Essays on Regulations in Peer-To-Peer Markets

Agraj Gupta

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Patrick L Bajari, Chair

Rajib K Doogar

Gregory Duncan

Yanqin Fan

Yuya Takahashi

Program Authorized to Offer Degree:
Economics

University of Washington

Abstract

Essays on Regulations in Peer-To-Peer Markets

Agraj Gupta

Chair of the Supervisory Committee:
Professor Patrick L Bajari
Department of Economics

Peer-to-Peer (P2P) markets have shown tremendous growth in the post-recession era, and three reasons contribute to their success. First, the economic surplus is generated for both buyers and sellers from the trade of a large selection of traditional products and services. Second, there exists easy instant access to the market through modern devices and internet technologies. Third, these markets self-correct failures by removing information asymmetry through reputation systems. Despite increasing efficiency in the markets of traditional goods and services, many P2P businesses and their markets face scourge from lawmakers because of two reasons. First, the P2P markets enable many users to operate where their operation conflicts with the existing regulatory regime. On the other hand, lobbying efforts by the incumbents in the industry might influence the regulators.

Chapter 1 examines how a Peer-to-Peer market such as Airbnb - a short-term rental marketplace - responds to supply restrictions prompted by a series of regulatory threats and incumbent pressure. The growth of online P2P marketplaces, their potential to influence traditional industries, and their operational dissent with outdated (yet existing) laws have made these markets a subject of substantive policy debate. In anticipation of regulation, Airbnb had purged thousands of rentals from its New York City market over the past few years. I use unique daily data of over 100 thousand rentals that appeared on the Airbnb website between September 2014 and March 2017. I identify changes in price and quantity of

accommodation due to the purge by exploiting the patterns of rental loss from the platform observed in the data. Next, I estimate the residual demand faced by each rental, and under assumed monopolistic competition market structure, I recover the marginal costs. These estimates are used to compute the welfare of producers (hosts/rental owners) and consumers (guest/travelers), which I find out to be $392M$ USD and $566M$ USD respectively in a 31 month period. Using a counterfactual setting where purge was not carried out, I recompute the welfare and find that rental owners lost nearly $26M$ USD while consumers lost $25M$ USD as a result of the purge.

Chapter 2 revisits the residual demand estimation problem in the presence of high dimensional confounding demand signals. I refrain from subjectively selecting the variables entering the model and automate this process by using Machine Learning (ML) techniques. I apply the Double Machine Learning (DML) approach proposed by Chernozhukov et al. (2018) and Chernozhukov et al. (2017), to a massive data of Airbnb rentals in the New York City (NYC) and estimate the demand facing each rental. The paper exploits the intrinsically high-dimensional text data from rental amenities, descriptions, and reviews, which confound the true demand relationship producing biased estimates. In converting text data into meaningful features and constructing a computationally manageable data, I make several design choices. The findings in this paper show a huge improvement on the previous estimates of demand, which completely ignored the time-varying reviews as controls. The demand estimate jumped to -14.364 with the standard error of 1.283, when we included a host of features constructed from reviews. The second stage results are stable across different choices of Penalized Regression methods. The findings also indicate that a good model fit in the first stage is crucial to the validity of the final estimates. Finally, the chapter shows how biases from ignoring high dimensional features in estimating residual demand, transmit themselves to the supply side and in the welfare calculation.

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would begin by expressing my deepest gratitude to my primary adviser Prof. Patrick Bajari without whom this dissertation would not have been possible. I sincerely thank my committee members Prof. Gregory Duncan, Prof. Yanqin Fan, and Prof. Yuya Takahashi, for their constructive guidance in shaping this research. A special thanks to Prof. Rajib Doogar for his continuous support at both academic and non-academic levels.

I will always be grateful to Prof. Praveen Kulshrestha (my undergraduate university professor) and Dr. Devesh Roy (my colleague/friend/mentor) for their unwavering support to this career choice. I am also grateful to Prof. Fahad Khalil (then Graduate Director of UW Economics) for putting his trust in me and offering me admission to the PhD program.

I thank my father, my mother, and my grandmother, for their solid backing, and unparalleled support throughout the PhD. A big thanks to my younger brother Anuj for providing me valuable technical advice and practical suggestions on core computations in this research. I cannot go without thanking my wife-to-be Raj for her immense patience with me.

Last but not least, I am thankful to my friends and colleagues for always boosting my morale and especially those in Seattle who cooked me several dinners throughout my PhD.

DEDICATION

To my grandfather,
(Late) Shri Shankar Prashad Gupta.

Chapter 1

REGULATION IN PEER-TO-PEER MARKETS: THE CASE OF AIRBNB

1.1 Introduction

This paper examines how a Peer-to-Peer(P2P) market responds to supply restrictions prompted by a series of regulatory threats and incumbent pressure. The growth of online P2P marketplaces, their potential to influence the traditional industries and their operational dissent with outdated (yet existing) laws have made these markets a subject of substantive policy debate. Airbnb, a market for short-term rentals has attracted scrutiny from regulators, affording housing advocates and hotel owners, demanding supply restriction on Airbnb in the New York City(NYC) market over externalities such as safety oversight, failure to collect taxes, and housing crisis among others [1][2]. On the other hand, Airbnb has emerged as a significant competitor of a concentrated (and regulated) hotel industry by enabling fragmented sellers to compete in the lodging market. In building a decentralized competitive market, Airbnb has successfully managed to keep search, matching, and transaction costs low and prevent market failures by keeping the marketplace safe. Thus, in the spirit of Stigler (1971) and Peltzman (1976), limiting competition by imposing barriers to entry, innovation, and entrepreneurship can also be viewed as an outcome of regulatory capture by an organized incumbent like the hotel industry.

Although regulation in the context of P2P businesses has received considerable attention in law and economics circles (Kaplan and Nadler (2015), Cohen and Sundararajan (2015),

[1] https://www.wsj.com/articles/for-uber-and-airbnb-new-york-city-turns-foe-1533843330?mod=searchresults&page=3&pos=17

[2] https://www.investopedia.com/articles/investing/112414/airbnb-brings-sharing-economy-hotels.asp

Koopman, Mitchell, and Thierer (2014)), empirical evidence is quite limited primarily because of the lack of proprietary data on Airbnb users. This paper takes up the case of Airbnb in NYC to measure the welfare effects of forcing fragmented sellers out of the market. Policy measures of similar nature such as mandatory registration requirements, quotas, caps on the operation, and bans have been enacted or being considered by lawmakers to limit the supply of short term rentals in several US markets and abroad. Supply restrictions through registration, permission requirements, or limiting the number of days of hosting poses barriers to entry for many potential hosts and pushes the existing hosts out of the short term rental market. While Airbnb is the most popular platform for short-term rentals, it is certainly not the only P2P marketplace facing regulatory and incumbent pressure. Other P2P markets such as VRBO, Homestay, etc. among short-term and vacation rentals, Uber and Lyft among ride-sharing platforms also face similar regulatory restriction. In this paper, I measure the impact of forcing rentals out of Airbnb market on the welfare of consumers (guests/tourists/travelers) and producers (peer sellers/local hosts).

In particular, this paper exploits a discrete set of events when in anticipation of regulation, Airbnb has been purging allegedly "illegal" rentals from its NYC market as a part of its "One Host, One Home" campaign[3] to measure the impact on welfare or consumers and sellers. The effect of Airbnb's self-inflicted supply restriction will enhance our understanding of competition within P2P markets and will serve as a yardstick in policy debates concerning supply restrictions in such markets. If the supply regulations are enacted with a public interest in mind to overcome negative externalities identified by regulators, then the welfare changes can be perceived as the cost of correcting market failures. On the other hand, if supply restrictions are a result of regulatory capture by traditional hotels who have high stakes associated with such supply regulation on Airbnb, then the impact on welfare can be viewed as a burden of government failure.

In order to measure the welfare effects of the purge prompted by regulatory threats, the

[3] https://www.airbnbcitizen.com/one-host-one-home-new-york-city/

paper proceeds in the following steps: First, I estimate the residual demand function faced by each rental in a highly differentiated market, taking the supply of all other rentals as given and calculate consumer surplus. Assuming monopolistic competition market structure and using the demand estimates and I recover marginal costs, markups, and producers profit. Next, I use a discrete set of spatial and temporal events of the rental purge to estimate changes in equilibrium prices and quantity of Airbnb accommodation. Finally, I recompute welfare in the counterfactual scenario without purge using the changes in price and quantity of the rentals.

I estimate a linear-log specification of residual demand, assuming that the shape of the demand curve stays the same for all rentals. This specification allows price sensitivity to decrease with prices. The estimation of demand suffers from endogeneity problem as I am restricted to the equilibrium data. I use IVs constructed using the rentals decision to operate in the past (previous markets). I argue that a rental's past operation decisions capture information which is correlated with the rental costs and therefore prices on a given day. Conditional on other factors, these IVs vary independently of the current demand. The estimated price sensitivity parameter is -6.79. The total accrued surplus of the consumers is $392M\$$. An average rental on a given day contributes about 13\$ to consumer welfare, while a purged rental used to contribute about 28\$ when operational.

Using the demand side parameters and observed prices, I recover marginal costs using Lerner's formula. I find evidence that sellers find it hard to charge higher markups during the periods of high demand as more number of flexible sellers enter the market. Low-cost rentals operate during low demand season while high costs rental find it easier to rent during peak seasons. The resulting seller's profits are $566M\$$, where an average rental earns about 19\$ per day while purge rentals earned about 37.7\$ each day.

Next, to estimate the changes in prices and quantity of Airbnb accommodation, I model equilibrium prices and quantity of rental as a function lost capacity and distance from a purged rental. The identifying assumption is that a rental remains unaffected a purged rental if the latter is located 10 miles away from the former. I use different proximity buffer regions

around each rental to identify the impact on equilibrium price and quantity. I find a distinct increase in the price and decrease in the quantity of accommodation. This effect size becomes smaller as the distance between existing rental and the purged rental increases.

Finally, I recalibrate a counterfactual scenario without the purge and calculate consumer and producer surplus. In a 17 month duration, I find that consumers are hurt not only because of the price increase due to reduced competition ($2.3M\$$) but also from the loss of selection ($22.8M\$$). On the other hand, the owners of the purged rentals lose sales after being eliminated from the market ($9.5M\$$). However, the rentals that were not purged gained about ($34.9M\$$) from less competition.

Background : Airbnb currently holds more than 3 million accommodation units that have catered to 200 million guests in 65 thousand cities spread across 195 countries [4]. The popularity of Airbnb stems from the benefits it provides its users in three primary ways. First, it generates value for visitors by providing them access to cheaper and personalized alternatives to stay compared to hotels in a city. Second, it creates a surplus for residents of the city by enabling them to earn money from their partially or entirely unused livable space. Third, Airbnb keeps the market safe for all participating users and has a well-established quality control system through reviews.

The rapid growth of Airbnb suggests the huge benefits its consumers (the guests) receive in terms of price, quantity, quality, and variety. Participation of numerous residents who supply accommodation spaces on Airbnb drive competition within the platform which benefits the consumers. In contrast to the traditional hotel industry, the Airbnb marketplace provides lower entry cost for its suppliers (the peer hosts) into the lodging market in a competitive manner. When hotels and motels respond to the demand surge with high prices, Airbnb hosts respond with increased supply and therefore undercutting hotels. Besides prices, Airbnb rentals compete with hotels and motels at different levels of quality and variety (see Zervas, Proserpio, and Byers (2017)). Although Airbnb has injected healthy competition into the

[4] https://www.airbnb.com/about/about-us

lodging industry, its rapid growth has sparked several interest groups to demand regulation that restricts entry into the platform. Hotels, affordable housing advocates, neighborhood associations and other anti-Airbnb activists have accused the company of promoting "illegal hotels", compromising the safety, encouraging racial discrimination, reducing affordable local housing stock, disrupting residential buildings and neighborhoods. While often denying any intimidation of competition from Airbnb, the hotel industry has to lead a focused lobbying campaign along with other interest groups to persuade regulators to restrict Airbnb's activities in the city. Many Airbnb markets around the world (not just in the US) have been struggling with regulators who have responded favorably to intense lobbying by considering to regulate Airbnb. The restriction already imposed or considered are mandatory registration/ permission requirements, quotas on the number of rentals, caps on the number of days of operation, and bans.

NYC has witnessed the persistent struggle between Airbnb on one side and regulators, a hotel industry lobbyist, affordable housing advocates, and neighborhood association on the other. Since October 2014, pressure mounted on Airbnb when New York State Attorney General used Airbnb data to claim that Airbnb is promoting unsafe "illegal hotels" that distort city neighborhoods and make housing less affordable. In New York, according to "Multi-Dwelling Law", it is illegal to rent an apartment in a building with three or more units for less than 30 days unless the owner is physically present in the apartment. This law brings a majority of rentals under scrutiny, including those which are rented occasionally and do not cause the housing crisis (see Barron, Kung, and Proserpio (2018)). Fearing growing Anti-Airbnb agitation and perceived regulation in NYC, Airbnb began purging its rentals that were operated by "bad actors". These "bad actor" or "commercial operators" rent multiple apartments on Airbnb across NYC. Not only are these rentals are believed to violate "Multi-Dwelling Law" but also exasperate the housing crisis in the city.

Related Literature : This paper builds on the growing empirical research on online P2P marketplaces. Proserpio and Tellis (2017) provide an extensive overview a rapidly expanding research on P2P markets. The existing literature examines many (but related)

facets of these markets such as entry and competition with hotels (Zervas, Proserpio, and Byers (2017), Farronato and Fradkin (2018)), changing patterns of ownership and usage of resources (Fraiberger, Sundararajan, et al. (2015), Horton and Zeckhauser (2016), Gong, Greenwood, and Song (2017)), pricing mechanisms and elastic market supply (Cullen and Farronato (2014), Chen and Sheldon (2016), Hall, Horton, and Knoepfle (2018)), reputation systems (Zervas, Proserpio, and Byers (2015), Fradkin, Grewal, and Holtz (2018), Proserpio, Xu, and Zervas (2018)) and platform design (Fradkin (2015), Cramer and Krueger (2016), Hall, Kendrick, and Nosko (2015)) among others.

The primary contribution of this paper lies in quantifying the welfare changes of sellers and consumers because of a potential regulatory restriction. A fraction of literature studies the effect Airbnb on hotels such as Guttentag and Smith (2017), Zervas, Proserpio, and Byers (2017), and Farronato and Fradkin (2018). Zervas, Proserpio, and Byers (2017) studies the effect of Airbnb on the incumbent hotel industry in Texas and show that competition pushes prices downwards, which benefits consumers. Farronato and Fradkin (2018) derive similar results and quantify surplus for a broader range of US markets as Airbnb competes with traditional hotels with constrained supply. The current paper expands these ideas to understand the nature of competition among the peer sellers within an Airbnb market. In particular, this paper looks at the price-quantity response of peer sellers when other competing sellers are forced out of the market due to an intervention.

The empirical findings on the issue of regulation and Airbnb are quite limited. Filippas and Horton (2017) model and test a regulatory hypothesis which allows the building owners to decide whether to permit subletting in the building. They conclude that doing so does not affect the rental prices, which is equivalent to the social planner outcome. Quattrone et al. (2016) provide correlational relationships between Airbnb's spread in London on socio-economic factors and suggest a data-driven algorithmic approach to regulation that considers the spatial and temporal evolution of such markets. In the context of regulation, Edelman and Geradin (2015), Kaplan and Nadler (2015), Cohen and Sundararajan (2015), Koopman, Mitchell, and Thierer (2014), Rauch and Schleicher (2015) and Rogers (2015)

provide theoretical and conceptual perspective on externalities accompanying these P2P markets. Cohen and Sundararajan (2015) stresses on self-regulation of these platforms with limited government interventions. Koopman, Mitchell, and Thierer (2014) suggests a possibility of regulatory capture by incumbents and argue against arbitrary legal restrictions on these platforms. It has a significant implication in the NYC as most of the "illegal" rentals in the city are due to the existence of "Multiple Dwelling Law"[5] of 2010. This paper provides complementary evidence to the debate surrounding regulations on P2P markets.

Another stream of papers studies the externalities of Airbnb operation in a city, particularly those that impact housing markets. Barron, Kung, and Proserpio (2018) provides empirical evidence that the growth of Airbnb raises the rents (and therefore house prices) since Airbnb incentivizes owners to switch from the long-term rental market to short-term. Similar evidence of impact of Airbnb on housing appear in Horn and Merante (2017) for Boston, Lee (2016) for Los Angeles , and Sheppard, Udell, et al. (2016) for NYC. Airbnb's influence on the housing market has been the most convincing political argument used by lawmakers for restricting Airbnb in NYC. The profitability computed in this paper can capture the extent of such incentives that push rentals from long-term to short-term markets. Though Sheppard, Udell, et al. (2016) arrive at the same results in terms of rents and house prices, they are careful enough to point to the policymakers that ignoring the efficiency with which homeowners utilize their space while drafting laws to make housing affordable may not be "welfare improving" necessarily.

Some other papers view Airbnb from a social perspective which is often used as a justification for regulation. For example, Edelman and Luca (2014) provides evidence of racial discrimination on Airbnb.

The rest of the paper is organized in the following manner: Section 1.2 describes the data and the NYC market of Airbnb. I first look at features and outcomes at the aggregate level across time/market and then at the rental level. Next, I examine the conduct of the purged

[5] https://www1.nyc.gov/assets/buildings/pdf/MultipleDwellingLaw.pdf

rentals and compare them with all rentals in the market. Section 1.3 presents an empirical framework to analyze the Airbnb market. In this section, I specify a demand function, layout IV strategy to estimate the price coefficient and describe an empirical strategy to identify changes in price and quantity as purged rentals leave the market. Estimation results, surplus calculations, and counterfactual analysis are described in section 1.4. Finally, section 1.5 discusses the findings and concludes the paper.

1.2 Market Description and Data

The daily data on rentals of the NYC market has been obtained from Airdna[6], a private data company that frequently scrapes and tracks Airbnb rentals across several cities around the world. The data provides information on prices[7], daily availability and booking status[8]. This data on price, quantity, and the decision to operate in the market is augmented by information on rental attributes such as space type[9], property type[10], number of guests a rental can accommodate, bedrooms, bathrooms, check-in time, check-out time, neighborhood,

[6] https://www.airdna.co/

[7] These prices (per night) appear on Airbnb website and may not reflect the exact transaction price because of the possibility of discounts and negotiation that happens after potential guest contact the host of a rental.

[8] In the data, a rental on a given day can have one of three statuses - Blocked, Available and Reserved (Booked). A rental is "Blocked" by the host when he chooses not to advertise his rental on Airbnb. In this paper, I consider "Blocked" status as a rental's decision to not to operate on a given day/market. The status "Available" implies that rental had been on the market but was not "booked" by anyone.

[9] There are three types of spaces offered by rentals on Airbnb. "Entire home and apartment"(EHA), "Private room"(PR) and "Shared room"(SR) which differ in terms of size, privacy and freedom a guest can avail during the stay. EHA is rental space where a guest has an entire living space to himself during the stay and can enjoy maximum space and privacy. PR, on the other hand, provides limited space, privacy, and freedom limited to a private bedroom but rest of the living space is generally shared either with the host or other Airbnb guests in the entire property space. SR is an arrangement with minimal privacy, and a guest is required to share living space as well as bedroom space. A host may choose whether to market their rental space as an entire home, private room or a shared room since space types are inter-convertible amongst themselves and are constrained only by the total space within the property. For instance, an EHA space can be broken and advertised as multiple private rooms, given the number of rooms in the property.

[10] The property type can be on the following : Apartment, Bed & Breakfast, Boat, Boutique hotel, Bungalow, Cabin, Camper/RV, Castle, Cave, Chalet, Condominium, Dorm, Earth House, Entire Floor, Guest suite, Guesthouse, Hostel, House, Hut, In-law, Island, Lighthouse, Loft, Plane, Serviced apartment, Tent, Townhouse, Train, Tree-house, Vacation home, Villa, Yurt, etc..

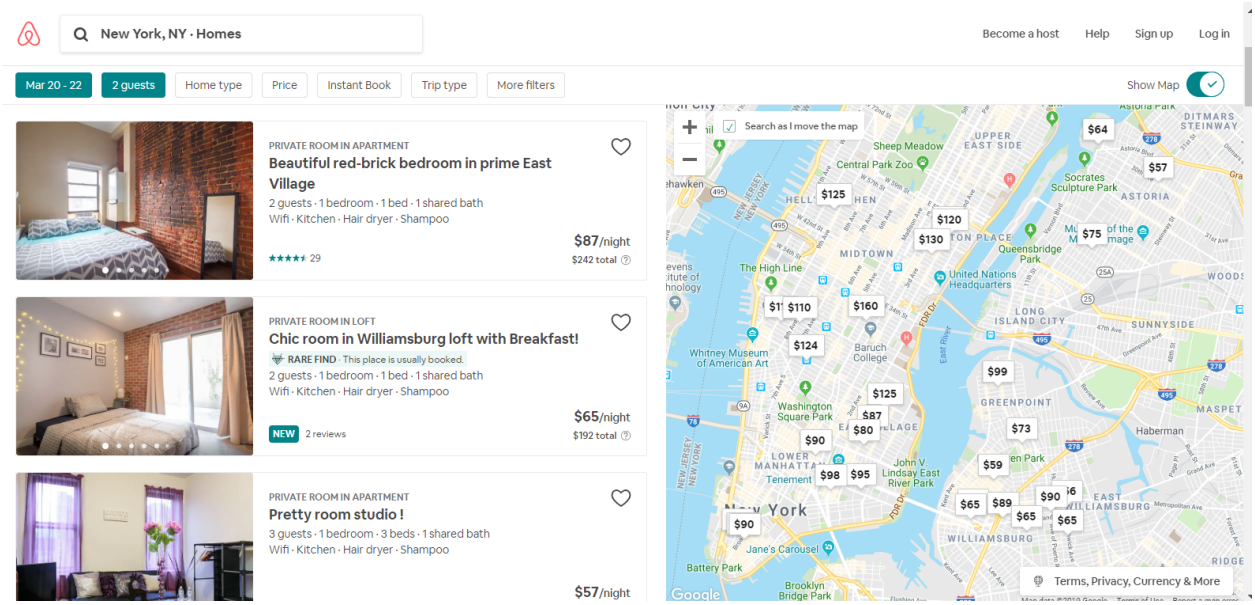


Figure 1.1 – Search Page on Airbnb Website

location in latitude and longitude, among other information. The data spans for over 10 quarters (943 days) from September 2014 to March 2017 and captures all of the 78, 362 hosts and 108, 212 rentals that appeared on Airbnb website during this period. I consider each day as a separate market where rental capacity is traded between hosts and guests.

