 (110.82)	-327.37 (123.02)	-145.35 (120.87)
$\theta_{c,1}$ (x)		19.07 (3.33)	18.47 (3.35)
$\theta_{c,2}$ (x^2)		0.019 (0.024)	0.041 (0.025)
$\theta_{c,3}$ (ADR_c)	4.55 (1.05)	1.126 (0.244)	0.382 (0.328)
$\theta_{c,4}$ (ADR_0)	0.496 (1.201)	-0.669 (0.262)	-1.16 (0.310)
$\theta_{c,5}$ ($\text{std}(\varepsilon_c)$)	225.38 (11.86)	98.67 (2.78)	77.91 (4.99)
$\eta_{0,0}$ (constant)	200.96 (48.60)	200.96 (48.60)	176.57 (38.83)
$\eta_{0,1}$ (x)	-2.89 (1.42)	-2.89 (1.42)	-2.27 (1.13)
$\eta_{0,2}$ (x^2)	0.035 (0.010)	0.035 (0.01)	0.031 (0.008)
$\eta_{0,3}$ ($\text{std}(z_0)$)	28.67 (0.73)	28.67 (0.73)	26.02 (0.97)
$\eta_{c,0}$ (constant)	265.37 (55.20)	265.37 (55.20)	228.27 (50.45)
$\eta_{c,1}$ (x)	-4.56 (1.59)	-4.56 (1.59)	-3.55 (1.43)
$\eta_{c,2}$ (x^2)	0.051 (0.011)	0.051 (0.011)	0.044 (0.009)
$\eta_{c,3}$ ($\text{std}(z_c)$)	31.55 (0.85)	31.55 (0.85)	27.99 (1.07)
δ_2 ($\text{std}(\varepsilon)$)	0.00 (0.00)	0.00 (0.00)	3.22 (0.13)
R^2, d_0	.426	.752	.740
R^2, d_c	.528	.910	.896
R^2, p_0	.502	.502	.501
R^2, p_c	.573	.573	.571
Log-likelihood	-10465.1	-9748.9	-9718.5

Table 5: Optimality and Equilibrium-constrained maximum likelihood estimates of hotel regression model

Parameter	Optimality constrained	Equilibrium constrained
$\theta_{0,0}$ (constant)	-10.48 (43.94)	315.03 (49.82)
$\theta_{0,1}$ (x)	2.99 (1.16)	-2.86 (1.27)
$\theta_{0,2}$ (x^2)	0.007 (0.008)	0.056 (0.009)
$\theta_{0,3}$ (ADR_0)	-1.196 (0.108)	-0.952 (0.125)
$\theta_{0,4}$ (ADR_c)	0.988 (0.112)	1.9×10^{-28} (0.132)
$\theta_{0,5}$ ($\text{std}(\varepsilon_0)$)	39.51 (1.25)	49.24 (2.30)
$\theta_{c,0}$ (constant)	458.70 (170.77)	566.37 (175.43)
$\theta_{c,1}$ (x)	1.81 (4.69)	-0.054 (4.68)
$\theta_{c,2}$ (x^2)	0.156 (0.035)	0.173 (0.035)
$\theta_{c,3}$ (ADR_c)	-4.36 (0.31)	-4.53 (0.32)
$\theta_{c,4}$ (ADR_0)	4.24 (0.51)	4.05 (0.48)
$\theta_{c,5}$ ($\text{std}(\varepsilon_c)$)	144.39 (6.56)	147.42 (7.00)
$\eta_{0,0}$ (constant)	94.81	126.40
$\eta_{0,1}$ (x)	-0.63	-1.50
$\eta_{0,2}$ (x^2)	0.024	0.030
$\eta_{0,3}$ ($\text{std}(z_0)$)	30.17 (1.41)	29.83 (1.48)
$\eta_{c,0}$ (constant)	101.17	-13.56
$\eta_{c,1}$ (x)	-1.19	-0.68
$\eta_{c,2}$ (x^2)	0.035	0.032
$\eta_{c,3}$ ($\text{std}(z_c)$)	35.24 (2.24)	36.30 (2.31)
c_0	-20.8 (19.45)	-78.01 (34.91)
c_c	-98.01 (21.29)	-85.45 (22.40)
R^2, d_0	.684	.508
R^2, d_c	.806	.798
R^2, p_0	.449	.462
R^2, p_c	.467	.434
Log-likelihood	-10130.5	-10279.4

Table 6: Maximum likelihood estimates of mixed trinomial demand model

Parameter	Observed x	Latent x
α_0	-2.61 (0.007)	-2.59 (0.007)
α_c	-0.91 (0.007)	-0.88 (0.007)
β	-0.291 (0.085)	-0.335 (0.085)
$\beta_{\emptyset}$	-0.298 (0.087)	-0.258 (0.086)
δ_2	0.00 (0.00)	0.032 (0.003)
$\eta_{0,0}$ (constant)	187.74 (46.48)	187.54 (38.55)
$\eta_{0,1}$ (x)	-2.50 (1.37)	-2.50 (1.12)
$\eta_{0,2}$ (x^2)	0.032 (0.009)	0.032 (0.008)
$\eta_{0,3}$ (std(z_0))	28.70 (0.71)	26.75 (0.80)
$\eta_{c,0}$ (constant)	251.08 (52.35)	250.54 (43.32)
$\eta_{c,1}$ (x)	-4.13 (1.50)	-4.12 (1.22)
$\eta_{c,2}$ (x^2)	0.048 (0.010)	0.048 (0.008)
$\eta_{c,3}$ (std(z_c))	31.52 (0.86)	28.44 (0.98)
$\gamma_{1,1}$ ($I\{A = 1500\}$)	137.49 (3959)	137.54 (20704)
$\gamma_{1,2}$ ($I\{A = 2000\}$)	122.00 (3957)	122.91 (20690)
$\gamma_{1,3}$ ($I\{A = 2500\}$)	99.00 (3954)	98.58 (20672)
$\gamma_{1,4}$ ($I\{A = 3000\}$)	88.97 (3953)	89.29 (20666)
$\gamma_{1,5}$ ($I\{A = 3500\}$)	57.59 (3951)	58.00 (20660)
$\gamma_{1,6}$ ($I\{A = 4000\}$)	29.02 (3950)	32.09 (20654)
$\gamma_{1,7}$ ($I\{A = 4500\}$)	13.99 (3950)	9.70 (20651)
$\gamma_{1,8}$ ($I\{A = 5000\}$)	-51.31 (3951)	-38.37 (20657)
$\gamma_{1,9}$ ($I\{A = 5500\}$)	-88.55 (3953)	-94.71 (20678)
$\gamma_{1,10}$ ($I\{A = 6000\}$)	-127.03 (3955)	-133.20 (20703)
$\gamma_{1,11}$ ($I\{A = 6500\}$)	-188.77 (3963)	-189.66 (20756)
$\gamma_{2,1}$ ($x * I\{A = 1000\}$)	-244.05 (18021)	-244.62 (29503)
$\gamma_{2,2}$ ($x * I\{A = 1500\}$)	-484.57 (37.42)	-488.03 (135.67)
$\gamma_{2,3}$ ($x * I\{A = 2000\}$)	-440.83 (30.68)	-446.32 (133.86)
$\gamma_{2,4}$ ($x * I\{A = 2500\}$)	-391.30 (26.58)	-393.11 (140.02)
$\gamma_{2,5}$ ($x * I\{A = 3000\}$)	-372.81 (26.86)	-376.04 (144.12)
$\gamma_{2,6}$ ($x * I\{A = 3500\}$)	-317.27 (23.13)	-320.98 (132.94)
$\gamma_{2,7}$ ($x * I\{A = 4000\}$)	-270.76 (21.44)	-278.64 (134.46)
$\gamma_{2,8}$ ($x * I\{A = 4500\}$)	-248.14 (20.57)	-245.22 (136.44)
$\gamma_{2,9}$ ($x * I\{A = 5000\}$)	-161.17 (17.12)	-180.10 (122.54)
$\gamma_{2,10}$ ($x * I\{A = 5500\}$)	-114.81 (14.00)	-109.30 (91.63)
$\gamma_{2,11}$ ($x * I\{A = 6000\}$)	-69.03 (10.92)	-63.12 (61.65)
Log-likelihood	-11666.2	-11608