1.2.1 NYC market

The NYC market comprises of a variety of short-term rental spaces hosted by their suppliers (the hosts[11]) who list one or more rentals on Airbnb's website. Each rental with distinct spaces, amenities, quality, location, etc. offers a capacity of accommodation (measured by the number of persons). The existing rental inventory (capacity) available in the market (on a given day/days) is viewed, examined and then booked online through the platform[12]

[11]These hosts can be legal owners of the property themselves. They may also be renters who primarily live in the property but occasionally sublet their space on Airbnb. Some reports suggest that hosts may also be property managers who rent the properties they manage on Airbnb and otherwise

[12]For every transaction between a consumer and the host Airbnb receives a share of the transaction amount from both sided. Roughly 6% to 12% from the guest depending on the number of days booked and 3%

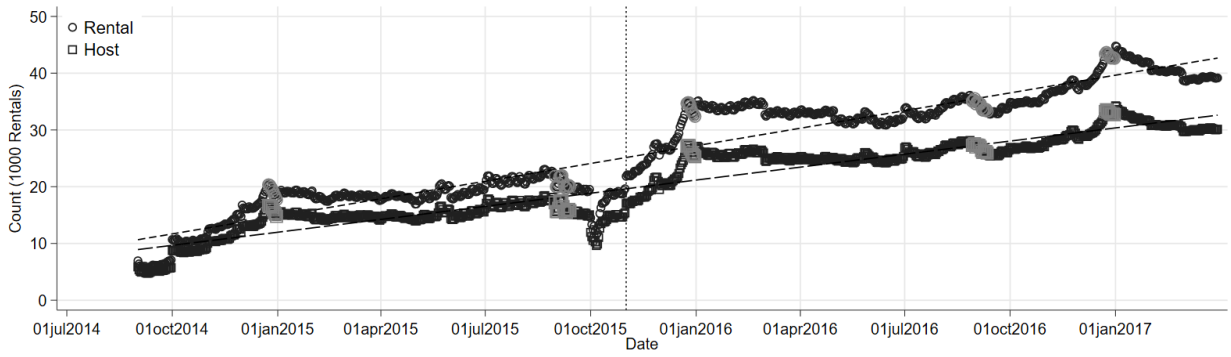


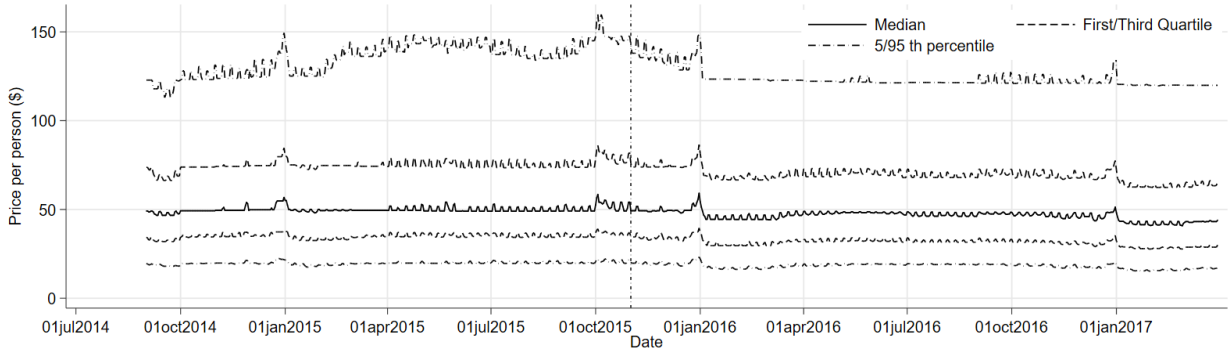
FIGURE 1.2 – NUMBER OF HOSTS AND RENTALS
Notes: The sudden drop at the beginning of October 2015 is potentially due to error in data collection through scraping caused by changes made by Airbnb to the rental listing page on its website. The gray points denote US Open Tennis Grand Slam period and the last week of the year.

(Figure 1.1) by the consumers (the guests). Over the years, the number of rentals, hosts and the bookings have been rapidly increasing (see figure 1.2) and so has the pressure from local regulators and lawmakers, backed by several interest groups, demanding a restriction on Airbnb's supply within cities.

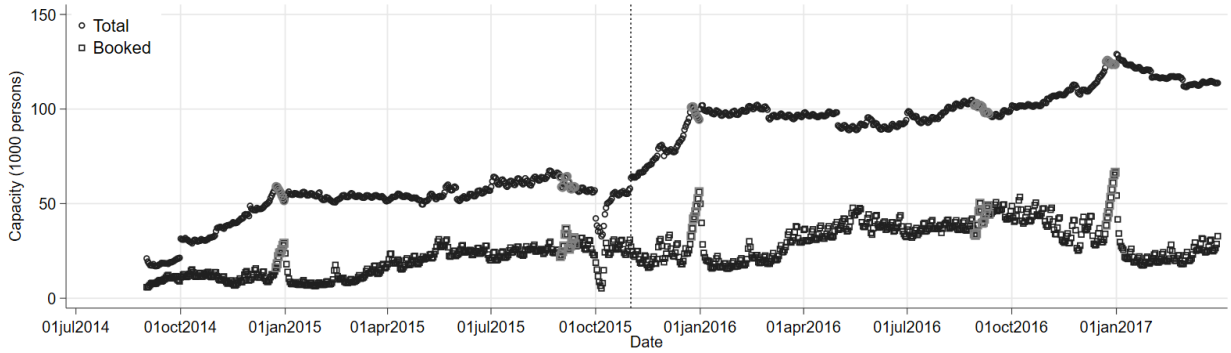
A crucial feature of Airbnb market (and other P2P markets as well) is its ability to change market supply elastically. Airbnb allows almost free entry(/exit) of short-term rentals into the market since there are minimal upfront costs to hosting on Airbnb (see Einav, Farronato, and Levin (2016), Farronato and Fradkin (2018)). Figure 1.3 shows movements in price and quantity across markets with elastic supply. Prices increase on the weekends and other times when the demand is high particularly around the holiday season and the new years. Any change in prices occurs simultaneously with changes in accommodation supply. While overall capacity booked has been gradually increasing over time with growing popularity of Airbnb, price per person has been decreasing. This is particularly because of the flexible supply enables competition within the market, keeping the prices low.

The NYC market on any given day is characterized by a large number of sellers (hosts) who advertise one or multiple rentals on Airbnb. Figure 1.2 shows the number of active hosts

from the host



(A) Price

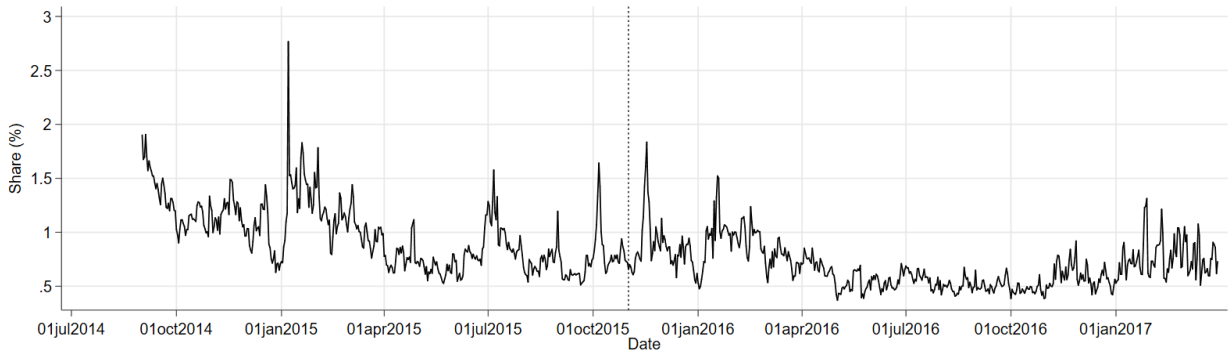


(B) Quantity Supplied and Rented

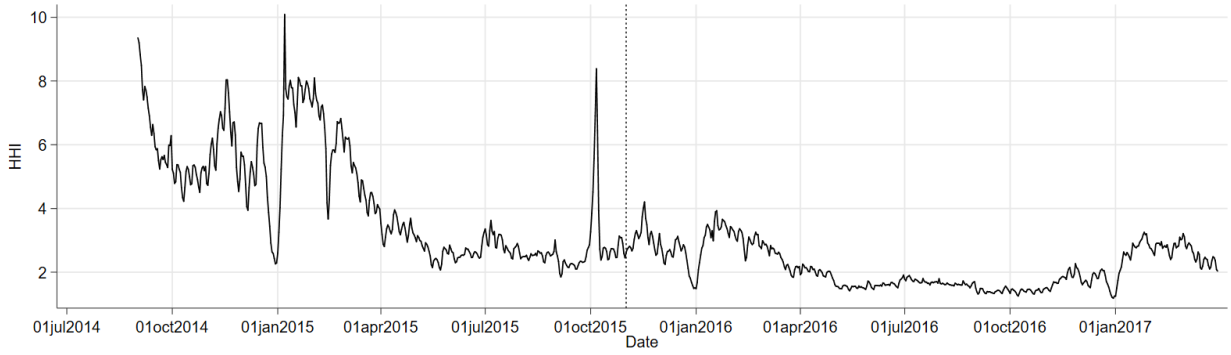
FIGURE 1.3 – MARKET PRICE AND QUANTITY

and rentals on a given day/market. Though many hosts operate more than one rentals[13], market on any given day is highly fragmented where top 4 hosts contribute to less than 3% on any given day (see Figure 1.4a). In fact, on average across all the days, the revenue share by top 4 firms is a little less than 1%. The Herfindahl Index also indicates that this market in extremely un-concentrated (Figure 1.4b). Interestingly, the concentration drops drastically during the high demand periods, for example, the holiday season, as shown in figure 1.4. This is because flexible peer sellers respond to rising prices during periods of high demand by supplying more accommodation.

[13]A host may operate more than one rentals, either contemporaneously or over time. Airbnb provides its hosts the choice of when to list their space on Airbnb without any transaction cost.



(A) Revenue Share Of Top 4 hosts



(B) Herfindahl Index

FIGURE 1.4 – MARKET CONCENTRATION
Notes: The revenue of the top 4 firms and the Herfindahl Index have been calculated considering each day as a separate market.

The search for rentals on Airbnb is based on consumer preference for prices, attributes, rating and reviews of a rental. Search results and recommendations presented to a potential guest on the website may have little to no dependence on a specific host. This is because, as a platform, Airbnb's objective is to increase the transactions since its profits are tied to each transactions[14]. Besides, rentals on a given day are booked at different times. Given these circumstances, I argue that the rentals managed by the same host do not coordinate and compete independently on prices with all other rentals.

[14]Airbnb collects 3% from the host for each transaction amount and $6\% - 12\%$ from the guest

1.2.2 Rental Information

The NYC market offers a breadth of differentiated rentals to travelers visiting the city. These rentals differ on several dimensions such as space type, property type, location, number of bedrooms and bathrooms, check-in time, check-out time and amenities. Besides, consumers differentiate rentals through pictures and written description posted by the owner on the online rental listing. Trust and reputation system which includes reviews and feedback additionally creates differences in consumer perception of rentals. Table 1.1 shows the number of rentals observed in the data and their distribution across a few attributes. 50% rentals offer "Entire Home/Apartment" space. Owners of Apartments, Condos, Lost and Townhouse supply about 92% or rentals. Notice that, according to the "Multi-Dwelling Law" rentals at the intersection of "Entire Home/Apartment" space-type and Apartments, Condos, Lost and Townhouse property-type are mostly "illegal".

Table 1.2 present basic summary statistic of rental operation activity in NYC from September 2014 to March 2017. The rentals decide to enter/exit the market at different times depending upon the market conditions. Out of a total of 943 days, on average rentals operate for 233 days. Average rental in the market has a capacity of 2.7 persons and receives a booking 29% of the times only.

Rentals supply behavior also varies across space types, property type and locations. The "Entire Home/Apartment" rentals operate for a shorter duration (226 days) but in a larger capacity (3.6 persons) compared to other two categories. Though "Entire Homes/Apartment" rentals operate for a smaller duration, they capture more bookings. Occupancy rates indicate that some rentals are preferred more than the others. In the space-type classification, the "Entire Home/Apartment" have a higher occupancy (32.32%) than rentals that offer "Private Room" (26.71%) or "Shared Room" (21.93%) spaces.

These operational differences exist across property types also. Boats, cabins, camper, RV and other similar property types that offer confined spaces and remain an idle resource for most of the times in a year stay on Airbnb longer (328 days) than residential apartments

Table 1.1 – Rentals

	(1)	(2)
	Number of Rentals	
	All Rentals	Purged Rentals
All rentals	108,212	4,127
Space-Type		
Entire Home/Apartment	55,790	4,127
Private Room	48,183	
Shared Room	4,207	
Property-Type		
Apartment/Condominium/Loft/Townhouse	99,782	4,122
House	7,025	
Bungalow/Castle/Guesthouse/Vacation home/Villa	145	
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	89	
Location		
Bronx	1,570	32
Brooklyn	39,527	867
Manhattan	56,977	2,997
Queens	9,396	212
All Staten Island	733	19

Notes: These are all the rentals that appeared on Airbnb's website between September 2014 and March 2017. The Purged rentals are the rentals identified to be removed on the way after November 2015.

(227 days). Despite, higher market presence, such property types have smaller occupancy (23%). These patterns suggest that the rental spaces with high opportunity cost do not stay in the market as long as those with low opportunity cost. From the location perspective, rentals in Manhattan are advertised less number of days (225) on average. However, the mean rental occupancy in Manhattan (30.66%) is higher than other boroughs of NYC. These evidence suggest that rentals are not only heterogeneous in terms of their characteristics but also their supply and operational behavior. These differences in operation also suggest that the distribution of rentals may vary across time/market.

Next, we look at the variation in prices, average revenue generated by a rental on a given day and total revenue in 31 months. Average per person price in the data is 57.82$. The prices charged for "Entire Home/Apartment" rentals (64.55) is more that "Private Room" (51.21$) or "Shared Room" (43.46$) rentals. Though, the averages calculated do not account

Table 1.2 – Operation and Occupancy Across Rentals

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Operation (Days)				Operation (Capacity)				Occupancy (%)			
	All Rentals		Purged Rentals		All Rentals		Purged Rentals		All Rentals		Purged Rentals	
VARIABLES	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
All rentals	232.5	217.8	322.5	186.3	2.791	1.799	4.362	2.356	29.41	29.41	30.81	26.80
Space-Type												
Entire Home/Apartment	226.6	214.8	322.5	186.3	3.592	1.962	4.362	2.356	32.32	29.76	30.81	26.80
Private Room	237.8	222.7			1.922	0.914			26.71	28.97		
Shared Room	250.5	197.5			2.137	2.241			21.93	25.12		
Property-Type												
Apartment/Condominium/Loft/Townhouse	227.2	215.0	322.6	186.3	2.747	1.674	4.354	2.336	29.62	29.52	30.84	26.80
House	297.7	239.9			3.325	2.722			27.53	28.10		
Bungalow/Castle/Guesthouse/Vacation home/Villa	219.8	206.6			2.945	2.581			28.78	28.25		
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	328.8	205.8			2.831	2.399			23.07	27.13		
Location												
Bronx	282.9	221.6	260.7	168.0	2.732	1.853	5	2.064	22.49	25.97	28.69	27.60
Brooklyn	235.6	221.0	325.2	188.9	2.748	1.858	4.796	2.636	28.83	29.35	27.91	25.66
Manhattan	225.6	214.9	322.0	185.4	2.835	1.730	4.224	2.265	30.66	29.71	32.25	27.09
Queens	248.1	216.2	320.0	180.9	2.688	1.891	4.486	2.251	25.77	27.96	22.99	25.11
All Staten Island	289.4	243.6	413.5	255.5	3.201	2.278	4.053	2.041	25.04	26.99	26.47	24.31

Notes: Each variable has been first calculated at rental level and reported basic statistics of all rentals. The total number of days in the data were 943. The occupancy has been calculated by dividing the number of days booked by the number of days it existed on the market.

for seasonality, table 1.3 suggests that there is significant variation in prices across rentals. It is also evident that some rentals earn more revenue than others. While higher prices contribute to the rentals revenue, we find that occupancy and capacity generated by rentals too have a large role to play in seller's revenue. As a result, the revenue across rental varies more than prices. For example, "Entire Home/Apartment" average revenue (57.79$) is about three times more than "Private Room" rentals and more than five times that of a "Shared Room" rental. Overall, the revenue generated in the NYC market in 31 months is $1.1B$$. Most of the revenue is generated by the "Entire Home/Apartment" rentals ($831.1M$$) most of which are subject to regulatory scrutiny.

1.2.3 Response to Anticipated Regulation

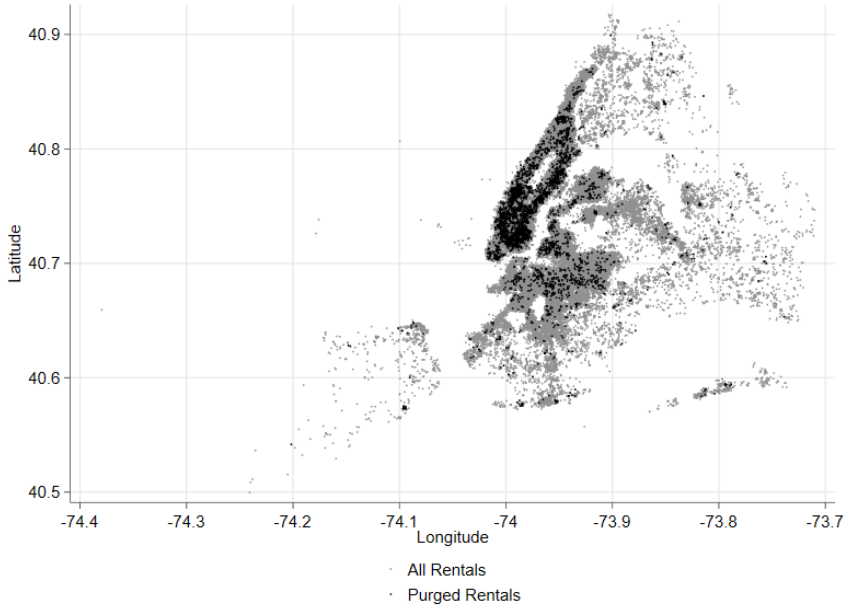
Anticipating a regulation, Airbnb began removing rentals from its website since November 2015 as a part of their "One Host, One Home" campaign. The company had removed over 4000 rentals from its NYC market by March 2017 and continues to do so. These rentals were "Entire Home/Apartment" operated by "commercial" hosts. The data on purged rentals was not exclusively available and therefore one of the tasks was to identify the rentals that were purged by Airbnb. To identify the purged properties, I use information provided by Airbnb press release and patterns in the data on rentals that were purged. Specifically, the rentals that were removed had two distinct features as reported in the press release. First, they are "Entire home/Apartment" rentals. Second, they are operated by hosts who manage more than one "Entire home/Apartment" rental.

The rentals have been purged from different geographic location and at different times. Figure 1.5a shows spatial distribution of all the purged rentals. A majority of rentals were removed from Manhattan borough. These rental were also evicted form the market at different points in time (Panel 1.5b). Despite of being prematurely evicted from the market, the purged rentals on average operated for a longer duration (323 days). The occupancy of the purged rentals (30.81%) too was better than the average rental on Airbnb and they also operated at higher capacity (4.36). Though purged rentals were priced more that average rental,

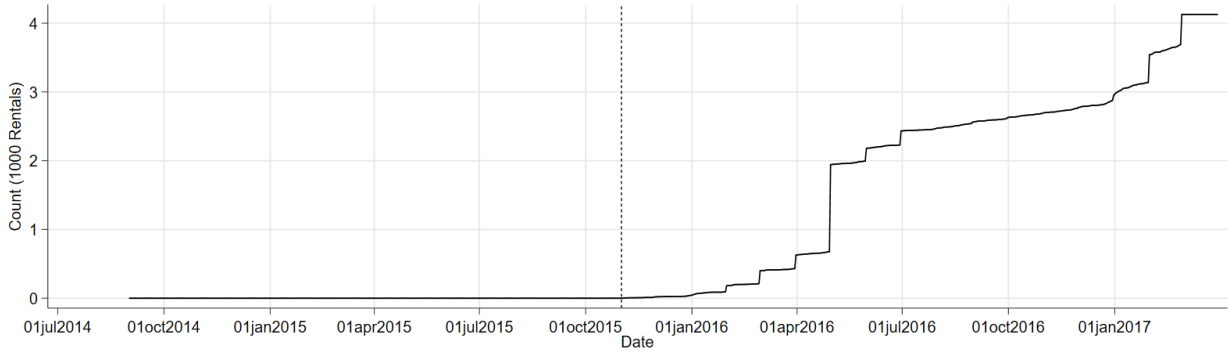
Table 1.3 – Price and Revenue Across Rentals

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Price ($)				Daily Revenue ($)				Total Revenue ($1M$)	
	All Rentals		Purged Rentals		All Rentals		Purged Rentals		All Rentals	Purged Rentals
VARIABLES	mean	sd	mean	sd	mean	sd	mean	sd	sum	sum
All rentals	57.82	59.36	61.94	51.79	39.37	54.20	64.00	68.10	1,100	90.89
Space-Type										
Entire Home/Apartment	64.55	57.23	61.94	51.79	57.79	66.25	64.00	68.10	831.1	90.89
Private Room	51.21	60.69			20.48	25.85			256.5	
Shared Room	43.46	57.91			11.62	18.77			11.99	
Property-Type										
Apartment/Condominium/Loft/Townhouse	58.97	57.21	62.00	51.79	40.22	54.20	64.07	68.10	1,020	90.89
House	39.48	45.54			29.75	54.60			68.77	
Bungalow/Castle/Guesthouse/Vacation home/Villa	49.52	63.34			26.31	40.22			0.864	
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	59.50	51.77			17.74	20.69			0.554	
Location										
Bronx	36.41	36.52	26.03	9.636	14.65	19.54	29.88	31.74	6.646	0.218
Brooklyn	46.08	44.99	45.25	34.24	30.55	42.14	47.26	50.55	310.6	13.81
Manhattan	69.28	62.03	68.72	55.77	49.12	62.69	71.67	72.78	719.9	74.11
Queens	41.75	81.45	39.05	27.07	21.88	29.79	29.75	33.29	54.53	2.221
All Staten Island	51.37	44.42	69.15	48.72	33.68	58.93	58.07	67.21	7.919	0.537

Notes: The prices and daily revenue has been first averaged within each rental and the summary statistic have been calculated across rentals. The total revenue has been computed by aggregating across all rental-time pair.



(A) Geographic Location of Rentals



(B) Purged Rentals Across Time

FIGURE 1.5 – SPATIAL AND TEMPORAL DISTRIBUTION OF PURGED RENTALS

they were slightly cheaper than their "Entire Home/Apartment" counterparts. Still, the revenue of purged rentals is higher than that of their "Entire Home/Apartment" counterparts mainly because of operating at higher capacity. Though these rentals were purged, they have contributed a little less than 1% of the total revenue and therefore reflect a small but significant portion of Airbnb's supply side. Given the nature of the NYC marketplace (many

fragmented sellers, differentiated products and free entry and exit), I assume a monopolistic competition structure in each market separated by time. Consequently, in each market/time, a rental sets its prices against the residual demand they face. The possibility of collusion on pricing is rare since the market is highly diluted and consumer search is likely to be independent of a specific host. The rentals decide to operate and rent as long as their profits are greater than zero. Also, in such a market, rentals will continue to enter as long as there are economic profits.

1.3 Empirical Strategy

I index my specification by host i, rental j and time (night/market) t. I consider Airbnb as a market for short-term accommodation where each flexible (or peer) seller of rental j produces a capacity (q_{jt}) of accommodation for each night t. Although the booking at a given time t occurs at the rental level, I choose to define q_{jt} as a measure of capacity in terms of number guests a rental can accommodate rather than the overall count of rentals. The reason is that first, it allows us to model buyers and sellers decision at a finer dis-aggregated level and thereby it renders more variability in the data. Second, it enables us to circumvent the computational burden that accompanies with binary outcome regressions in a panel setting with multi-dimensional fixed effects and instrumental variables. Consequently, price per unit of accommodation (p_{jt}) for each rental can be defined as the nightly price per available capacity and is calculated by dividing the nightly price of the entire rental by its capacity of accommodating guests.

In the rest of this section, I describe a framework for empirical estimation and welfare calculations. I begin by specifying the demand faced by each rental in a market characterized by the monopolistic competitive structure. I lay out an instrumental variable (IV) strategy to account for potential price endogeneity as we are confined to equilibrium level data. Then, I discuss the supply side structure and how I recover the underlying marginal cost of Airbnb accommodations. Next, I describe an approach to estimate changes in price and quantity of Airbnb accommodations as a result of the purge by exploiting the spatial and temporal

patterns of removed rentals observed in the data.

1.3.1 Demand

I begin by specifying a residual demand function facing each rental while taking supply response of other rentals in the market as given. Focusing on residual demand allows us to flexibly obtain the elasticity (and therefore, the degree of market power) in a suitable monopolistic competition setup without indulging into estimating a notoriously large number of cross-price elasticities (Baker and Bresnahan (1988)). I refrain from using discrete choice methods of demand estimation because the Airbnb market has evolved over time. Since rentals can enter/ exit costlessly on Airbnb, market boundaries are difficult to define. Moreover, the NYC market for tourist accommodation can geographically extend to neighboring cities in New Jersey and to the hotels which are not the primary focus of this paper.

The quantity demanded (q_{jt}) of rental j in market/time t as:

$$q_{jt} = \alpha_0 + \alpha_1 ln(p_{jt}) + \alpha_2 X_{jt} + \mu_j + \lambda_t + \nu_{jt} + \eta_{jt} + \epsilon_{jt} \tag{1.1}$$

where is the quantity demanded p_{jt} is the price (inflation adjusted) and X_{jt} are the time-varying rental attributes. I include rental fixed effect (μ_j) to capture all the time-invariant features of a rental such as images and pictures. The rental fixed effects control for the heterogeneity across the rentals. The demand also shifts by each day and λ_t captures market level and seasonal variations in the demand. Finally, I use neighborhood-week fixed effects ν_{jt} and neighborhood-day of the week η_{jt} to capture the evolution of rentals in different parts of the city. These fixed effect also account for factors that affect the supply side at the broader neighborhood level. In estimating the residual demand α_1 is the parameter capturing price sensitivity of demand.

The estimation of demand suffers from potential endogeneity issue as I am confined to equilibrium price and quantity data. I account for this endogeneity by employing measures of supply-side decisions of each rental in previous markets/time (similar to Hausman (1996)).

In particular, I argue that past production decision of whether to rent (operate) on Airbnb or not and in what capacity is correlated with rental's inherent costs and therefore prices set by rentals. For example, a rental which operates 7 days in the past week is likely to have a lower opportunity cost than a similar rental operating only 2 days in the past week. Also, rentals operating in a different capacity for the same number of days will have different underlying costs. Thus, if the opportunity costs are high, then a rental is more likely to stay inactive on Airbnb or operate in a smaller capacity.

For exclusion restriction, I claim that the current demand for a rental should not vary with the past decision of making a rental available in the market conditional on recent past bookings and customer reviews. Bookings made on Airbnb by a customer happen on a periodic basis which typically spans over a few days or weeks. So, the past hosting behavior will be correlated with current demand if the customer is looking to book a rental for more than one day. Hosting also depends on reviews which impacts the current demand since older reviews influence both past and current bookings. It is therefore crucial to control for past bookings and reviews. In this paper, I do control for past bookings, but not for past reviews that require extensive Text Mining and Natural Language Processing (NLP).

1.3.2 Supply Side

Next, I turn to the supply side of the market. I assume that mode of conduct on the supply side follows monopolistic competition where a host i decides whether to list rental j on Airbnb in a given market/time t and the price per night p_{jt}[15]. Pricing and the decision to list a rental (the shutdown decision) depends on the profit π_{jt}, given the residual demand faced by the rental. The setup assumes that each host knows the residual demand for his rental in any given market/time. This assumption is reasonable because a host can view all the other rental listing in the market on the Airbnb website (including their detailed information) and the rental owner also remains informed of market demand from past information and general

[15]Each rental can have only one owner. However, one host may have more that one rental.

awareness. Moreover, if in case the owner does not know his own rentals demand early, he still has the freedom to adjust his price accordingly as he learns his own residual demand.

Each host holds private information about the costs associated with their rentals on a given date. The total cost (C_{jt}) incurred by rental j at time (market) t consists of fixed and variable costs denoted by F_{jt} and V_{jt} respectively. The costs, fixed or variable, are fragmented and measured at a daily level to fit in the econometric specification. Since a host can dedicate different fractions of their property to Airbnb rentals, the costs in consideration belong only to portions of the property dedicated to rental. For example, the property tax is paid for the entire property. However, the costs of renting on Airbnb shall include a fraction of property tax paid by for the portion of the property dedicated to Airbnb rental and not the entire property.

In the case of Airbnb rentals, the fixed costs are small as there is no upfront cost of building a certain capacity of accommodation. Airbnb's major role as a market designer is to eliminate advertising costs and therefore the only fixed cost associated with a rental is in the form of time spent on listing the rental and maintaining rental website. The second stream of costs is incurred depending upon the hosts' decision on the number of guests his rental can accommodate. These costs can be thought of as the opportunity cost of the dedicating space and amenities towards the rental. This opportunity cost incorporates alternate use of the space/amenities and can take a variety of forms depending upon rental and the time it is being rented.

To understand these costs better, consider the following example. A host can be a single consultant (without a family) working four weekdays of a week on site (not in NYC) considers renting his entire apartment on Airbnb. The opportunity cost of his entire apartment will be high when he is using the apartment. However, during weekdays when he is not in town the opportunity cost of the rental space is little to such a host. On the contrary, a local NYC resident considering to rent on Airbnb while he is away for a vacation has a small opportunity cost of the rental space during his vacation while higher costs otherwise when he/she is residing in the apartment. Another and unarguably more considerable Airbnb rentals from

a policy perspective are those rented by owners and property managers who are alleged to shift rental space from long term to the short-term rental market and therefore causing the housing crisis. These controversial rentals are also subject to certain costs. The opportunity cost of such rentals to an owner or property manager is equal to the rents their space would have fetched in a competitive long term rental market. Cleaning costs, water, and electricity usage costs also add to the variable costs. Finally, there are costs associated with the risk of having strangers in one's own property. Some implicit costs are accrued in the form of time spent in communicating with guests at the time of booking and during the time of stay and catering to the guest's need whenever they arise. I assume that the variable cost for each rental in a given market/time is linearly related to the guest capacity, i.e. ($V_{jt} \propto q_{jt}$). I can, therefore, write the cost function as:

$$C_{jt} = F_{jt} + V_{jt} = F_{jt} + c_{jt}q_{jt} \tag{1.2}$$

where c_{jt} denotes the marginal cost of dedicating a rental space catering to an additional guest. As discussed before the marginal costs of hosting will vary with the rental and time.

Given the cost structure, the host's profit from operating rental j and time t can be expressed as:

$$\pi_{jt}(p_{jt}) = (p_{jt} - c_{jt})\hat{q_{jt}}(p_{jt}) - F_{jt} \tag{1.3}$$

A host faces a downward sloping demand ($\hat{q_{jt}}(p_{jt})$) described in section 1.3.1 and sets the price sets a nightly price per accommodation (p_{jt}) that maximizes profits each day. The optimal price (p^*_{jt}) satisfies the following first order condition (FOC):

$$(p^*_{jt} - c_{jt})\frac{\partial \hat{q_{jt}}(p^*_{jt})}{\partial p^*_{jt}} + \hat{q_{jt}}(p^*_{jt}) = 0 \tag{1.4}$$

Inverting the FOC allows us to recover the marginal costs of rental j and time t when the rentals were operational on Airbnb. Basic economic theory dictates that a firm continues to operate until price falls below average variable cost. In this case, where marginal cost of

rental is assumed to be constant, I can characterize a host's decision (s_{jt}) to operate rental j on date t as :

$$s_{jt} = \begin{cases} \text{Operate} & \text{if } p^*_{jt} \geq c_{jt} \\ \text{Do not operate} & \text{otherwise} \end{cases} \tag{1.5}$$

Equation 1.5 essentially suggests that a host will operate on Airbnb when residual demand of his rental is high enough and marginal costs (opportunity cost of rental space, amenities and operational costs and discomfort and risk of hosting strangers) is low enough.

1.3.3 Equilibrium Price and Quantity

In this section, I describe the strategy for estimating the impact of the purge on equilibrium price and quantity. Since rentals and their capacity were eliminated from the market at different points in time (see figure 1.5b) and from diverse spatial locations in the NYC (see figure 1.5a), the impact can not be captured at an aggregate level. Instead, I try to capture the local impacts using variation in the eliminated capacity and their proximity to existing rentals (that were not purged by the time the purged rental was removed). In order to obtain these impacts, I specify the equilibrium outcome y_{jt} of existing rental j in market t as a function purged rental $n \in \mathcal{N}$ where $\mathcal{N}$ is a set of all purged rentals. Following the methodology similar to Muehlenbachs, Spiller, and Timmins (2015), I model the equilibrium outcome of rental j at time t as a function of the all $\|\mathcal{N}\|$ eliminated rentals in the set $\mathcal{N}$.

$$y_{jt} = \beta_0 + \sum_{n=1}^{\|\mathcal{N}\|} \delta_{jn} w_{nt} + \Gamma X_{jt} + \mu_j + \lambda_t + \nu_{jt} + \eta_{jt} + \varepsilon_{jt} \tag{1.6}$$

where $w_{nt} = 1$ if rental n is eliminated from the market by time t during the course of the purge. X_{jt} are other time varying rental characteristics and cost side controls determining equilibrium price and quantity. I also include rental fixed effects (μ_j) to capture any unobserved rental characteristics that affects rental outcome but stays constant over time. In the case of Airbnb rentals, I observe most of the time invariant characteristics of a rental with the exception

of pictures and images of rentals which makes rental FE extremely crucial. Next, I include market/time FE (λ_t) to absorb the market conditions such as overall demand and seasonality. I also include neighborhood-week FE (ν_{jt}) and neighborhood-day of week FE (η_{jt}) to absorb differences in neighborhood at weekly and day of week level. For example, during US Open tennis tournament which takes place at Flushing Meadows–Corona Park in Queens from last Monday of August and continues for two weeks, higher demand and supply is expected in surrounding neighborhoods during this period, especially in the last few days of the tournament. These set of FE account for all such confounding factors and events occurring across the NYC. δ_{jn} can be interpreted as the impact on an existing rental j from removal of rental n. This impact depends on the capacity and proximity of the purged rental. If d_{jn} denotes distance (in miles) between rental j and purged rental n and s_n denotes the capacity (number of guests) purged rental n accommodates, then δ_{jn} can be decomposed as:

$$\begin{aligned}\delta_{jn} = s_n[&\beta_{1,(0-0.5)}\mathbb{I}(0 \leq d_{jn} < 0.5) \\ &+ \beta_{1,(0.5-1)}\mathbb{I}(0.5 \leq d_{jn} < 1) \\ &+ \beta_{1,(1-2)}\mathbb{I}(1 \leq d_{jn} < 2) \\ &+ \beta_{1,(2-5)}\mathbb{I}(2 \leq d_{jn} < 5) \\ &+ \beta_{1,(5-10)}\mathbb{I}(5 \leq d_{jn} < 10) \\ &+ \beta_{1,(10-\infty)}\mathbb{I}(10 \leq d_{jn} < \infty)]\end{aligned} \tag{1.7}$$

where $\mathbb{I}(a \leq d_{jn} < b) = 1$ if the distance d_{jn} between j and n falls in the a given interval $[a, b)$. The corresponding parameter $\beta_{1,(a-b)}$ captures the average effect of purged capacity

within proximity range of a to b miles. Substituting 1.9 into 1.6, we get:

$$
\begin{aligned}
y_{jt} = \beta_0 + & \\
& + \beta_{1,(0-0.5)} \sum_{n=1}^{\|\mathcal{N}\|} \mathbb{I}(0 \leq d_{jn} < 0.5) s_n w_{jt} \\
& + \beta_{1,(0.5-1)} \sum_{n=1}^{\|\mathcal{N}\|} \mathbb{I}(0.5 \leq d_{jn} < 1) s_n w_{jt} \\
& + \beta_{1,(1-2)} \sum_{n=1}^{\|\mathcal{N}\|} \mathbb{I}(1 \leq d_{jn} < 2) s_n w_{jt} \\
& + \beta_{1,(2-5)} \sum_{n=1}^{\|\mathcal{N}\|} \mathbb{I}(2 \leq d_{jn} < 5) s_n w_{jt} \\
& + \beta_{1,(5-10)} \sum_{n=1}^{\|\mathcal{N}\|} \mathbb{I}(5 \leq d_{jn} < 10) s_n w_{jt} \\
& + \beta_{1,(10-\infty)} \sum_{n=1}^{\|\mathcal{N}\|} \mathbb{I}(10 \leq d_{jn} < \infty) s_n w_{jt} \\
& + \Gamma X_{jt} + \mu_j + \lambda_t + \nu_{jt} + \eta_{jt} + \varepsilon_{jt}
\end{aligned}
\tag{1.8}
$$

Notice that for any given interval $[a, b)$, the term $\sum_{n=1}^{\|\mathcal{N}\|} \mathbb{I}(a \leq d_{jn} < b) s_n w_{jt}$ represents the capacity lost from buffer zone enclosed between the geographical circles of radius a and $b(> a)$ with center at the location of rental j by time t. Since, loss in capacity at a large enough distance from a rental will have nearly no effect on equilibrium price and quantity. I assume $\beta_{1,(10-\infty)} = 0$ i.e. if the purged rental will have no effect on the price and quantity of an existing rental if former is more than 10 miles away from the later. This assumption is crucial for the identification of other parameters in the above equation. After gathering all

the information and assumption together, the estimating equation takes the following form.

$$
\begin{aligned}
y_{jt} = \beta_0 + & \\
& + \beta_{1,(0-0.5)}(\text{Capacity lost between 0 and 0.5 miles})_{jt} \\
& + \beta_{1,(0.5-1)}(\text{Capacity lost between 0.5 and 1 miles})_{jt} \\
& + \beta_{1,(1-2)}(\text{Capacity lost between 1 and 2 miles})_{jt} \\
& + \beta_{1,(2-5)}(\text{Capacity lost between 2 and 5 miles})_{jt} \\
& + \beta_{1,(5-10)}(\text{Capacity lost between 5 and 10 miles})_{jt} \\
& + \Gamma X_{jt} + \mu_j + \lambda_t + \nu_{jt} + \eta_{jt} + \varepsilon_{jt}
\end{aligned}
\tag{1.9}
$$

where parameters $\beta_{1,(0-0.5)}$, $\beta_{1,(0.5-1)}$, $\beta_{1,(1-2)}$, $\beta_{1,(2-5)}$, and $\beta_{1,(5-10)}$ measure the impact on equilibrium outcome for different buffer zones. y_{jt} can be replaced by log of prices ($ln(p_{jt})$) or quantity (q_{jt}).

Caveat : It is important to note that I do not observe prices and quantities of rentals on days when they do not appear in the market. This does not occur randomly, but instead when hosts perceive non positive expected profits, i.e. when price or quantity demanded are low, they choose to stay out of the Airbnb market. Thus, sufficiently low price and quantity remain unobserved depending on rental costs and thereby causing Sample Selection problem. Therefore, the naive estimates of changes in price and quantity are biased. The ideal way to deal with such a problem is to use Heckman's Correction where selection depends on the cost side variables of the rentals. However, running a Probit regression for selection is computationally infeasible for the data of this magnitude at this moment. Intuitively, since only very low prices and quantities tend to remain unobserved for a rental, the changes in price and quantity (in absolute terms) are underestimated. Thus, we have reasons to believe that final welfare losses can only be larger than currently estimated impacts.

1.4 Results

This section presents the estimation results and welfare calculations. I begin by estimating a residual demand curve for each rental and compute the willingness to pay. Using the demand estimates and observed prices in the data, I calculate the consumer surplus derived from each rental in the market. Then, I use the demand function and the assumed market structure to recover marginal cost and markups. These along with equilibrium outcomes enable me to obtain the profits of each rental. In order to realize the effect of the purge on the welfare of guests (consumers) and the hosts (sellers), I first estimate the extent to which the equilibrium price and quantity (of accommodation) respond to the elimination of competing rentals from the market. Finally, I re-calibrate the counterfactual using the price-quantity response to purge. Assuming that the price parameter of the demand does not change due post purge, I compute changes in the welfare due to the purge.

1.4.1 Demand and Consumer Surplus

I estimate demand parameters in equation 1.1 using daily panel data on the quantity and price of rental accommodation. The quantity variable is zero if a rental is not booked on a given day/market. The quantity equals the accommodation capacity (number of persons) of rental when it is "booked". In particular, I assume that consumer demand is capacity specific and that a consumer booking a rental on Airbnb utilizes the entire capacity. The price of a unit accommodation is calculated by splitting the price each night of a rental into the number of guests it can accommodate.

Trust and reputation system is an integral part of Airbnb market design. The system is extensively maintained in the form of reviews and feedback that past travelers share about the rental on the website. Not only do these reviews remove information asymmetry which is otherwise difficult to accomplish between two strangers (host and guest) meeting online, they also differentiate rentals on the quality of space and services for future customers. Each rental has a growing stock of reviews which plays a vital role in consumers demand and thus the

Table 1.4 – Demand Estimation Results

	(1) Quantity	(2) Quantity	(3) Quantity(IV)	(4) Quantity(IV)	(5) Quantity(IV)	(6) Quantity(IV)	(7) Quantity(IV)	(8) Quantity(IV)
ln(Price per guest)	-0.13***	-0.13***	-8.37***	-6.63***	-6.86***	-0.86***	-6.67***	-6.79***
	(0.012)	(0.012)	(0.21)	(0.14)	(0.14)	(0.021)	(0.14)	(0.14)
Quantity (lag 1)				0.70***	0.69***	0.75***	0.69***	0.69***
				(0.0015)	(0.0016)	(0.0013)	(0.0015)	(0.0016)
Quantity (lag 2)					0.011***	0.063***	0.012***	0.011***
					(0.0012)	(0.0012)	(0.0012)	(0.0012)
Rental Age		-622822.1	-3695374.4	-3145193.2		0.000042***	-277927575.5	-3131707.1
		(11181544.5)	(17199363.9)	(12710591.5)		(0.0000098)	(478512411.5)	(12924817.8)
Host Age		622821.8	3695373.4	3145192.6		-0.0000065	269561883.0	3131706.5
		(11181686.6)	(17199887.8)	(12710799.2)		(0.0000086)	(463489480.4)	(12925032.2)
Past booked days (rental)		-0.0014***	-0.0018***	-0.0015***		0.00073***	-0.0015***	-0.0015***
		(0.00010)	(0.00018)	(0.00014)		(0.000057)	(0.00014)	(0.00014)
Past booked days (host)		-0.00034***	-0.00020***	0.000065		-0.00039***	0.000078	0.000070
		(0.000042)	(0.000066)	(0.000050)		(0.000020)	(0.000049)	(0.000052)
Past guests count (rental)		0.0091***	0.013***	0.0087***		-0.000021	0.0090***	0.0088***
		(0.00044)	(0.00073)	(0.00055)		(0.00026)	(0.00054)	(0.00056)
Past guests count (host)		0.00011	0.000048	-0.00028*		0.00054***	-0.00037**	-0.00029*
		(0.00014)	(0.00021)	(0.00016)		(0.000070)	(0.00015)	(0.00017)
Property FE	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Neighborhood-Week FE	Yes	Yes	Yes	Yes	Yes	No	No	Yes
Neighborhood-DOW FE	Yes	Yes	Yes	Yes	Yes	No	No	Yes
Observations	25153461	25153461	25153461	25153461	25153461	25154404	25153664	25153461
r2	0.48	0.48	0.32	0.33	0.30	0.67	0.32	0.31
r2_within	0.00032	0.0024	-0.30	-0.30	-0.36		-0.31	-0.34
r2_a	0.48	0.48	0.33	0.33	0.29	0.67	0.32	0.31

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Notes: (1) The Within R-Square computes the R-Square of a regression where every variable has already been demeaned with respect to all the fixed effects. (2) If R-square is negative, it is because in the regression the variables are demeaned first with respect to all fixed effects and then demeaned outcome variable is regressed on demanded regressors without a constant term. (3) These results have been obtained using Stata command *reghdfe* developed by Correia (2016).

price of the rental. Exclusion of reviews as exogenous covariates in the demand specification will, therefore, bias the price sensitivity. Though Rental FE absorbs all variation (observed and unobserved) in demand due to time-invariant factors, they can not substitute for the temporal nature of reviews growth for each rental. In this paper, I use a proxy for reviews information by using covariates not directly observed by consumers on the website but are expected to be correlated with reviews. These include the age of rental and host on Airbnb, the number of bookings and the number of days rental was booked, the number of days of operation in the market both at rental as well as at host level. Circumventing reviews in this manner is only a simplification and not a foolproof solution. I shall address this issue later.