Table 7: Maximum likelihood estimates of mixed trinomial demand model with random coefficients

Parameter	Observed x	Latent x
α_0	-2.81 (0.022)	-2.91 (0.020)
α_c	-1.15 (0.023)	-1.09 (0.020)
β	-1.44 (0.37)	-3.37 (0.27)
$\beta_\emptyset$	-2.55 (0.35)	-4.31 (0.27)
δ_2	0.00 (0.00)	0.040 (0.006)
$\omega_{0,1}$ (std(ξ_0))	0.164 (0.005)	0.153 (0.003)
$\omega_{c,1}$ (std(ξ_c))	0.056 (0.002)	0.043 (0.002)
$\eta_{0,0}$ (constant)	39.57 (16.32)	38.33 (13.66)
$\eta_{0,1}$ (x)	201.62 (19.04)	200.61 (16.66)
$\eta_{0,2}$ (ξ_0)	1.64 (42.57)	1.59 (228.94)
$\eta_{0,3}$ (std(z_0))	30.00 (0.81)	28.24 (0.95)
$\eta_{c,0}$ (constant)	26.02 (18.03)	24.92 (17.19)
$\eta_{c,1}$ (x)	260.53 (21.64)	259.67 (21.65)
$\eta_{c,2}$ (ξ_c)	4.04 (634.88)	4.05 (4107)
$\eta_{c,3}$ (std(z_c))	33.98 (1.08)	31.83 (1.46)
$\gamma_{1,1}$ ($I\{A = 1500\}$)	126.71 (8209)	128.59 (21435)
$\gamma_{1,2}$ ($I\{A = 2000\}$)	110.73 (8208)	105.98 (21423)
$\gamma_{1,3}$ ($I\{A = 2500\}$)	97.62 (8205)	97.37 (21418)
$\gamma_{1,4}$ ($I\{A = 3000\}$)	85.03 (8203)	85.04 (21409)
$\gamma_{1,5}$ ($I\{A = 3500\}$)	63.46 (8202)	62.90 (21404)
$\gamma_{1,6}$ ($I\{A = 4000\}$)	29.44 (8200)	28.79 (21390)
$\gamma_{1,7}$ ($I\{A = 4500\}$)	12.32 (8199)	11.92 (21387)
$\gamma_{1,8}$ ($I\{A = 5000\}$)	-37.66 (8200)	-36.91 (21401)
$\gamma_{1,9}$ ($I\{A = 5500\}$)	-91.53 (8207)	-89.99 (21416)
$\gamma_{1,10}$ ($I\{A = 6000\}$)	-133.79 (8217)	-131.31 (21445)
$\gamma_{1,11}$ ($I\{A = 6500\}$)	-193.06 (8226)	-193.09 (21494)
$\gamma_{2,1}$ ($x * I\{A = 1000\}$)	-242.92 (11881)	-242.92 (31639)
$\gamma_{2,2}$ ($x * I\{A = 1500\}$)	-479.06 (154)	-478.42 (2432)
$\gamma_{2,3}$ ($x * I\{A = 2000\}$)	-431.79 (152)	-433.61 (2431)
$\gamma_{2,4}$ ($x * I\{A = 2500\}$)	-397.96 (152)	-398.33 (2432)
$\gamma_{2,5}$ ($x * I\{A = 3000\}$)	-373.88 (151)	-374.14 (2430)
$\gamma_{2,6}$ ($x * I\{A = 3500\}$)	-333.15 (152)	-333.53 (2428)
$\gamma_{2,7}$ ($x * I\{A = 4000\}$)	-274.87 (149)	-275.54 (2426)
$\gamma_{2,8}$ ($x * I\{A = 4500\}$)	-247.80 (149)	-248.53 (2425)
$\gamma_{2,9}$ ($x * I\{A = 5000\}$)	-179.77 (148)	-179.47 (2419)
$\gamma_{2,10}$ ($x * I\{A = 5500\}$)	-111.30 (146)	-110.04 (2417)
$\gamma_{2,11}$ ($x * I\{A = 6000\}$)	-61.55 (144)	-59.49 (2418)
Log-likelihood	-9609.3	-8697.8