To account for endogeneity in price, I use instruments constructed from the past rental activity at an operational level (and not at booking level). It includes the number of days of operation and its interaction with the capacity of operation in a certain time intervals in the past. The data of a rental on a given day exists in three states - "Available", "Booked" and "Blocked". Being "Available" or "Booked" implies that a given rental is operational (active) in the market. Specifically, I pick intervals of time say past one week ($[t-7, t-1]$) and for each rental j construct dummy indicating whether the rental was operational p days of that interval where $p \in \{0, 1, \ldots, 7\}$. I interact these variable with the capacity of past operation. I construct similar dummy measures for larger intervals $[t-15, t-1]$, $[t-22, t-1]$, $[t-29, t-1]$ and $[t-36, t-1]$ along with their interactions with capacity directly from the data . Conditional on other factors that may affect the quantity demanded, these set of variables, I argue, will vary with the opportunity cost of hosting and therefore with the prices.

Table 1.4 reports the estimates of demand. The specification in columns (1) and (2) includes all the FEs and time-varying covariates but estimate demand without accounting for price endogeneity. The demand parameter turns out to be -0.13 in both specifications suggesting that the demand is highly inelastic even at a lower level of capacity demanded. This seems unrealistic in a market with such a large number of imperfect substitutes. Given the specification of demand (linear-log), a small estimated price parameter also yields ridiculously

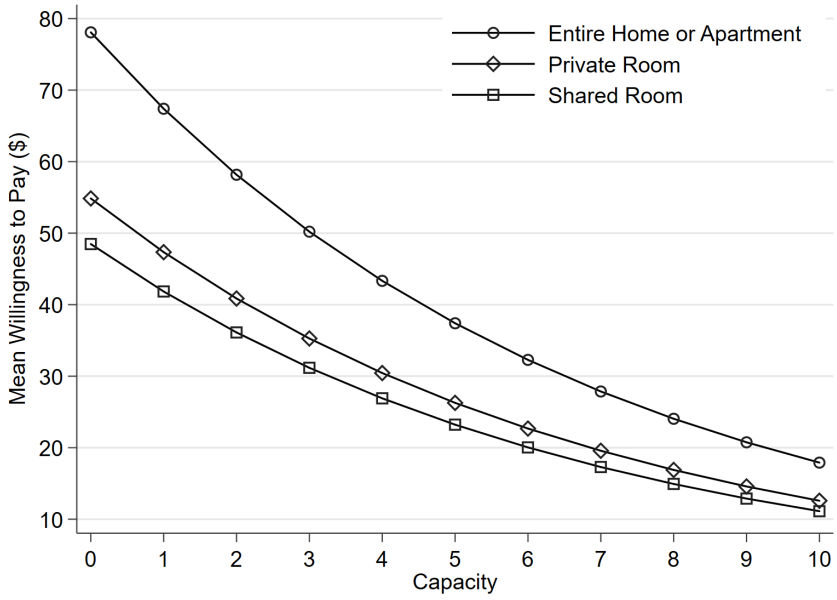


Figure 1.6 – Willingness To Pay
Notes: The WTP has been averaged over all rentals and time within each rental type category.

high welfare numbers. The coefficient falls to a more realistic estimate of −7.60 when I use cost side instruments for the price as shown in column (3), and therefore indicating the importance and validity of instruments in demand estimation. In column (4), I add lagged demand in the specification to incorporate the fact that current demand depends on the lagged demand as a booking made by one customer may span for more that one consecutive day. The next few columns test the relevance of covariates and FEs in the IV estimation. In the final specification (8), I include two lagged demand terms and time-varying rentals characteristics to proxy for varying quality of rental with time and multiple sets of FEs described in section 1.3.3. The coefficient of price sensitivity parameter is estimated to be −6.79 with a standard error of 0.14.

Next, we look at some implications of the estimated demand curve parameters. Figure 1.6 shows the average willingness to pay (WTP) as a function of accommodation capacity (number of persons). Within each rental type category differentiating rentals by space and

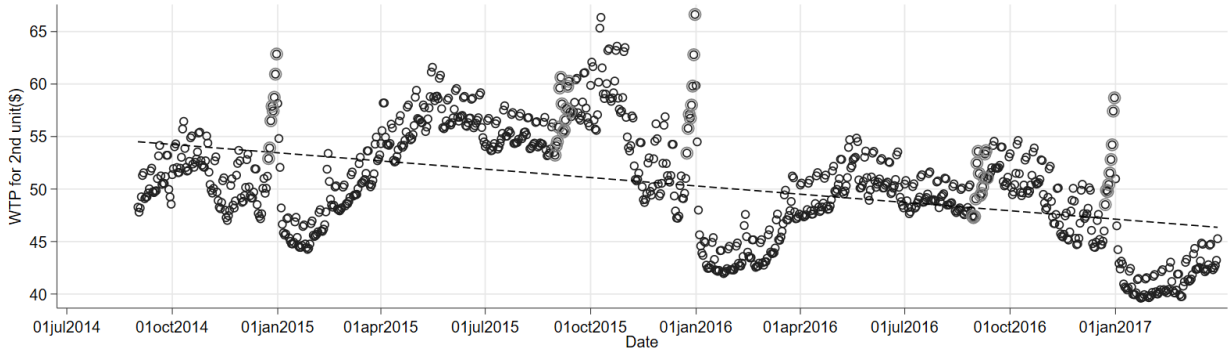


FIGURE 1.7 – WILLINGNESS TO PAY ACROSS MARKETS
Notes: For each time/market, the willingness to pay for 2^{nd} unit of accommodation has been averaged and then plotted against time.

privacy they offer, the WTP has been averaged over rentals and time. As expected, the "Entire Home/Apartment" rental which offers larger spaces and maximum privacy within a rental has the higher WTP compared to "Private Room" and "Shared Room" rental. For 2^{nd} unit of capacity, the WTP for "Entire Home/Apartment" rental is about 56$ while "Private Room" and "Shared Room" rentals have 40$ and 33$ respectively (Table 1.5). These differences in WTP is indicative of consumers preference for quality in terms of space and privacy and that the demand specification captures heterogeneity carefully. Willingness to pay for rentals on average increases during periods of high market demand such as in September and October and during the holiday season in late December as shown in Figure 1.7. From the same figure, it is also evident that as the number of operational rentals increases with time, the WTP for per rental gradually falls at a given level of quantity and that the residual demand for each rental shrinks with time.

Table 1.5 describes the distribution of WTP across rentals. The table shows that residual demand faced by rentals varies across space type, property type, and location. For example, Manhattan has a higher demand for rentals compared to other boroughs. Notice that purged rentals are not very different in terms of their demand from the non-purged rentals. Within "Entire Home/Apartment" rental type, the WTP for a purged rental is only 1$ less than their contemporaries suggesting that consumers view purged rentals similar to competing

Table 1.5 – Willingness To Pay Across Rentals

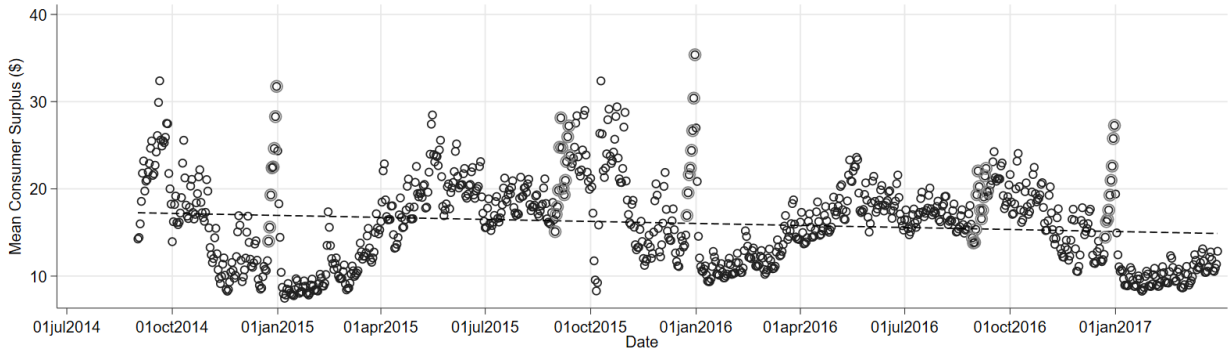
	(1)	(2)	(3)	(4)
	WTP ($)			
	All Rentals		Purged Rentals	
VARIABLES	mean	sd	mean	sd
All rentals	47.96	43.90	54.75	37.82
Space-Type				
Entire Home/Apartment	55.77	42.49	54.75	37.82
Private Room	40.15	43.68		
Shared Room	33.28	43.08		
Property-Type				
Apartment/Condominium/Loft/Townhouse	48.89	42.15	54.81	37.80
House	33.70	35.88		
Bungalow/Castle/Guesthouse/Vacation home/Villa	40.11	47.70		
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	46.49	38.34		
Location				
Bronx	28.87	26.89	23.96	7.808
Brooklyn	38.07	31.14	40.41	27.00
Manhattan	57.75	46.21	60.72	39.78
Queens	33.83	60.65	33.24	20.08
All Staten Island	42.71	35.91	59.28	38.98

Notes: The average willingness to pay for the 2^{nd} unit of stay per night has been calculated for each rental and then summary statistics have been reported across rentals and by rental categories.

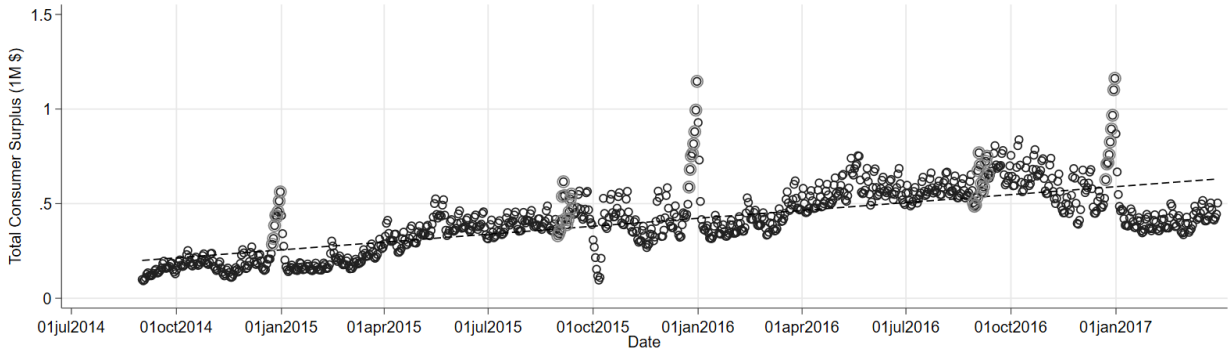
non-purged rentals.

I compute the consumer surplus by integrating under the demand function for a given rental-time (jt) pair. Figure 1.8 plots the aggregate trends in consumer surplus over time. In panel 1.8a, I average the surplus generated by rentals in each market which decreases as we move across market/time. The next panel (figure 1.8b) plots the total surplus to consumers generated by rentals in a given market. While the total surplus has been increasing indicating an increase in the total market demand, decreasing average surplus per rental suggests that the market demand is being distributed among a larger number of sellers with time.

Table 2.5 presents the variation in daily-average surplus and total-surplus across different rental types. Average daily consumer surplus of purged rentals (27.66$) is higher than all rental average (12.93$) in the market, primarily because of former's above average quality and



(A) Average Consumer Surplus



(B) Total Consumer Surplus

FIGURE 1.8 – CONSUMER SURPLUS ACROSS MARKETS

their ability to operate at a higher capacity in terms of providing accommodation. Despite being able to operate less due to forced elimination from the market, the purged rental contribute significantly more to the welfare of consumers. In "Entire Home/Apartment" category, the purge rentals have generated about 8$ more surplus average rental in the same category.

The total consumer welfare is calculated by aggregating surplus across all rental-time (jt) pairs. Consumers in the 31 month period have gained $392M$\$ by booking rentals on Airbnb. Column (5) of table 2.5 reports the total consumer welfare and its decomposition across space type, property type and location. A major fraction of surplus (about 82%) originate from "Entire Home/Apartment" rentals out of which 12.5% ($40M$\$) of consumer surplus came from

Table 1.6 – Consumer Surplus Across Rentals

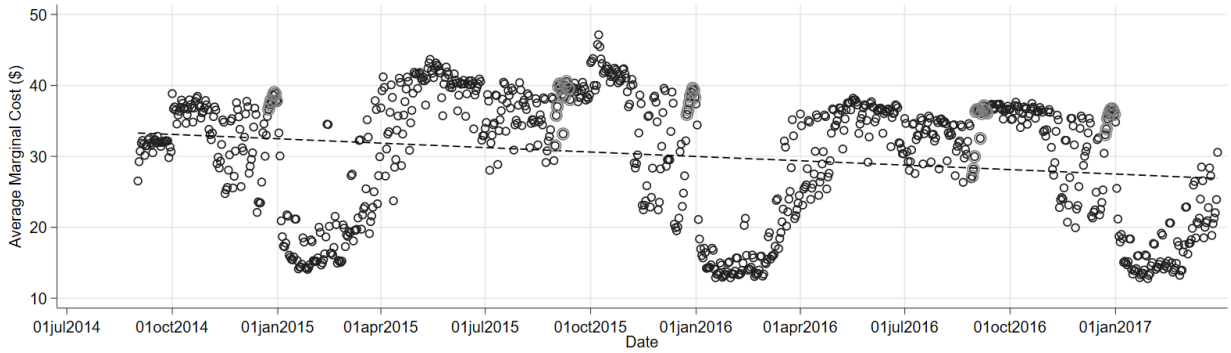
	(1)	(2)	(3)	(4)	(5)	(6)
	Average CS ($)				Total CS (1000$)	
	All Rentals		Purged Rentals		All Rentals	Purged Rentals
VARIABLES	mean	sd	mean	sd	sum	sum
All rentals	12.93	31.69	27.66	48.35	391,841	40,272
Space-Type						
Entire Home/Apartment	20.58	41.05	27.66	48.35	321,728	40,272
Private Room	4.904	12.65			66,138	
Shared Room	3.394	7.110			3,971	
Property-Type						
Apartment/Condominium/Loft/Townhouse	12.86	29.69	27.69	48.37	353,998	40,269
House	14.40	51.81			33,518	
Bungalow/Castle/Guesthouse/Vacation home/Villa	9.874	28.77			312.9	
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	4.814	7.297			149.1	
Location						
Bronx	4.582	8.668	14.37	25.04	2,148	84.24
Brooklyn	9.943	27.28	19.97	33.33	110,430	5,994
Manhattan	16.16	35.82	31.14	52.94	256,167	33,020
Queens	7.339	20.58	12.65	23.48	20,081	996.2
All Staten Island	11.90	31.85	18.86	21.69	3,012	177.5

Notes: The consumer surplus has been first averaged within rentals before reporting the summary statistics. The total consumer surplus is calculated by aggregating over all rental-time combination.

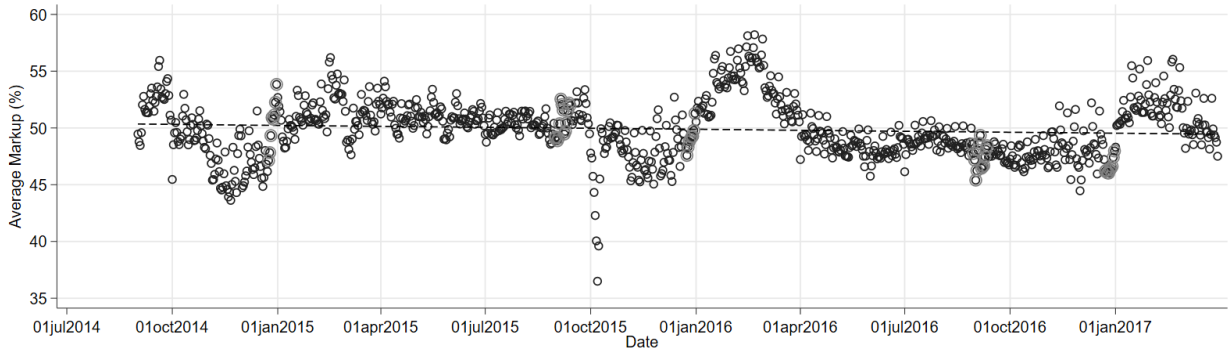
purged rentals before being purged.

1.4.2 Costs Markup and Producers Surplus

I recover markup, and the marginal cost of rentals using the FOC described in section 1.3.2, assuming monopolistic competition structure of Airbnb NYC marketplace. The marginal cost of creating capacity for an additional person/guest is about 28.66$ on average which is about 864$ aggregated over a month - the cost incurred by an Airbnb host when he decides to operate his rentals for an entire month with just one guest. High marginal costs are in line with the fact that living spaces have a high opportunity cost in the city. The distribution of the marginal cost across space type, property type and location resonates with this feature of the NYC market. For example, the cost of creating capacity is higher for a better quality



(A) Average Marginal Cost



(B) Average (Sales Weighted) Markup

FIGURE 1.9 – MARGINAL COST AND MARKUP ACROSS MARKETS
Notes: The markups are sales weighted where the weights are calculated for each market/time.

rental in terms of privacy and space. The "Entire Home/Apartment" rentals have average marginal costs of 31.99$ while "Private Rooms" and "Shared Rooms" cost about 25.44$ and 20.89$ respectively. Marginal costs vary with locations and neighborhoods. Rentals in Manhattan where living costs are highest compared to other boroughs also face higher costs (34.22$) of renting on Airbnb. Notice that the purged rentals (Table 1.7) are cost-efficient compared to market average which explains why they operate at higher capacity and generate more consumer welfare.

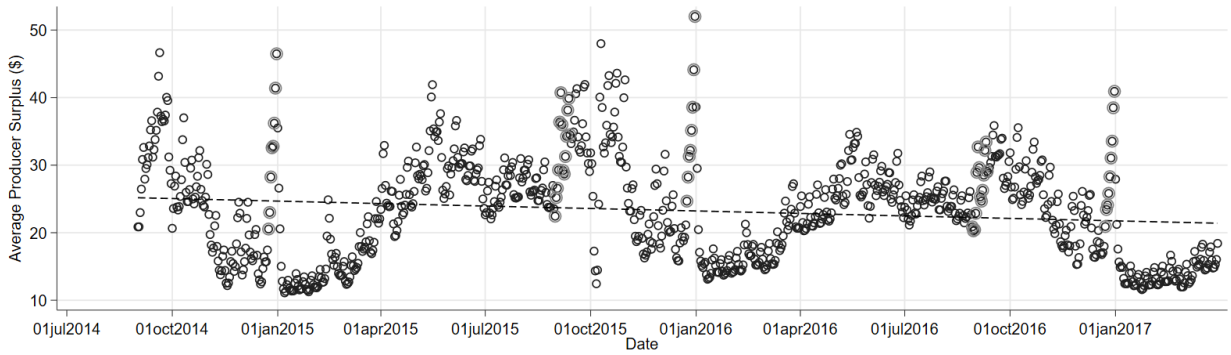
Figure 1.9a show trends in recovered marginal cost across time. These aggregate trends (averaged for each market/time) in the marginal cost indicate that the least cost (more

Table 1.7 – Markup and Marginal Cost Across Rentals

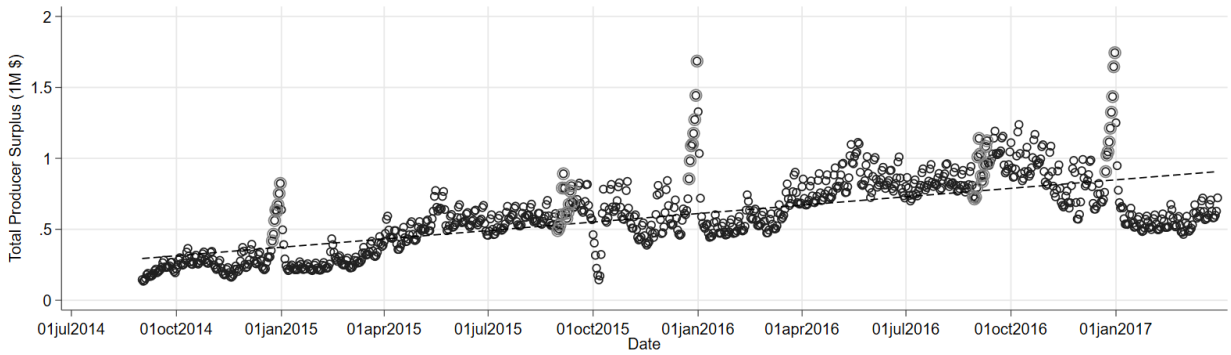
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Marginal Cost ($)				Markup (%)			
	All Rentals		Purged Rentals		All Rentals		Purged Rentals	
VARIABLES	mean	sd	mean	sd	mean	sd	mean	sd
All rentals	28.66	28.84	28.13	21.97	29.15	20.57	44.50	23.21
Space-Type								
Entire Home/Apartment	31.99	29.34	28.13	21.97	35.40	22.61	44.50	23.21
Private Room	25.44	27.73			22.44	15.43		
Shared Room	20.89	27.92			23.20	17.66		
Property-Type								
Apartment/Condominium/Loft/Townhouse	29.27	28.16	28.16	21.96	29.05	20.38	44.50	23.21
House	19.42	22.30			30.72	22.82		
Bungalow/Castle/Guesthouse/Vacation home/Villa	24.24	33.25			30.19	18.71		
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	28.67	20.44			23.32	17.42		
Location								
Bronx	17.61	17.01	13.06	5.212	25.71	19.56	40.89	25.47
Brooklyn	23.11	20.63	22.00	17.26	28.08	20.17	43.35	23.27
Manhattan	34.22	31.83	30.65	23.12	30.40	20.89	45.27	23.19
Queens	20.40	33.86	19.28	14.19	26.77	20.01	39.42	22.40
Staten Island	25.40	23.92	34.87	25.17	28.72	20.06	39.35	20.19

Notes: The reported markups are sales weighted averages. But the sales weights are actually calculated within a rental.

efficient) rentals can survive and operate in the market when the demand is low especially during January and February. On the other hand, high-cost (less efficient) rentals can enter/operate (along with low-cost rentals) when demand is high enough. For instance, a large number of sellers including those with larger costs operate in the market to benefit from high demand during the holiday season (last week of December). This is possible since Airbnb has reduced the cost of entry into the market. As we move across the market, it also appears that rentals in the markets are becoming more cost-efficient. A part of the reason for decreasing costs could be increasing experience. However, with increasing, long term rents (suggested by) in NYC, the opportunity cost of renting on Airbnb should increase. A decrease in marginal cost could also result from the bias transmitted from the demand side since I did not account for the time-varying reviews while estimating demand.



(A) Average Profits



(B) Total Producer Surplus

FIGURE 1.10 – PRODUCER SURPLUS ACROSS MARKETS

Next, we look at the markups charged by the rental operators. The aggregated trends in (sales weighted) markups show that sellers enjoy decent markup (averaging about 45% to 55%) on their rented spaces (Figure 1.9b). Though the markups are high, the sellers find it hard to raise the markups further during the peak demand season as more peer sellers enter the market and increase the supply of rental accommodation. It is in contrast to hotels in NYC during the peak seasons where capacity is constrained, and hotels respond by increasing price and charge higher markup (see Farronato and Fradkin 2018). Table 1.7 presents the variation in the markups across rentals. Higher markups (35.40% on "Entire Home/Apartment") are associated with higher quality. Boat, Cabin, Camper, RV, Hut, Lighthouse, Planes and other forms of small space rentals enjoy significantly lower markups than other property types.

Table 1.8 – Profit and Producer Surplus Across Rentals

	(1)	(2)	(3)	(4)	(5)	(6)
	Average Profits ($)				Total Profits (1000$)	
	All Rentals		Purged Rentals		All Rentals	Purged Rentals
VARIABLES	mean	sd	mean	sd	sum	sum
All rentals	19.20	33.91	37.70	48.25	566,404	54,299
Space-Type						
Entire Home/Apartment	29.86	43.15	37.70	48.25	452,132	54,299
Private Room	8.079	12.11			107,788	
Shared Room	5.473	9.560			6,478	
Property-Type						
Apartment/Condominium/Loft/Townhouse	19.42	33.24	37.74	48.26	521,380	54,297
House	17.24	42.78			39,189	
Bungalow/Castle/Guesthouse/Vacation home/Villa	13.09	26.60			460.2	
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	7.413	8.232			233.3	
Location						
Bronx	7.155	11.49	19.09	27.73	3,395	121.4
Brooklyn	14.65	26.47	27.37	34.34	158,014	8,121
Manhattan	24.14	39.50	42.36	52.19	372,799	44,423
Queens	10.59	19.14	17.44	24.02	28,048	1,350
Staten Island	17.36	37.36	30.26	34.48	4,142	283.0

Notes: The producer surplus has been first averaged within rentals before reporting the summary statistics. The total producer surplus is calculated by aggregating over all rental-time combination.

This evidence suggests that in an imperfectly competitive market, the market power of rentals originate from consumer perceived differences in products and the number of competing rentals in the market. The purged rentals too enjoyed a higher average markup of 44.50% compared to the market average in "Entire Home/Apartment" category.

Lastly, I compute producer surplus by multiplying the difference between price and marginal cost ($p_{jt} - c_{jt}$) by the quantity of rental accommodation (q_{jt}) for each rental-time combination. Figure 1.10 displays average and total profits made by rentals on Airbnb. While the average profitability of rentals is decreasing over time, total profits generated by rentals combined depicts increasing trends. This observation is in sync long-run outcomes in a monopolistic competitive market where entry of sellers occurs until the economic profits of firms diminish to zero. Rentals are most profitable when the demand is high incentivizing

sellers to enter the market and gain from market demand. For instance, on new years eve of 2014, 2015 and 2016, total sellers profits were about $0.84M\$$, $1.72M\$$ and $1.78M\$$ respectively.

Table 1.8 reports the distribution of profits across different rental categories. The profitability of an average rental is about 19\$ while that of a purged rental is 37.70\$. The purge rentals are also more profitable than an average "Entire Home/Apartment" rental (29.68\$). Profitability also varies with the location with Manhattan rentals being the most profitable. The total producer surplus is calculated by aggregating the surplus across all rental-time(jt) pairs. The total surplus in 31 month period is about $566M\$$, most of which comes from the "Entire Home/Apartment" rental(80%). The purged rentals before being evicted from the market generated $54M\$$ profits which are about 10% of the total surplus generated in 31 months.

1.4.3 Effect of Purge on Price and Quantity

I follow the strategy laid out in section 1.3.3 to estimate the extent to which the equilibrium price and quantity of Airbnb accommodation have changed as rentals are eliminated from the market as a result of the purge. I model equilibrium price and quantity (number of persons) of each rental at a given time/market as a function of the net loss in rental capacity by that time and distance from the purged rental. A key identifying assumption is that any change in rental capacity (or rentals) that originates more than 10 miles away from a given rental seizes to affect equilibrium outcomes of that rental. Table 1.9 presents the results of specification (1.9) when outcome variable on the left hand side is log of price ($ln(p_{jt})$). The prices are adjusted to inflation using the monthly CPI data obtained from the Bureau of Labor Statistics. Column (1) through (5) reports coefficients that denote a percentage change in the prices as capacity (or the number of rentals) changes in multiple buffer zones at a different distance range.

I find that prices are positively associated with the loss of capacity in nearest buffer zones ($0 - 0.5$ mile and $0.5 - 1$ mile). For subsequent buffers, not only does the size of

Table 1.9 – Impact On Prices

	(1) $ln(Price)(X100\%)$	(2) $ln(Price)(X100\%)$	(3) $ln(Price)(X100\%)$	(4) $ln(Price)(X100\%)$	(5) $ln(Price)(X100\%)$
Purged capacity ($0 \leq distance < 0.5mile$)	0.0028***	0.0033***	0.0033***	0.0032***	0.0032***
	(0.00086)	(0.00087)	(0.00087)	(0.00087)	(0.00088)
Purged capacity ($0.5 \leq distance < 1mile$)		0.0016***	0.0016***	0.0015**	0.0015**
		(0.00061)	(0.00061)	(0.00061)	(0.00062)
Purged capacity ($1 \leq distance < 2mile$)			-0.00000068	-0.0000080	-0.0000095
			(0.00026)	(0.00026)	(0.00031)
Purged capacity ($2 \leq distance < 5mile$)				0.00021	0.00021
				(0.00014)	(0.00020)
Purged capacity ($5 \leq distance < 10mile$)					-0.0000019
					(0.00022)
Rental Age	-48108769.1	-48751843.8	-48751208.8	-47135913.6	-47136812.9
	(161634379.6)	(161614592.2)	(161611307.0)	(161649352.1)	(161660992.2)
Host Age	48108760.8	48751835.5	48751200.4	47135905.4	47136804.6
	(161639453.0)	(161619716.8)	(161616476.7)	(161654527.9)	(161666167.6)
Past booked days (rental)	-0.0067***	-0.0067***	-0.0067***	-0.0068***	-0.0068***
	(0.0021)	(0.0021)	(0.0021)	(0.0021)	(0.0021)
Past booked days (host)	0.0018**	0.0018**	0.0018**	0.0019**	0.0019**
	(0.00079)	(0.00079)	(0.00079)	(0.00079)	(0.00079)
Past guests count (rental)	0.057***	0.057***	0.057***	0.057***	0.057***
	(0.0084)	(0.0084)	(0.0084)	(0.0084)	(0.0084)
Past guests count (host)	-0.0019	-0.0020	-0.0020	-0.0021	-0.0021
	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0025)
Rental FE	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes
Neighborhood-Week FE	Yes	Yes	Yes	Yes	Yes
Neighborhood-DOW FE	Yes	Yes	Yes	Yes	Yes
Supply side controls	Yes	Yes	Yes	Yes	Yes
Observations	25153461	25153461	25153461	25153461	25153461
r2	0.92	0.92	0.92	0.92	0.92
r2_within	0.0042	0.0043	0.0043	0.0043	0.0043
r2_a	0.92	0.92	0.92	0.92	0.92

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Notes: (1) Purged capacity variable denote the rental capacity purged from each buffer zone around the a given rental by a given time. (2) The Within R-Square computes the R-Square of a regression where every variable has already been demeaned with respect to all the fixed effects. (3) If R-square is negative, it is because in the regression the variables are demeaned first with respect to all fixed effects and then demeaned outcome variable is regressed on demanded regressors without a constant term. (4) These results have been obtained using Stata command *reghdfe* developed by Correia (2016).

the coefficient is small; it also becomes insignificant for inference. Under the identifying assumption, column (5) reports causal estimates of the impact on the price of existing rentals due to capacity adjustments after the purge. Price per unit of accommodation increases most (0.0032%) with a unit drop in capacity in geographically closest buffer ($0 - 0.5$ mile). The effect size shrinks as the distance between the purged rental and existing rental increases. For every unit lost (capacity in terms of the number of persons), the price increases on average by 0.0074% in the $0.5 - 1$ mile buffer. Although the coefficients of increasing buffer size beyond one-mile distance are insignificant, they indicate that the effect size diminishes with decreasing proximity. This finding also reassures the validity of the identifying assumption that assumes the effect on price to be negligible (zero) when the purged rental is more than 10 miles away from existing rental. From columns (1) through (5), the coefficients do not change significantly across specifications with the inclusion of additional and distant buffer zones from the previous column. It indicates a smaller response of rental prices to changes in capacity at more considerable distances. The price coefficient also reflects the average impact on the prices of all the rentals falling in a given proximity range of a purged rental. So, even though the estimates may seem small, they capture a unanimous (direct and indirect) impact on prices of all the rentals in the proximity range of a purged rental.

Next, table 1.10 presents regression results of quantity (q_{jt}) on capacity lost in different buffer areas. Each column represents the effect of loss in capacity in each buffer zone considering the buffer not included in regression as the base. Although the effect is negative throughout suggesting a decrease in the equilibrium levels of quantity, the inference is difficult as the standard errors are large. As with prices, column (5) reports causal estimates of the effect on equilibrium quantity and under the same identifying assumption. The effect is negative and diminishes as the buffer distance (i.e., the distance between purged and existing rental) increases.

The regression of price-quantity response assumes that the distribution of reviews is iid not only across rentals but also within rentals. It may not be particularly accurate, and thus the estimates are subject to bias. We look at this issue in detail later. Moreover,

Table 1.10 – Impact On Quantity

	(1) Quantity	(2) Quantity	(3) Quantity	(4) Quantity	(5) Quantity
Purged capacity ($0 \leq distance < 0.5mile$)	-0.000063	-0.000074*	-0.000077**	-0.000075*	-0.000078**
	(0.000039)	(0.000039)	(0.000039)	(0.000039)	(0.000039)
Purged capacity ($0.5 \leq distance < 1mile$)		-0.000040	-0.000039	-0.000036	-0.000040
		(0.000026)	(0.000026)	(0.000026)	(0.000027)
Purged capacity ($1 \leq distance < 2mile$)			-0.000013	-0.000013	-0.000017
			(0.0000099)	(0.0000099)	(0.000012)
Purged capacity ($2 \leq distance < 5mile$)				-0.0000065	-0.000010
				(0.0000052)	(0.0000075)
Purged capacity ($5 \leq distance < 10mile$)					-0.0000056
					(0.0000096)
Rental Age	2312112.2	2327982.3	2340210.7	2291463.6	2288825.0
	(11110430.2)	(11110867.3)	(11111019.6)	(11110318.0)	(11111213.5)
Host Age	-2312112.3	-2327982.4	-2340210.8	-2291463.7	-2288825.1
	(11110554.3)	(11110991.5)	(11111142.4)	(11110441.2)	(11111336.5)
Past booked days (rental)	-0.0012***	-0.0012***	-0.0012***	-0.0012***	-0.0012***
	(0.00010)	(0.00010)	(0.00010)	(0.00010)	(0.00010)
Past booked days (host)	-0.00034***	-0.00034***	-0.00034***	-0.00034***	-0.00034***
	(0.000042)	(0.000042)	(0.000042)	(0.000042)	(0.000041)
Past guests count (rental)	0.0080***	0.0080***	0.0080***	0.0080***	0.0080***
	(0.00043)	(0.00043)	(0.00043)	(0.00043)	(0.00043)
Past guests count (host)	0.00016	0.00016	0.00016	0.00016	0.00016
	(0.00014)	(0.00014)	(0.00014)	(0.00014)	(0.00014)
Rental FE	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes
Neighborhood-Week FE	Yes	Yes	Yes	Yes	Yes
Neighborhood-DOW FE	Yes	Yes	Yes	Yes	Yes
Supply side controls	Yes	Yes	Yes	Yes	Yes
Observations	25153461	25153461	25153461	25153461	25153461
r2	0.49	0.49	0.49	0.49	0.49
r2_within	0.013	0.013	0.013	0.013	0.013
r2_a	0.49	0.49	0.49	0.49	0.49

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Notes: (1) Purged capacity variable denote the rental capacity purged from each buffer zone around the a given rental by a given time. (2) The Within R-Square computes the R-Square of a regression where every variable has already been demeaned with respect to all the fixed effects. (3) If R-square is negative, it is because in the regression the variables are demeaned first with respect to all fixed effects and then demeaned outcome variable is regressed on demanded regressors without a constant term. (4) These results have been obtained using Stata command *reghdfe* developed by Correia (2016).

within each rental, the reviews should also be correlated across time and thereby leading to serial correlation in error term. To obtain correct standard errors arising from serially correlated errors structure, I compute and report cluster-robust standard errors. Without such clustering, all the coefficients in the price and quantity regressions were highly significant as the standard errors were too low.

These results suggest that the price-quantity response has been highest when competitors closest to an existing rental are removed from the market. They signify that closer substitutes of purged rentals will show higher response despite imprecisions in the estimation which stems from not accounting for exact substitution patterns. Nevertheless, I use these estimated coefficients to compute counterfactual prices and quantities that would have prevailed in the market, had there been no purge.

1.4.4 Counterfactual Analysis

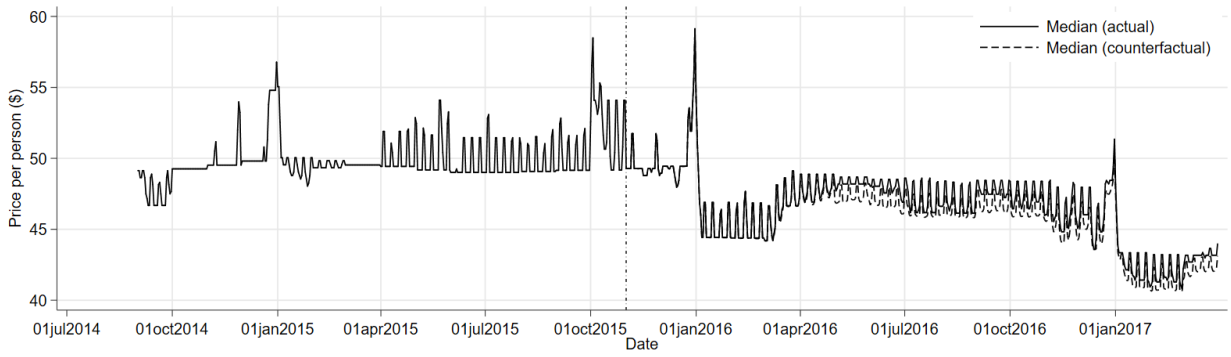
I recalibrate a counterfactual scenario without the purge and describe its effects on the welfare of consumers and rental owners. Each market/time in the counterfactual contains both purged and non-purged rentals competing with each other in a monopolistic competition setup. The counterfactual analysis makes a couple of assumption. I first assume that the price sensitivity parameter of the demand does not change post purge, thus preserving the shape of the demand curve. It implies that for a given quantity of accommodation, the own price elasticity remains the same in the counterfactual setting. The only change that happens to the demand curve post purge is in the form of shift such that the counterfactual price and quantity lie on the respective demand curves. Next, I assume that the marginal cost of any rental does not change as a result of the purge.

I proceed with the counterfactual analysis in the following steps. First, I recompute the consumer and producer surplus derived from rentals that were not purged under the set of new price and quantity computed using the estimates in section 1.4.3. This step enables us to measure the changes in welfare as a result of reduced competition in the market. Next, I shift focus to the purged rentals and approximate the demand and marginal cost of these

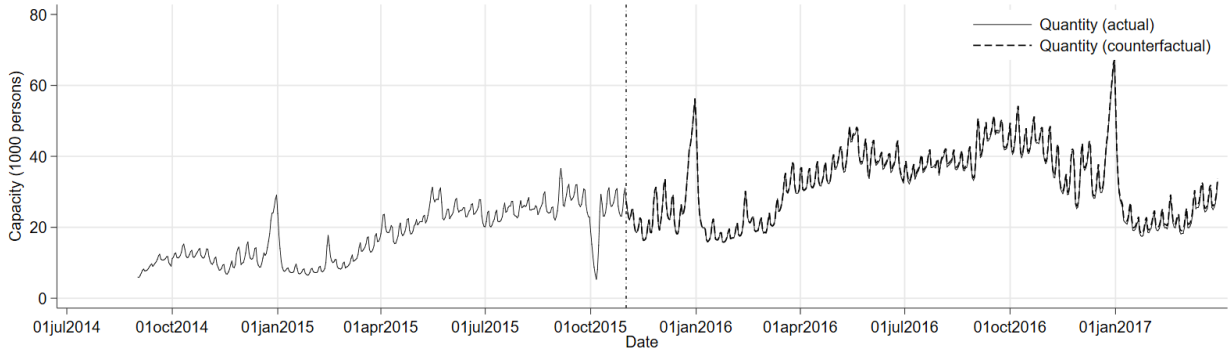
Table 1.11 – Changes in Consumer Welfare

	(1)	(2)	(3)	(4)	(5)	(6)
	Number of rentals		CS_{actual}	CS_{cf}	$\Delta_{price}CS$	$\Delta_{total}CS$
	All	Purged	(1000$)	(1000$)	(1000$)	(1000$)
All rentals	108,212	4,127	452,489	483,290	-1,127	-30,801
Space-Type						
Entire Home/Apartment	55,790	4,127	355,564	386,102	-863.1	-30,537
Private Room	48,183	0	89,374	89,622	-248.0	-248.0
Shared Room	4,207	0	7,550	7,565	-15.35	-15.35
Property-Type						
Apartment/Condominium/Loft/Townhouse	99,782	4,122	396,520	427,278	-1,084	-30,759
House	7,025	0	50,189	50,221	-32.60	-32.60
Bungalow/Castle/Guesthouse/Vacation home/Villa	145	0	408.2	408.6	-0.403	-0.403
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	89	0	232.1	232.4	-0.296	-0.296
Location						
Bronx	1,570	32	4,986	4,986	-0.195	-0.195
Brooklyn	39,527	867	158,775	162,341	-145.5	-3,566
Manhattan	56,977	2,997	249,939	276,404	-961.6	-26,465
Queens	9,396	212	35,071	35,669	-9.446	-597.8
Staten Island	733	19	3,707	3,878	-9.802	-171.4

Notes: The consumer surplus and changes have been aggregated across all the rental-time pair by each rental category.



(A) Prices



(B) Quantities

FIGURE 1.11 – COUNTERFACTUAL AGGREGATE TRENDS IN EQUILIBRIUM OUTCOMES
Notes: The counterfactual price-quantity has been calculated using by superimposing the estimated changes and spatial-temporal patterns on the actual price-quantity

rentals during the times/markets when they were forcibly kept out of the market. For each purged rental, I use the extrapolated demand and marginal cost to solve for the price and quantity that would if the given rental was not purged. Though, the price-quantity has been calculated for purged rentals during the time they stayed out of the market individually, notice that they do not adequately reflect the counterfactual considering all purged rental exist in the market. They merely are the outcomes that would have been observed for a given purged rental had it escaped the purge while all other rentals were eliminated as in the actual setting. To incorporate the existence of all purged rentals and obtain counterfactual price and quantity of purged rentals, I adjust these outcomes of purged rentals by exploiting the

Table 1.12 – Changes in Producer Welfare

	(1)	(2)	(3)	(4)	(5)	(6)
	Number of rentals		PS_{actual}	PS_{cf}	$\Delta_{price}PS$	$\Delta_{total}PS$
	All	Purged	(1000$)	(1000$)	(1000$)	(1000$)
All rentals	108,212	4,127	302,243	329,882	4,829	-27,639
Space-Type						
Entire Home/Apartment	55,790	4,127	243,718	272,217	3,970	-28,499
Private Room	48,183	0	53,261	52,458	803.0	803.0
Shared Room	4,207	0	5,263	5,207	56.66	56.66
Property-Type						
Apartment/Condominium/Loft/Townhouse	99,782	4,122	262,930	290,800	4,598	-27,870
House	7,025	0	35,471	35,282	189.4	189.4
Bungalow/Castle/Guesthouse/Vacation home/Villa	145	0	264.7	262.8	1.934	1.934
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	89	0	165.0	163.8	1.210	1.210
Location						
Bronx	1,570	32	3,862	3,859	2.679	2.679
Brooklyn	39,527	867	109,575	112,756	734.4	-3,181
Manhattan	56,977	2,997	161,072	184,830	4,003	-23,758
Queens	9,396	212	25,077	25,703	48.82	-625.6
Staten Island	733	19	2,648	2,725	40.16	-77.51

Notes: The table presents aggregate producer surplus and the changes calculated for different categories of rentals.

spatial and temporal patterns of all other purged rentals and the changes estimated in section 1.4.3. Using an approximate demand curve for each purged rental that lies on counterfactual price and quantity, I calculate the consumer surplus by integrating under the demand curve. The producer's profits are subsequently calculated by multiplying markup with quantity in the counterfactual state.

I begin by looking at how different are the counterfactual price and quantity from the actual. Since I have modeled equilibrium outcome changes to vary with distance from purged rentals, the counterfactual outcomes in a given market/time differ most for rentals that experienced the most substantial capacity loss in their vicinity. Figure 1.11 shows the aggregate trends in counterfactual prices and quantities over time in comparison to the actual. After November 2015, the counterfactual prices are lower than what we observe. These differences increase as the number of rentals exiting the market increases. The trends in

counterfactual quantities (1.11b) are slightly higher than the actual.

The consumer surplus in actual and counterfactual setting are summarized in table **??**. The total loss to consumers in about 17 months since the purge first began is calculated to be around $25M\$$. Most of the rentals (about 75%) have been purged from Manhattan, which is also one of the most sought after borough on Airbnb for its locational appeal. Thus, it is not surprising that a major portion of consumer welfare loss (about 83%) comes from this borough. Though reduced competition and changes in prices led to a decrease in consumer surplus by $2.3M\$$, the significant loss to their welfare ($22.93M\$$) came from the unavailability of products in the market. This is because if the purged rentals had not been purged, they would have continued to serve the demand of the consumers. The changes in price and quantity also imply the loss in surplus to spread across rentals which are not subject to regulatory restrictions. These include "Private Rooms" ($0.54M\$$) and "Shared Rooms" ($0.03M\$$) among different space type and all property type variants except "Apartment/Condominium/Loft/Townhouse".

Table **??** present the summary of welfare effects of the purge on the rental owners. The sellers whose rentals were not removed gained $9.45M\$$ as a result of the purge. This is because remaining rentals in the market were able to charge a higher markup upon the exit of their competitors. Most of these gains (about 88%) were accrued in the Manhattan borough which came at the expense of consumers and the rentals which were purged. Purged rentals within Manhattan lost a total of $28.47M\$$ (81% of total $34.95M$) as a result of losing sales.

1.5 Discussion and Conclusion

Table 1.13 summarizes welfare results. The welfare impacts of removing rentals from the market are significantly large ($55M\$$) both for consumers and sellers. Though the given case of purge is relatively much smaller than the actual regulations being considered in NYC, we can expect similar but larger impacts from such regulations.

In this paper, I have quantified the surplus generated in the Airbnb marketplace of NYC. The benefits of Airbnb as a marketplace is substantial for both rental owners as well as

Table 1.13 – Welfare Summary

	(1)	(2)	(3)
Surplus	Market Presence	Setting	Welfare ($1M$)
Counsumer	Existing	Actual	391.84
Counsumer	Existing	Counter-factual	394.17
Counsumer	Purged	Counter-factual	22.83
Producer	Existing	Actual	566.40
Producer	Existing	Counter-factual	556.95
Producer	Purged	Counter-factual	34.95

Notes: The Existing rentals are those which were not purged. These are aggregated figures summed up over all rental-time observation in each setting.

travelers, which explains the growth and popularity of Airbnb. This paper views Airbnb as a monopolistic-competitive market with minimal entry costs. The results are in line with this view. We observe that while overall sellers continue to increase in the market, the residual demand faced by every single rental is shrinking over time and so are the profits of the sellers.

A supply-side regulation in the form of bans, quotas or caps on the number of rentals or days of operation will raise the prices in the market above the competitive equilibrium levels, leading to a loss in the welfare. I demonstrate the impact of such supply restriction on the welfare using a series of small episodes of rental purge carried out by Airbnb on its NYC market. I find evidence that peer sellers though small and fragmented respond by raising prices. To an increase in price, the consumers react by decreasing the quantity of accommodation (measured as the number of persons). As a result of a loss in rental capacity in the market, there are changes in the welfare of both consumers and sellers through separate channels. The first channel through which consumer welfare changes is the increase in prices. However, while an increase in prices hurts consumers, it benefits the sellers that are allowed to stay in the market.

Airbnb markets supply a breadth of personalized selection of rentals to the consumers. Each unique rentals faces a residual demand when present in the market. The purged rentals would have had a demand in the markets/times when they were forcibly kept out of the

market. I quantify the loss to the consumers from the loss of selection by approximating the demand of purged rentals. Substantial losses are incurred from unsatisfied consumers demand as a section of rentals are unavailable for consumption. On the other hand, the purged rentals ended up losing their revenue in the market.

Findings in this paper suggest that market participants respond to changing supply conditions in the market. Merely a small scale purge carried out by Airbnb on its market was sufficient to induce prices to increase and significant losses in surplus. This paper focuses on a self-inflicted supply purge mainly because it mimics the very nature of proposed regulations (such as mandatory registration/approval requirements, quotas, and caps on the operation, and bans) not only in NYC but also in other Airbnb markets. This paper contributes to the sparse empirical literature on P2P markets and to ongoing policy debates surrounding such businesses primarily in the context of regulations that directly target the supply side.

This paper highlights the welfare losses induced by the disruption caused by a mere threat of regulation. We also find that such a supply restriction helps the sellers by removing the competition. One can expect similar and more significant impacts under actual regulations which restrict supply. These threats of regulation also infuse fear of wrongdoing among the hosts who may not enter the market at all. Though the evidence suggests that supply regulations reduce consumer welfare through a reduction in price competition and loss of selection, it also impacts market thickness, search and matching frictions, crucial from the market design perspective. While these issues open lines for further empirical research, the current research is informative of the ramifications of regulatory policy in this regard.

Chapter 2

ESTIMATING AIRBNB RENTAL DEMAND USING DOUBLE MACHINE LEARNING APPROACH

2.1 Introduction

This paper estimates the residual demand function faced by Airbnb rentals in the presence of high dimensional confounding demand signals. I refrain from subjectively selecting the variables entering the model and automate this process by using Machine Learning (ML) techniques. I apply the Double Machine Learning (DML) approach proposed by Chernozhukov et al. (2018) and Chernozhukov et al. (2017), to a massive data of Airbnb rentals in the New York City (NYC) and estimate the demand facing the rentals. The DML approach enables low-dimensional inference by sample splitting and constructing control functions of high dimensional covariates using ML methods. The method uses orthogonalization to alleviate regularization bias and sample splitting for guarding against over-fitting.

Consumers decide on rentals after carefully examining the rental page on the Airbnb website, where they view, gather, and process all the available information. This information includes rental characteristics such as space type, property type, the number of bedrooms and bathrooms, location, amenities, rental description, and images, among others. Consumers also view and evaluate each rental through reviews posted by the past guests on the rental web-page. The rich information on the rental web-pages tends to be very high-dimensional, and though available, it is often unknown to the researcher as to how and which information matters in the consumer's decision-making process. Moreover, rental owners understand the quality of their rental pages through the same visible collection of information and strategically set the nightly prices of their rentals. Thus, the high dimensional information, in turn, confounds the demand relationship between price and quantity.

The recovery of the demand parameter is frustrated by the presence of vast amounts of information available to the econometrician and needs to be controlled. A valid inference is difficult in the presence of high-dimensional controls, which affect both price and quantity as they lead to high variance on coefficients estimates. On the contrary, performing a naive inference while using ML methods for variable and model selection concurrently yields biased estimates. It is because of regularization, where penalty on the size of coefficients or restriction on the complexity of the model throws out variables that are weakly correlated to the outcome but may have a strong association with the variable of interest (Belloni, Chernozhukov, et al. (2013), Belloni, Chernozhukov, and Wei (2016), Belloni, Chernozhukov, and Hansen (2014)). Thus, the resulting model begins to violate the conditional independence assumption.

On the other hand, the subjective selection of variables or models is often prone to errors on the part of the researcher and thereby violating the conditional independence assumption again. In such situations, when it is not straightforward to assume strict exogeneity of price, it is hard to recover a valid demand parameter. The DML approach attempts to maintain conditional independence assumption while excluding covariates, which do not affect either price or quantity by constructing approximately orthogonal residuals.

In this paper, I exploit the richness of the Airbnb data gathered initially by scrapping rental web pages and cast them into a set of a high-dimensional feature vector. The high-dimensional set of controls arises from four distinct aspects of the Airbnb rental web-page, namely rental characteristics, amenities, description, and past customer reviews. The characteristic space is high-dimensional since there are many attributes that rentals use to distinguish themselves from other rentals such as space type, property type, neighborhood, bedrooms, and bathrooms, among others. Most of these attributes are categorical such as neighborhood, which increases the dimensionality by the total number of categories within each variable (one-hot encoding). Also, since there is no assumption on how and which attributes enter the specification, I add higher-order interaction terms of the base variables to the specification. These terms further increasing the dimensionality, particularly when the interaction of categorical variables is

being considered.

Subsequently, I add other pieces of information typically present in the form of text, which exhibits an intrinsic high-dimensional character. These include rental amenities and descriptions and past customer reviews. I try to capture rental level heterogeneity using the rich (though unstructured) text data on rental amenities and description, which is assumed to be time-invariant but vary across rentals. In addition, reviews collected by each rental on its web page generally increase on the website over time. This paper views customer reviews as a stock of reputation associated with a rental at any given point of time. I, therefore, cast reviews into high-dimensional quantitative features such that features reflect time-varying and qualitative aspects of the rental. The resulting data is at a rental-day level with about $25M$ observations and thousands of covariates/features, which confound the underlying demand relationship. In recovering this relationship, I face a 3-way trade-off between the size of the data, computational capacity, and time. I address these issues during the course of this paper.

I pose a Partially Linear Instrument Variable model for rental's residual demand where the price is endogenous. The high-dimensional confounders enter through a nuisance function, which, in turn, enters the specification linearly. Although, I have an extensively rich data on the determinants of demand and prices from the rental web-page, yet the data lacks many demand signals which are usually accessible to an industry researcher. For example, I do not have data on the amount of interest a particular rental receives, the information usually recorded in communication messages between the host and potential guests prior to finalizing a booking. Moreover, I do not have images of the rentals posted by their owners and used by consumers to evaluate the rentals. Lastly, I am restricted to equilibrium data where observed prices and quantity result from movements in both supply and demand. I, therefore, use instruments for price constructed from past hosting behavior of each rental and argue that they are good proxies of cost-shifters (see Section 1.3.1).

The DML approach counters the issue of high-dimensional controls derived from rich data available to us through public web pages of Airbnb rentals. It is a 2-step estimation procedure that requires sample-splitting and obtaining exogenous variation in the residuals

of price and other variables using out-of-sample predictions post a Machine Learning (ML) estimation. Doing so allows us to use simple Generalized Method of Moments (GMM) to estimate the low-dimensional demand relationship in the second step consistently.

In this paper, I implement the DML procedure by partialling out the high-dimensional control variables (orthogonalization) from each of the low-dimensional variables on split data. For each low-dimensional variable, including instruments, I then select the best ML model within each class of popular ML algorithms to obtain out-of-sample predictions and corresponding residuals. The ML methods, I consider, include Penalized Regression, Random Forest, Boosting, Neural Networks, and Stacked Ensembles. The final choice of the class of ML algorithms depends on the assumed structure of the nuisance function for each variable in the low-dimensional regime. In practice, the selection of the exact ML model is left to the out-of-sample performance of all the different models. I pick the model with the lowest Mean Squared Error (MSE) for each variable in the low dimension. In order to estimate the demand relationship in the second stage of the DML procedure, I use GMM on the orthogonal residuals of the auxiliary sample not used for ML training.

As the size of the data is too big for our current computation resources, I restrict the full sample ML exercises to Regularized Regression methods such as LASSO and Elastic-Net. These models are comparatively easier to optimize and, therefore, faster to train. The DML procedure necessitates excellent predictive performance to allow valid inference. Thus, while setting up the DML algorithm, I make several design choices to enable reliable predictions. First, I cast the unstructured text data into a set of quantitative independent variables using frequency-based methods. I describe several choices made to handle unstructured text data in detail in Section 2.2.1. Second, I experiment with multiple classes of ML functions used in the first stage of the DML procedure, including the ensembles method (Appendix A). Third, I use a data-driven approach to find a set of hyperparameters for each class of ML models that perform best in terms of out-of-sample predictions.

The findings in this paper show a considerable improvement on the previous estimates of demand which completely ignored to control for the time-varying reviews. The demand

estimate jumped to −14.364 with the standard error of 1.283 when we included a host of features constructed from reviews text. The results of the full sample show stability across the choice of different Penalized Regression models to estimate the nuisance function in the first stage. The findings also indicate that discovering a good model fit in the first stage is crucial to the validity of the final DML estimates. The paper also provides a small sample result of DML exercise to experiment with more complex ML methods such as Random Forest, Boosting, and Neural Network in Appendix B. Though not very representative, the results of this experiment re-establish the importance of the excellent performance of ML estimation in the first stage.

Related Literature : This paper is complementary to the literature on theoretical advances in econometrics of high-dimensional data. For more details see Varian (2014), Mullainathan and Spiess (2017), Athey and Imbens (2017), Athey (2018), and Athey and Imbens (2019). Research on Machine Learning in Econometrics has taken different routes in terms of their approach. For example, Machine Learning has been used in estimating heterogeneous treatment effects using causal trees (Athey and Imbens (2016)) and causal forest (Athey, Tibshirani, Wager, et al. (2019)). These methods first learn the patterns of heterogeneity using observed variables without assuming any specific functional form on heterogeneity on one partition of the data and then use groups and subgroups from the first stage to identify the treatment effect within these groups.

A separate stream of research deals with the inference of parameters in the low-dimension in the presence of high-dimensional confounders using data-driven methods. Though these methods are excellent for prediction exercises, especially when the data is high-dimensional, they are unsuitable for inference of model parameters that we are mostly interested in as economists (Leeb, Pötscher, et al. (2006) and Leeb and Pötscher (2008)). This line of theoretical work exploits multi-equation orthogonalization concept where they study the asymptotic performance of Post-Selection Least-Squares estimators (Belloni, Chernozhukov, et al. (2013), Belloni, Chernozhukov, and Hansen (2014)), and Generalized Linear models (Belloni, Chernozhukov, and Wei (2016)). These methods involve the selection of the model

using Penalized Regression to obtain an approximate estimate of the nuisance part of the main estimating equation with high-dimensional controls. These methods enable obtaining valid post model selection inference and stress how estimators fail to perform valid inference when the estimation of the nuisance part and the main parameter of interest is naively carried out using ML methods in a single step.

Most of the empirical work in economics that uses Machine Learning is limited to prediction exercises (see Donaldson and Storeygard (2016), Henderson, Storeygard, and Weil (2012), Lobell (2013), Blumenstock, Cadamuro, and On (2015), Glaeser et al. (2018)). This paper extends the empirical work involving Machine Learning beyond prediction. We apply the general methods developed in Chernozhukov et al. (2018) and Chernozhukov et al. (2017), which rely on the "Neyman Orthogonality" property. The authors argue that this property enables recovery of the valid parameters by approximately estimating the nuisance function. The methodology allows a data-driven approach to selecting the variables and models while still enabling the inference of parameters of interest post model selection. With the exception of Knaus (2018), the empirical literature is quite limited in the application of this approach. In their paper, Knaus (2018) use the DML to estimate the effect of playing music on 9^{th}-grader's cognitive and non-cognitive skills, where high-dimensional confounding factors are socio-economic backgrounds of the students. In estimating the firm-facing residual demand function, the current paper uses a much larger sample and set of confounding factors derived from higher-order terms and interactions of categorical variables and inherently high-dimensional unstructured text data.

The methods this paper uses for prediction and model selection in the first stage of DML procedure come from a wider Machine Learning literature (see Hastie et al. (2005) James et al. (2013), Friedman, Hastie, and Tibshirani 2001 and Hastie, Tibshirani, and Wainwright (2015)). These ML algorithms are desirable in the sparse settings where only a small number of variables are important, while most variables have little or no effect on the target variable in the low dimension. While these methods are excellent at prediction, we can not draw inference naively from these methods.

This paper further adds to the econometrics literature that uses unstructured text as data (see Gentzkow, Kelly, and Taddy (2017) for a comprehensive overview). Most of the existing studies using text as data are also limited to using ML methods for prediction. For example, Hoberg and Phillips (2016) uses the text of business description in 10-K filings to flexibly classify firms over time, which allows them to study competition in the product space. Other papers in Finance use text in 10-K filings (Kogan et al. (2009)), text from Wall Street Journal (Tetlock (2007)), and text from online message boards such as Yahoo Finance (Antweiler and Frank (2004)) for prediction of stock market outcomes. In IO and marketing, Kang et al. (2013) builds a classification model using Support Vector Machines to predict offenders using data on hygiene inspection outcomes and the customer reviews of restaurants from Yelp.

Inference using the text data in the literature too has the flavor of prediction, but approaches have been usually subjective. Recent work by Stephens-Davidowitz (2014) use racially charged text searched on Google to create an index of racial animus and subsequently estimates the loss of vote share caused by racism. Gentzkow and Shapiro (2010) refrain from the subjective selection of text to predict media slant. Instead, they use all words, 2-word, 3-word phrases of newspapers and compare them to Congressional speeches. The predicted index is then used to characterize the demand and supply of newspapers that contain these terms. In the current paper, I combine the use of text from reviews, rental amenities, and descriptions to obtain excellent predictive models for price and quantity. Subsequently, I use them to obtain high-quality point estimates of demand parameters in an entirely data-driven manner.

The rest of the paper is organized as follows: Section 2.2 describes the data and its construction. I explain the high-dimensional nature of the data built from higher-order interactions of categorical variables and unstructured text data. Section 2.3 outlines the econometric setup, the DML estimation algorithm, and the first stage ML methods used in the procedure. The next section (2.4) presents the results in which, I provide full sample DML results using Penalized Regression methods such as Lasso and Elastic-Net. I also report

the results from a small sample exercise that uses more complicated ML methods such as Random Forest, Boosting, and Neural Network in Appendix B. Next, in this section, I use the results of demand estimation and recompute the consumer and producer surplus. Section 2.5 concludes the paper.

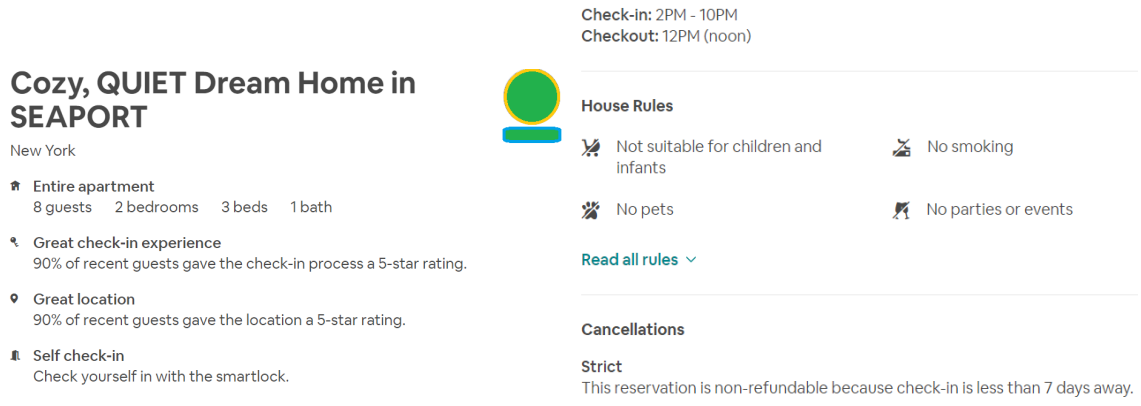
2.2 Data

The data central to this paper is a daily rental data on bookings and prices. This data contains sales and pricing information of more than 100 thousand rentals that have appeared on Airbnb market of NYC during the period from September 2014 through March 2017. The price-sales data comes from a private data-analytic company Airdna, that scrapes publicly available information on Airbnb rentals including bookings and availability calendar and rental characteristics from Airbnb website. I obtain text data on amenities, description and reviews data from Inside-Airbnb, an open-source website which also scrapes data from Airbnb website.

The low dimensional demand relationship between price and bookings is the main parameter of interest. However, this parameter is confounded by a host of factors such as rental characteristics, amenities, the information enclosed in the rental description, images, and reviews. The vast variety of information is presented on the rental web-page to facilitate consumers in deciding whether or not to book a rental. However, it is usually unknown to the researcher which information from a complete information set affects the quantity demanded. The rental owners also account for these signals while setting the prices, and thus, identification of demand parameter may be confounded by these factors and needs to be controlled.

I begin by combining daily data on prices and bookings to the rental characteristics data[1]. These characteristics include the type of space (Entire Home/Apartment, Private

[1]Price-Booking data has been obtained from a private provider AIRDNA. They have also provided data on many characteristics of the rentals such as approximate geographic location, number of room, bathrooms, listing type (Entire home or apartment, Private Room, or Shared Room), Property Type (apartment, condominium, townhouse, house, bungalow, boat among others)



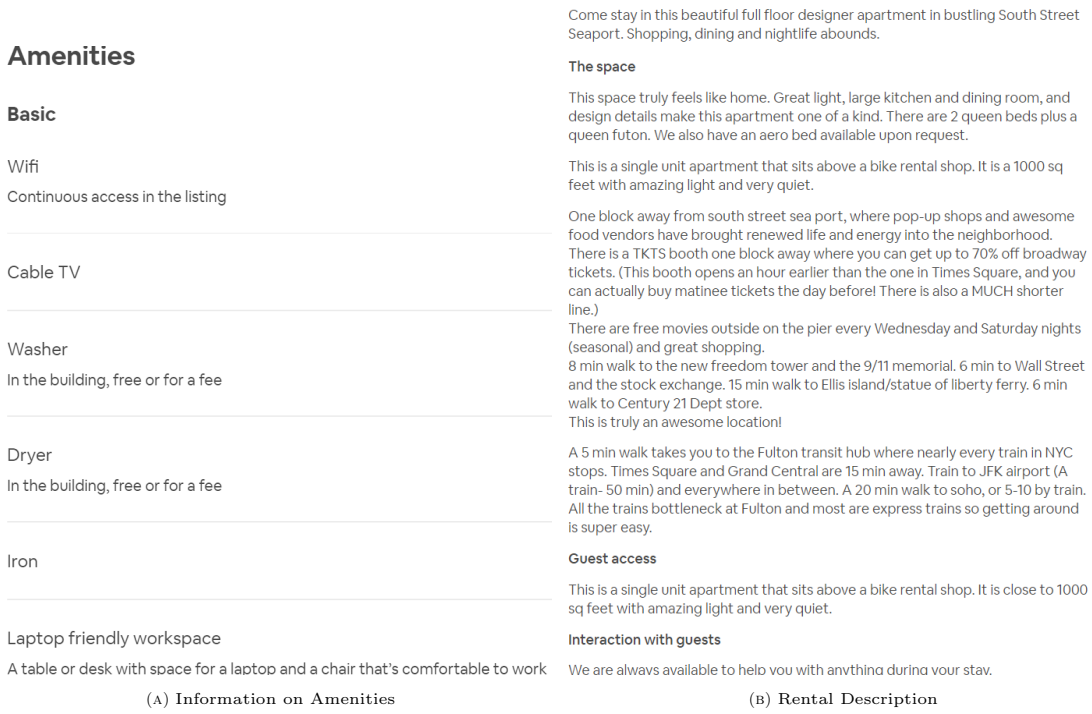
(A) Primary Information (B) Other Information

FIGURE 2.1 – RENTAL CHARACTERISTICS AS DISPLAYED ON AIRBNB WEB-PAGE

Room, Shared Room), property type (such as a house, bungalow, villa, apartment, cabin and boat among others), number of bedrooms and bathrooms, location, neighborhood, borough, and several others. Figure 2.1 shows the characteristics as they are presented on the rental web-page of Airbnb. Notice that most of these characteristics are categorical. For example, there are 40 different types of properties and about 200 different neighborhoods in the NYC market. Such categorical variables contribute to the creation of dummy variables which is equal to the number of categories in each category type minus one. I also include seasonal variables as categorical variables such as year (4 years), month (12 months), week (52 weeks), day of the week (7 days). They further add to the dimensionality of the data.

Since I do not know the functional form through which these variables enter the model, I consider using higher-order terms and multi-way interactions of the above set of variables. The interactions further multiply the number of covariates entering the specification. For instance, if I want to capture the changing patterns across neighborhood over time and want to include a 3-way interaction of neighborhood, year and week, I end up with about 30 thousand more variables ($= (192 - 1) \times (4 - 1) \times (52 - 1)$).

Next, I augment the price-sales data to the pre-processed text data of rental description



(A) Information on Amenities (B) Rental Description

FIGURE 2.2 – AMENITIES AND DESCRIPTION

and amenities[2]. In the following section, I describe the text data in detail, the steps to convert the text into data and elaborate on the high dimensional nature of text data. The rental characteristics, amenities, and description do not vary over time. This allows us to capture the rental level heterogeneity high dimensional set of covariates. Figure 2.2 depicts an example of a web-page advertising the amenities and describing the rental. There are many different types of amenities presented to the consumers, some more important than the others for decision making. Since I do not assume which variables enter the demand model, I refrain from hand-picking the amenities and converting them into dummy variables. Instead, more generally, I consider all the text pertaining to amenities and create features directly from the text. Finally, I augment the resulting data with the pre-processed textual reviews

[2]This data is obtained from Inside-Airbnb, an independent and open source data provider. They compile publicly available information on Airbnb, including reviews and rental description.

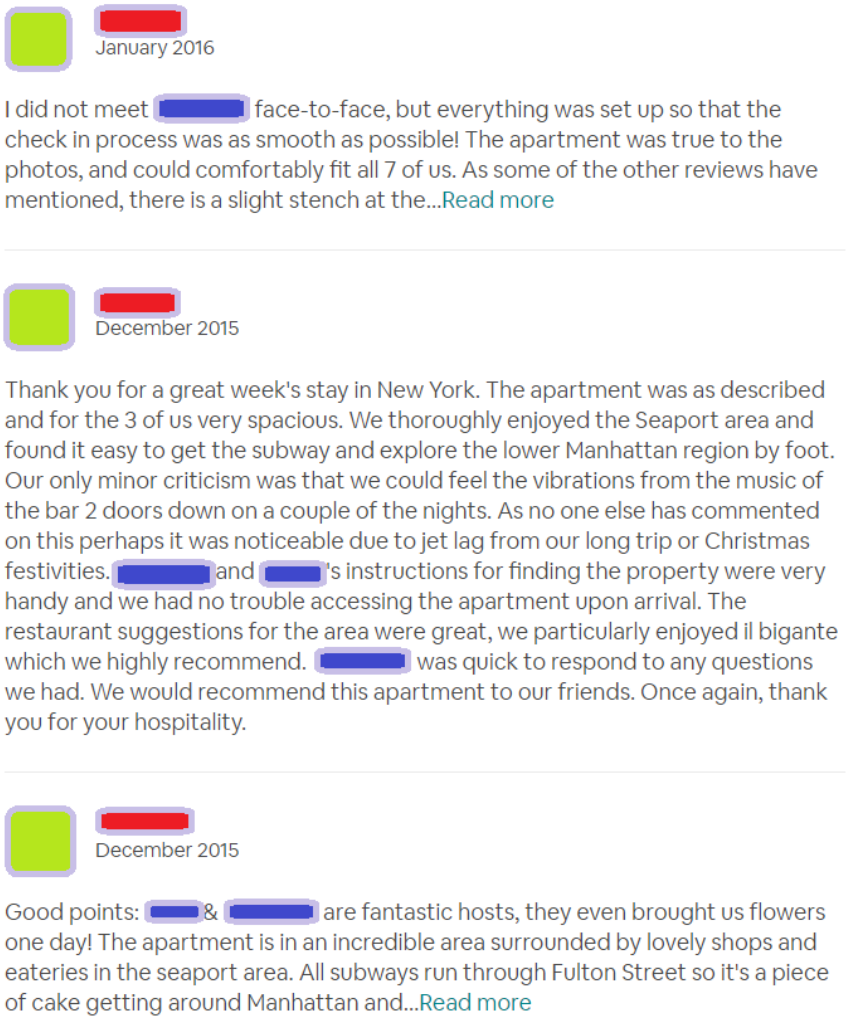


Figure 2.3 – Reviews as they appear on the Airbnb Rental page

data such that it captures time-varying demand signals for each rental.

2.2.1 High Dimensional Controls from Text

Human understanding of a raw text document (a single piece of written text such as a multi-sentence text describing a rental or a review posted by a past consumer) requires grammatical structure of the language and the context of related text in the same document

and others in the corpus (collection/sample of documents such as all the reviews of rentals in NYC from the beginning of time to the last data collection date). For example, a review is understood and evaluated by a consumer based on nouns, verbs, adjectives, and adverbs and how they are used with one another within the same review. They are simultaneously evaluated in contrast with other reviews by comparing usage of vocabulary and grammar across other reviews. Besides, reviews are often perceived differently by different readers.

In order to enable a computer to understand the text, one would require some text representation that provides the machine all the rules of grammar, vocabulary, and the context. Naturally, such representation of text as data can be very high dimensional for any modern-day computational capacity. The text data, therefore, usually requires pre-processing before it can be formally used in the analysis. The pre-processing step is a way to extract a significant portion of the information contained in the text useful for prediction of different outcomes. This process involves the creation of meaningful features which appropriately represents the text and context in which it has been used. I largely follow the steps detailed in Gentzkow, Kelly, and Taddy (2017).

In this paper, I work with more straightforward dictionary-based representation where I primarily use the count of words (uni-gram) and adjacent words (bi-gram) to construct the features. The dimensions of the data in a simple dictionary-based representation increase with the number of words in the vocabulary needed to represent the text. For example, consider a vocabulary with words - “Good”, “Bad” and “Ugly” - only and a representation where I only count words and two-word phrases. Given such vocabulary, a simple representation like counting uni-gram can be expressed in 3 dimensions. Similarly, a vocabulary with just 100 words would require 100 dimensions to be represented and so on as I increase the size of vocabulary of the used text. Moreover, the dimensions of quantitative data increases exponentially when multi-word phrases are considered.

Keeping the exploding dimensionality of text data in mind, I now describe the steps taken to convert human-readable raw text data (description, amenities, and reviews) into a manageable machine-readable quantitative data. For the most part, the raw text data on

description, amenities, and reviews are treated similarly. However, I slightly differ in my approach towards reviews data because unlike reviews, description, and amenities vary at the rental level and do not have time-varying component. In fact, I view customer reviews a rental accumulates on its Airbnb web-page by a given time as a stock of the rental reputation. Therefore, I take a few extra steps on reviews text data so that the quantitative features constructed from this data reflects properties of stock variables.

I begin by adding some structure to otherwise unstructured text data. For each of the text data - description, amenities, and reviews - I first define a boundary of the document. I define a document at the rental level for description and amenities data. It is because I assume that rental description and amenities do not change over time. So in Figure 2.2b, the rental description defines a document for this data. I similarly proceed with the amenities data where all the written text on amenities (Figure 2.2a) for a rental is treated as one document.

On the other hand, the reviewer and the date when a review was posted defines a document boundary for reviews data. Figure 2.3 shows an example of presentation of reviews on the website, which can also be viewed as three individual documents of reviews corpus. Alternatively, I could have defined a document as an aggregation of all reviews a rental received on a given date (see Gentzkow, Shapiro, and Taddy (2016) for other examples). I choose former definition over latter because I am interested in the frequency of phrases used in a document and the fact that I am going to aggregate the frequency to match the data at the rental day level, choosing one over the other does not make much conceptual difference. In fact, this is a design choice we must make because smaller sized documents are computationally much faster. It is a critical consideration given that the number of documents is enormous (more than $1M$).

After defining the document and its corpus, I perform preliminary text mining and dimension reduction steps. I start by tokenizing the documents. In this step, a document is broken into its elements separated by delimiters. Table 2.1 provides an example of the text pre-processing treatment of a single document from the reviews corpus. Despite being necessary for the grammatical structure of the text, punctuation and symbols do not carry

Table 2.1 – Text Preprocessing

Steps	**Example**
Raw Text Document	"Pretty, big and clean department 15 minutes away from the port authority bus terminal, safe neighborhood, buses available until 1 AM. You have groceries stores and restaurants just two blocks away."
Tokens	"Pretty", ",", "big", "and", "clean", "department", "15", "minutes", "away", "from", "the", "port", "authority", "bus", "terminal", ",", "safe", "neighborhood", ",", "buses", "available", "until", "1", "AM", ".", "You", "have", "groceries", "stores", "and", "restaurants", "just", "two", "blocks", "away", "."
Remove punctuation and symbols (Keep numbers)	"Pretty", "big", "and", "clean", "department", "15", "minutes", "away", "from", "the", "port", "authority", "bus", "terminal", "safe", "neighborhood", "buses", "available", "until", "1", "AM", "You", "have", "groceries", "stores", "and", "restaurants", "just", "two", "blocks", "away"
Lower case, remove stop-words and stemming	"pretty", "big", "clean", "department", "15", "minut", "away", "port", "authority", "bus", "terminal", "saf", "neighborhood", "bus", "availabl", "1", "groceri", "stor", "restaurant", "just", "two", "block", "away"
Uni-grams, Bi-grams and Bi-grams with skips	"pretty", "big", "clean", "department", "15", "minut", "away", "port", "authority", "bus", "terminal", "saf", "neighborhood", "bus", "availabl", "1", "groceri", "stor", "restaurant", "just", "two", "block", "away", "pretty_big", "pretty_clean", "big_clean", "big_department", "clean_department", "clean_15", "department_15", "department_minut", "15_minut", "15_away", "minut_away", "minut_port", "away_port", "away_authority", "port_authority", "port_bus", "authority_bus", "authority_terminal", "bus_terminal", "bus_saf", "terminal_saf", "terminal_neighborhood", "saf_neighborhood", "saf_bus", "neighborhood_bus", "neighborhood_availabl", "bus_availabl", "bus_1", "availabl_1", "availabl_groceri", "1_groceri", "1_stor", "groceri_stor", "groceri_restaurant", "stor_restaurant", "stor_just", "restaurant_just", "restaurant_two", "just_two", "just_block", "two_block", "two_away", "block_away"

Notes : The example presented above is of a single document (a review) from Reviews data. We perform the same treatment on the other documents in reviews as well as description and amenities data in order to convert them into informative features.

much meaning on their own. So, we remove all the punctuation and symbols in the document. Next, I convert all the tokens into the lower case and remove all the stop words present in the text. Though stop-words (articles such as "a" "an", and "the", conjunctions such as "and", "but", and "or") are often used in grammatical structuring of the text, they, like punctuation and symbols, also do not provide meaningful signals. With the remainder of the tokens, I perform word-stemming, which converts the words into their roots. For example, the words "picture", "pictures" and "pictorial" are reduced to their root "pict". All these steps help us reduce the dimensionality of resulting quantitative data by removing noisy features and preserving the ones that contain more signal for prediction.

Single words, such as in "Bag of Words" model, completely ignore the order of word occurrence in a document and thus each word as a feature is independent of the other. To

Table 2.2 – Document Frequency Matrices

Corpus	Number of Documents	Before Trimming		After Trimming	
		Number of Features	Sparsity	Number of Features (k%)	Sparsity
Amenities	108,212	1,591	97.2%	696(0.1%)	93.6%
Rental Description	108,212	3,612,269	> 99.99%	1,907(1.5%)	94.7%
Reviews	1,191,481	11,051,502	> 99.99%	2,529(0.3%)	98.7%

Notes : k denotes the minimum percentage values of a feature's document frequency, below which features have been trimmed out.

engineer better features and to allow some dependence between individual words, I turn to 2-word phrases in addition to single words. I include 2-word phrases of adjacently occurring words (see Gentzkow and Shapiro (2010)) and word pairs occurring on either side of a single word. For example, the tokens "carbon", "monoxide" and "detector" occurring in this order in a given text will be tokenized into "carbon", "monoxide" and "detector" as singular worded features. In addition, the 2-word phrases generated will be "carbon monoxide", "monoxide detector" (bi-gram without skip), and "carbon detector" (bi-gram with skip). Also, notice that, I do not remove numbers from the tokens because I use them to create some other useful features. These include phrases like "10 minutes", "2 days" or "1 block" which convey some sense of time and distance to the reader.

Though 2-word phrases as features provide us a high richness of data, they exponentially increase the dimensionality of the data compared to single-words. Allowing for an even higher dependence among words may not be suitable or even feasible given the computational burden it adds to an already cumbersome problem. It is important to note that the representation, I use in this paper, leaves out a considerable share of encoded information of these texts, and one can use richer representation (such as word-to-vec). It is yet another design choice to circumvent the computational burden and time required by richer representation, especially when the expected marginal gains in the accuracy of the results are small.

After creating one-word and 2-word tokens post-cleaning and dimension reduction steps, I convert these tokens into Document Feature Matrix (DFM) (transpose of Term Document

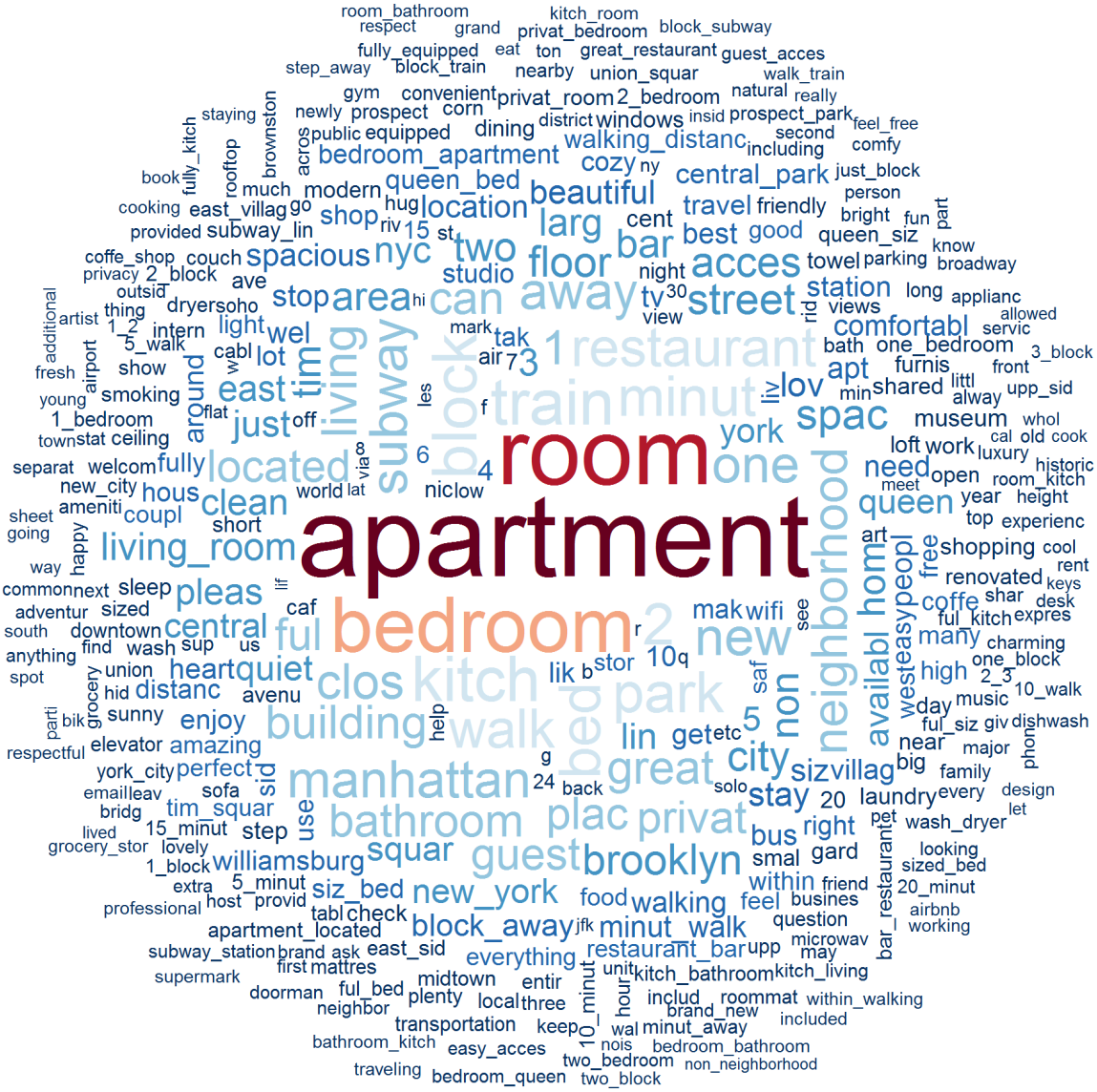


Figure 2.4 – Most Frequent Terms in Rental Description.

Notes : These 413 terms/tokens appear in atleast 6% of the documents of rental description corpus and 85.4% sparse. The size of each word or phrase in the word cloud is proportional to the frequency of its usage in the given corpus.

(A) Last set of included features

(B) Excluded features

FIGURE 2.5 – LESS FREQUENT TERMS IN RENTAL DESCRIPTION

Notes : In panel (*a*) The 272 terms/tokens appear in atleast 0.5% and and atmost 0.52% of the documents of rental description corpus and are 99.5% sparse. In panel (*b*) The 267 terms/tokens appear in atleast 0.2% and and atmost 0.202% of the documents of rental description corpus and are 99.8% sparse. The size of each word or phrase in the word cloud is proportional to the frequency of its usage in the given corpus.

Frequency Matrix). I denote a DFM by $\mathbf{C}$ where each element of matrix c_{ij} represents the number of times the token/feature j occurs in document i. I create one DFM for each of the textual data-sets i.e., Amenities, Description, and Reviews. Each of these matrices differs in the number of documents and features created after tokenizing and cleaning. The description and amenities have 108 thousand documents at the rental level while reviews have more than a million documents for each review posted since the beginning of time. Table 2.2 presents a summary of the each corpus. Notice that all three DFMs are high dimensional and sparse, especially the rental description and customer reviews corpus, despite cleaning and pre-processing of the text.

The use of all the features in the DFM will be a mammoth computational exercise, and incremental gains in accuracy and precision might be little. Thus, I perform feature selection exercises to obtain a comprehensive set of features that effectively capture the information within the text. I first begin by noticing that the features in all three DFMs are created from words and phrases, many of which are either very frequent or extremely rare in the corpus. Most common words (other than stop-words) are often related to the context of the discourse and may not be very informative to affect the price or demand. For example, the three most common words in the rental description are "apartment", "room", and "bedroom". Although, these words are necessary for advertising space rental has to offer, yet they add no value to a consumer's perception and therefore may not influence the demand. Figure 2.4 shows top 413 words by frequency used in rental description. These words appear in at least 6% of the documents in the description corpus. Generally, the more frequent words carry less qualitative information relative to less frequent words. This is also true for customer reviews corpus where, with exception of word "great", most common words are relatively less informative about the rental (Figure 2.6) such as "stay", "place", "apartment" or "host".

Though uncommon words do carry informational value, keeping them as features is a computational nuisance. To control the rapid growth of data, I use a popular strategy of trimming down the least frequent words in each of the corpora. In particular, for an arbitrary percentage k, I throw away all the features that occur in less than $k\%$ of the documents

Figure 2.6 – Most Frequent Terms in Reviews

Notes : These 443 terms/tokens appear in atleast 1.5% of the documents of rental description corpus and 95.3% sparse. The size of each word or phrase in the word-cloud is proportional to the frequency of its usage in the given corpus.

(A) Last set of included features

(B) Excluded features

FIGURE 2.7 – LESS FREQUENT TERMS IN REVIEWS

Notes : In panel (*a*) The 243 terms/tokens appear in atleast 0.5% and and atmost 0.6% of the documents of reviews corpus and are 99.5% sparse. In panel (*b*) The 214 terms/tokens appear in atleast 0.2% and and atmost 0.21% of the documents of reviews corpus and are 99.8% sparse. The size of each word or phrase in the word-cloud is proportional to the frequency of its usage in the given corpus.

of each corpus. I pick the value of k to be 0.1 for amenities corpus and 0.5 for the reviews and rental description corpus. The choice of k for each of the data set is subject to the computer memory and storage available since these features are converted to a data which require large amounts of computer memory while running the models. It is again a design choice that allows me to keep the problem computationally manageable by throwing away features, many of which could be potentially important. Nevertheless, I am still able to retain thousands of features which capture most of the salient features in each of the corpora. Besides, it is also possible that combinations of several excluded features could be highly correlated with those that are included. For example, an included uni-gram "laundry" could be correlated with some linear combination of bi-grams which include "laundry" such as "laundry rooftop", "laundry access", "laundry grocery", among others (Figure 2.5b and 2.7b). Besides, pooling more textual features may not provide any valuable gains in the precision of the final estimates.

I perform another simple step of information retrieval on the features based on the usage of words and phrases across the corpus. Instead of merely counting the number times a word or phrase j has appeared in a document i, I re-weigh the frequency (c_{ij}) by the number of documents (in the corpus) in which the term j appears. If a particular term appears in all the documents, then it provides no informational gain and therefore should not be weighed at all. For example, the terms "United States" or "America" are used in almost all the US presidential speeches. While in this particular example, "United States" is the subject of the discourse, it provides no qualitative information about the subject. Similarly, if two words/phrases occur the same number of times in a corpus, say the words "great" and "apartment" appear in the review corpus (Figure 2.6) and that "apartment" appears in more documents than "great" since "apartment" is the subject of the discourse, then "apartment" as a feature should carry a smaller weight than "great". This is because despite being frequently used when words such as "apartment" in reviews corpus appear in many documents, they provide small informational gain compared to the word "great".

I accomplish the re-weighing of the DFM, $\mathbf{C}$, using Term Frequency - Inverse Document

Frequency (TF-IDF), denoted by $d_{ij} \in \mathbf{D}$. If n is the total number of documents in the corpus and δ_j is the count of documents in which the term j appears atleast once i.e. $\delta_j = \sum_i \mathbb{1}_{[c_{ij}>0]}$, then I transform DFM to TF-IDF matrix $\mathbf{D}$ where each element of the matrix $d_{ij} = c_{ij} ln(n/\delta_j)$ for each document i and term j.

Once I have created TF-IDF matrices for all three pieces of text data along with the features, I combine them to the rental time panel. Since I assume that rental description and amenities do not change over time, I can directly combine them to the panel using rental ID as the key. However, since this paper treats reviews as the stock of reputation, I transform the TF-IDF matrix of reviews corpus such that every feature of reviews data is accumulated by each day. To achieve this, I cumulatively add the values of TF-IDF matrix from the beginning of time up to a given date within each rental. Thus, I preserve the independence of each document while creating features while still capturing the time-varying nature of covariates from reviews feature space.

The final data I use in DML estimation of demand parameter comprises of about $25M$ observations and thousands of confounding covariates including higher-order interacting categorical variables, features from text data of amenities, description, and reviews.

2.3 Methodology

I follow the DML procedure proposed by Chernozhukov et al. (2018) and Chernozhukov et al. (2017). I begin by setting up a Partially Linear econometric model of residual demand faced by Airbnb rentals where a nuisance function of high-dimensional confounding variables enter the model linearly along with logarithm of price. Next, I outline the computational algorithm that I code to implement the DML procedure. Finally, I describe the ML methods used to approximately estimate the high dimensional nuisance functions and predict the cross-fitted residuals.

2.3.1 Econometric Model

I consider a simple firm facing residual demand where high-dimensional confounding factors may enter the model through a non-linear function in a partially linear setup. The demand for rental j, q_{jt} (measured in terms of number of persons a rental accommodates), is given by :

$$q_{jt} = \alpha_1 ln(p_{jt}) + g_{j0}(\mathbf{X}_{jt}) + u_{jt}, \mathbb{E}(u_{jt}|\mathbf{X}_{jt}, \mathbf{z}_{jt}) = 0 \tag{2.1}$$

where p_{jt} is the price consumer pays for one individual's stay per night and $\mathbf{X}_{jt} = \{\mathbf{x}_{jt}, \Phi_t\}$. $\mathbf{x}_{jt}$ is a high-dimensional vector of controls and Φ_t denotes information set upto time t. Following the lines of Chernozhukov et al. (2017), I approximate Φ_t using past 3 lags of price and quantity. Since many rentals operate in the market for short span of time, taking several lags of price and quantity will lead to substantial loss of data. As described in Section 2.2.1, the remaining high-dimensional controls are derived from the information on the web-page including rental characteristics, amenities description and past customer reviews. $\mathbf{X}_{jt}$ enters the model through a complex (nuisance) function g_{j0}. The main parameter of interest is α_1.

In order to ease the computational burden, I limit the above specification 2.1 to own price effect and ignore cross-product price effect. Though the vast number of rentals in Airbnb NYC market can provide a potential justification for assuming insignificant cross rental substitution, this is still a strong assumption. This assumption can be relaxed as permitted by computational capacity and is open for future exploration.

I depart from Chernozhukov et al. (2017) in that I do not assume exogeneity of price conditional on available demand signals alone. Instead, I assume sequential exogeneity conditional on high-dimensional controls and the instruments $\mathbf{z}_{jt}$. This is because despite observing a huge set of factors that affect both price and demand, I do not observe all of them. These variables are typically observed by an Airbnb host and are accessible to an industry econometrician. For instance, I do not observe the amount of interest a rental receives before the actual booking. Usually, such information is observed through inquiries and conversations between a potential guest and the host prior to the transactions. I also do not observe all the

cultural, professionals, sports, or academic events occurring locally around the rental. Thus, it is reasonable to assume the prices to be endogenous because of unobserved factors that influence demand and instruments them using the past hosting behavior. For sufficing the exclusion restriction, I argue that the instruments, past decision to list a rental on Airbnb, are a good proxy for cost-shifters as they reflect the contemporaneous opportunity cost of renting space on Airbnb. Specifically, $\mathbf{z}_{jt} = [z_{jt}^1 \dots z_{jt}^B]'$ is a vectors of B instruments for price such that :

$$\mathbf{z}_{jt} = \mathbf{m}_{j0}(\mathbf{X}_{jt}) + \mathbf{v}_{jt}, \mathbb{E}(\mathbf{v}_{jt}|\mathbf{X}_{jt}) = 0 \tag{2.2}$$

where $\mathbf{m}_{j0} = [m_{j0}^1 \dots m_{j0}^B]'$ denotes functions through which high-dimensional controls enter the instrumenting equation and error $\mathbf{v}_{jt} = [v_{jt}^1 \dots v_{jt}^B]'$ such that conditional on $\mathbf{x}_{jt}$, instruments $\mathbf{z}_{jt}$ are exogenous.

Although, equation (2.1) and (2.2) define the system of econometric equations completely, it is useful to express the low-dimensional variables (quantity, price and instruments) as a function of high-dimensional variables (the reduced form) :

$$q_{jt} = l_{j0}(\mathbf{X}_{jt}) + \xi_{jt}, \mathbb{E}(\xi_{jt}|\mathbf{X}_{jt}) = 0 \tag{2.3}$$

$$ln(p_{jt}) = r_{j0}(\mathbf{X}_{jt}) + \zeta_{jt}, \mathbb{E}(\zeta_{jt}|\mathbf{X}_{jt}) = 0 \tag{2.4}$$

In equation (2.3) and (2.4), l_{j0} and r_{j0} represent the nuisance function that dictate how $\mathbf{X}_{jt}$ affects demand and price variable. The equations (2.3) and (2.4) combined with (2.2) complete the reduced form system of econometric equations. The zero expected conditional mean assumption on ξ_{jt}, ζ_{jt} and $\mathbf{v}_{jt}$ allows us to express the residuals of each variable in the LD regime as:

$$\begin{aligned}
\widetilde{q_{jt}} \quad &= q_{jt} - \mathbb{E}(q_{jt}|\mathbf{X}_{jt}) \quad &= q_{jt} - l_{j0}(\mathbf{X}_{jt}) \\
\widetilde{ln(p_{jt})} \quad &= ln(p_{jt}) - \mathbb{E}(ln(p_{jt})|\mathbf{X}_{jt}) \quad &= ln(p_{jt}) - r_{j0}(\mathbf{X}_{jt}) \\
\widetilde{\mathbf{z}_{jt}} \quad &= \mathbf{z}_{jt} - \mathbb{E}(\mathbf{z}_{jt}|\mathbf{X}_{jt}) \quad &= \mathbf{z}_{jt} - \mathbf{m}_{j0}(\mathbf{X}_{jt})
\end{aligned} \tag{2.5}$$

If nuisance functions l_{j0}, r_{j0}, and m_{j0} were known, the problem would have reduced to a simple GMM estimation where the moment condition is :

$$\mathbb{E}[\widetilde{\mathbf{z}_{jt}}(\widetilde{q_{jt}} - \alpha_1 \widetilde{ln(p_{jt})})] = 0$$

However, since nuisance functions are unknown, I estimate them using ML methods such as Penalized Regression methods, Tree Based methods and Neural Networks.

The issue with naively using ML methods for inference of low-dimensional parameter of interest is that regularization in ML methods induces bias in the estimated ML function. It is because the principal purpose of these methods is prediction where they face a trade-off between unbiased-ness and low variance. Regularization enables to reduce the variance at the cost of unbiased-ness by reducing the complexity of the model and prevent over-fitting. For example, estimating the nuisance function using Lasso regression would yield biased coefficients as $\ell 1$ penalty on coefficients pulls them towards zero (see Hastie, Tibshirani, and Wainwright (2015)). Even if I include only those variables that were qualified by Lasso to be important for predicting the dependent variable (similar to running a Restricted Linear Regression), the nuisance function may still remain biased. This is due to omission of variables that might be correlated with other dependent variables (Belloni and Chernozhukov (2011), Belloni, Chernozhukov, et al. (2013), Belloni, Chernozhukov, and Hansen (2014)). Moreover, over-fitting while estimating nuisance functions also causes bias in the low-dimensional parameter, which is alleviated by sample splitting. In subsequent sections, I detail the DML algorithm of Chernozhukov et al. (2018) and Chernozhukov et al. (2017) coded for this paper and briefly explain the usage of the ML methods in first stage estimation of nuisance parameters.

The DML approach tackles regularization bias by obtaining orthogonal residuals of each variable in low-dimension, q_{jt} , $ln(p_{jt})$ and $\mathbf{z}_{jt}$, after partialling out the effect of high-dimensional controls $\mathbf{X}_{jt}$ from them. Specifically, this is done by constructing ML estimators $\widehat{l}_{j0}(\mathbf{X}_{jt})$, $\widehat{r}_{j0}(\mathbf{X}_{jt})$, and $\widehat{\mathbf{m}}_{j0}(\mathbf{X}_{jt})$ respectively of nuisance functions:

$$\begin{aligned} \widehat{\widetilde{q_{jt}}} &= q_{jt} - \widehat{l}_{j0}(\mathbf{x}_{jt}, \Phi_t) \\ \widehat{\widetilde{ln(p_{jt})}} &= ln(p_{jt}) - \widehat{r}_{j0}(\mathbf{x}_{jt}, \Phi_t) \\ \widehat{\widetilde{\mathbf{z}_{jt}}} &= \mathbf{z}_{jt} - \widehat{\mathbf{m}}_{j0}(\mathbf{x}_{jt}, \Phi_t) \end{aligned} \tag{2.6}$$

I can now use the residuals estimated in the first stage using an ML method and plug them into GMM or 2 SLS setup to estimate the final demand parameter. However, Chernozhukov et al. (2018) point that over-fitting while estimating the nuisance functions introduce biases in the estimation of low-dimensional parameter and suggest the use of sample splitting to remove this bias. Since, sample-splitting causes significant loss of efficiency, they suggest a cross-fitting approach to restore it. The second stage GMM or 2 SLS estimation is carried out on the residuals estimated on the fraction of sample of data (main sample) which was not used for ML estimation (done on an auxiliary sample). In cross-fitting exercise, the main and auxiliary samples are flipped and re-estimated. If the samples are randomly picked the final estimate will be independent and can be averaged to obtain efficient estimates.

Caveat (Unobserved heterogeneity) : It is important to highlight that in the panel data set, unobserved rental heterogeneity could potentially bias the results. Specifically, unobserved heterogeneity enters the nuisance function as follows :

$$\begin{aligned} l_{j0}(\mathbf{X}_{jt}) &= l_0(\mathbf{X}_{jt}) + \tilde{\xi}_j \\ r_{j0}(\mathbf{X}_{jt}) &= r_0(\mathbf{X}_{jt}) + \tilde{\zeta}_j \\ \mathbf{m}_{j0}(\mathbf{X}_{jt}) &= \mathbf{m}_0(\mathbf{X}_{jt}) + \tilde{\mathbf{v}}_j \end{aligned} \tag{2.7}$$

where $\tilde{\xi}_j$, $\tilde{\zeta}_j$, and $\tilde{\mathbf{v}}_j$ denote the unobserved heterogeneity component of rental j in each nuisance function. Under weak sparsity assumption on unobserved heterogeneity, an appropriate

method is suggested by Chernozhukov et al. (2017). To keep things computationally simple in this paper, I ignore unobserved heterogeneity for now, but the effect of accounting for unobserved heterogeneity is open to further exploration. The fact that, with the exception of images, I do observe very rich set of data on individual characteristics, which provides a weak justification for proceeding without unobserved heterogeneity.

2.3.2 DML Estimation Algorithm

I translate the original DML procedure of sample splitting, ML estimation, cross-fitting, and GMM steps to adapt to the simple demand estimation problem. The steps of the algorithm are:

1. Randomly split the entire data, denoted by I, into K equal partitions. We denote each partition by I_k where $k \in \{1 \dots K\}$.

2. For each partition I_k:

   (a) For the main partition I_k, construct an auxiliary sample $I_k^c = I \setminus I_k$.

   (b) Using the auxiliary sample I_k^c, build multiple ML models for nuisance functions: $\widehat{l}_{j0}$, $\widehat{r}_{j0}$, and $\widehat{\mathbf{m}}_{j0}$ and select the one that performs best on cross-validated sample.

   (c) Obtain residuals on the main sample I_k using estimated functions and the set of equations 2.6 for all variables in the low-dimensional regime and store these residuals.

   (d) Compute the GMM estimates and standard error for the main sample I_k using residuals obtained in step (2c). Store the coefficient as $\widehat{\alpha}_{1,k}^{DML1}$ and standard error as $se(\widehat{\alpha}_{1,k}^{DML1})$.

3. **DML1 :** Compute $\widehat{\alpha}_1^{DML1} = K^{-1} \sum_1^K \widehat{\alpha}_{1,k}^{DML1}$ and $se(\widehat{\alpha}_1^{DML1}) = [K^{-2} \sum_1^K (se(\widehat{\alpha}_{1,k}^{DML1}))^2]^{1/2}$

4. **DML2 :** Compute GMM estimates ($\widehat{\alpha}_1^{DML2}$) and standard error ($se(\widehat{\alpha}_1^{DML2})$) by pooling residuals computed in (2c)

The key challenge while coding the above algorithm with ML methods on such a massive data set is to obtain parallel and scalable computational process that uses all CPU cores. This was achieved in R programming language and by using H2O, an open source ML platform, through R.

2.3.3 Regularized Linear Regression

Next, I discuss the Penalized Linear Regression methods used for estimating the nuisance parameters and their hyperparameters. In Appendix A, I describe hyperparameters and tuning issues for other ML methods such as Random Forest, Boosted Trees, and Neural Networks. Since the low-dimensional targets to be predicted using ML methods are continuous variables, I apply Supervised Regression-based ML methods for each target. In particular, I implement Penalized Linear Regression methods (Lasso and Elastic-Net) to estimate the nuisance parameter in full sample exercise.

These methods are used to perform the first stage prediction exercise of the DML Algorithm where the setting is such that the low-dimensional variable (for example, the demand, q or log of inflation-adjusted prices, $ln(p)$) is some complex parametric function of a large number of covariates (predictors/features). In this paper, these covariates are the high-dimensional confounders or control variables. I do not necessarily assume a very strict parametric form on these predictive models. I only assume them to belong to a given class of ML function and instead let the algorithm decide the exact functional form within that class using out-of-sample performance. I manually supply the hyperparameters to the algorithms for tuning the model.

While tuning within each class of models using a different combination of hyperparameters, the primary objective is to find a complex model which neither under-fits nor over-fits. Thus, the coded algorithm runs with different combinations of hyperparameters and selects that particular set which performs best in terms of prediction accuracy on new samples of data drawn from the same population but not seen by the training algorithm. The performance is measured by the cross-validated mean-squared error (MSE).

In each of the first stage ML exercise, the target is one of the low dimensional variables, in-

cluding the instruments. I now describe how I use these methods in practice to find a well tuned model. To keep things simple in this section, I denote first stage outcome/target/dependent variable by y_{jt} which can be one of the following low-dimensional variable: $\{q_{jt}, ln(p_{jt}), \mathbf{z}_{jt}\}$. I denote the predictors/features/independent variables $\mathbf{x}_{jt} = [x_{1,jt}, \ldots, x_{K,jt}]'$ that includes high-dimensional covariates from data (text and otherwise) plus the information set.

Regularized Regression methods represent a class of ML models in which a linear fit is obtained from a least-squares minimization exercise subject to some penalty on the size of the coefficients (see Hastie, Tibshirani, and Wainwright (2015) for more details). In this paper, I use Lasso (Tibshirani (1996)) and Elastic Net (Zou and Hastie (2005)). I ignore Ridge Regression (Hoerl and Kennard (1970)) in favor of Lasso and Elastic Net because though I suspect the presence of multi-collinearity that increases the variance of the estimator, I also assume the system to have a sparse solution especially because of the text data. Keeping that in mind, I use Lasso where $\ell 1$-penalty forces the regression coefficients to be zero. Since the performance of Ridge Regression dominates Lasso when there is a high correlation between predictors (Tibshirani (1996)), I also use Elastic Net that possesses components of $\ell 2$-penalty along with $\ell 1$-penalty. Nevertheless, I exploit the Elastic Net and its special case, Lasso, for their desirable properties and computational ease. In particular, the Elastic Net problem solves:

$$\underset{\beta}{\text{minimize}} \left\{ \frac{1}{2}(y_{jt} - \mathbf{x}'_{jt}\beta)^2 + \lambda \left[\alpha \|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2^2 \right] \right\} \tag{2.8}$$

where β is the vectors of coefficient on each independent variable. λ defines the strength of regularization to prevent over-fitting. Larger values of λ impose higher regularization forcing more coefficients towards zero. When $\lambda = 0$, the model shrinks to OLS, i.e. no regularization. α controls the weights on $\ell 1$ and $\ell 2$ penalties in the Elastic Net. When $\alpha = 0$, the model corresponds to Ridge Regression and $\alpha = 1$, the model corresponds to Lasso. λ and α are the hyper-parameters in the model, which are chosen from a user-defined list based on cross validated out-of-sample performance measured by MSE.

Essentially, by controlling hyper-parameters α and λ, I try to strike a balance between over-fitting versus under-fitting and sparsity versus high variance due to collinear independent variables. In practice, I may trivially choose a small set of λ and α and run regularized regression models and compare their out-of-sample performance for each combination of λ and α. In principle, such an approach could have worked if the underlying predictive model was somewhat known to the researcher. Unfortunately, in the presence of high-dimensional predictors, this is usually difficult to achieve. Therefore, in practice, I perform a more systematic and rigorous model discovery procedure, instead of a simple trial and error approach to find a good predictive model.

A typical way of finding a balance between under-fitting and over-fitting in Penalized Regression methods is to start from a very large value of λ where all coefficients are forced to zero and then slowly reduce λ and release the coefficients until the model over-fits. The algorithm would automatically detect when cross-validated MSE begins to increase again upon lowering the regularization. The only problem then remaining is the search for an appropriate value of α. Although the automatic λ search is more systematic, it tends to be more time consuming as the algorithm rigorously searches across many values of λ.

2.4 Results

This section presents the results of the application of the DML approach in estimating a simple residual demand function. I begin with the full-sample results that use Lasso and Elastic Net to estimate the nuisance parameters in the high-dimensional setup. I discuss the performance of the Regularized Linear Regression methods and then present the DML results corresponding to their predictions obtained in the first stage. Unfortunately, full sample analysis is limited to Linear Regression methods only, because other richer ML models such as Random Forest, Boosting Machines and Neural Networks, at this stage, are computationally infeasible given the size of the data. Thus, instead of performing a full sample DML procedure, I provide 1% sample estimates using these algorithms. See Appendix A for details on ML methods and their hyperparameters and Appendix B for the results

obtained on the sub-sample.

2.4.1 Lasso and Elastic Net Performance

In the first stage of the procedure, I provide a set of hyperparameter values to the Penalized Regression models in an attempt to discover a suitable model that fits well. I run a complete grid search to find optimal values of α and λ, where I run the models with all possible combinations of hyperparameters from a manually provided list and select the model with lowest cross-validated MSE. Table 2.3 reports the performance of the best Lasso and Elastic Net models selected for each target variable upon grid search. In Panel A, I restrict $\alpha = 1$ (Lasso) and experiment with $\lambda \in \{0.001, 0.01, 0.1, 1, 10, 100\}$. These are relatively weak regularization strengths, and yet the best performing model is the one with the least regularization strengths. Similarly, when I relax $\alpha \in \{0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99, 1\}$ to incorporate Elastic Net penalty (Panel B), the results are very similar to that in Panel A, where again the models with very least regularization strengths perform best on out-of-sample data. Additionally, for most of the variables, the grid search throws out models with small weights on $\ell 1$-penalty with the exception of instruments z_6 and z_9 when the cross-validation scheme is 2-fold. Nevertheless, these higher weights on $\ell 1$-penalty disappear under 5-fold cross-validation.

These evidence suggests low sparsity of high-dimensional covariates and that many of these variables are essential and must be included in the models. This is contrary to our intuition that the actual model is sparse, where a large number of features have small or no effect on price and demand variables. Besides, training MSEs are very small as the fitted models are too complex due to the inclusion of many variables, and thereby indicating an over-fitted model.

Panel C and D of Table 2.3 present the results of model selection through a more systematic approach where I enable the ML algorithm to search for an optimal λ systematically. The algorithm begins with a sufficiently high value of λ, where the complexity of the model is reduced to that of a null model by forcing all coefficients to zero. The algorithm then

Table 2.3 – Lasso and Elastic Net Performance

	2-fold			5-fold		
Target Variable	**Hyper-Paramenters**	**MSE (Training)**	**MSE (CV)**	**Hyper-Paramenters**	**MSE (Training)**	**MSE (CV)**
Panel A : $\ell 1$ Penalty or Lasso (Without λ search)						
q	$\alpha = 1, \lambda = 0.001$	0.88	0.91	$\alpha = 1, \lambda = 0.001$	0.88	0.9
$ln(p)$	$\alpha = 1, \lambda = 0.001$	0.04	0.04	$\alpha = 1, \lambda = 0.001$	0.04	0.04
z_1	$\alpha = 1, \lambda = 0.001$	1.87	1.91	$\alpha = 1, \lambda = 0.001$	1.87	1.9
z_2	$\alpha = 1, \lambda = 0.001$	8.43	8.79	$\alpha = 1, \lambda = 0.001$	8.43	8.59
z_3	$\alpha = 1, \lambda = 0.001$	19.5	20.45	$\alpha = 1, \lambda = 0.001$	19.5	20.02
z_4	$\alpha = 1, \lambda = 0.001$	34.84	36.48	$\alpha = 1, \lambda = 0.001$	34.84	35.88
z_5	$\alpha = 1, \lambda = 0.001$	53.74	56.03	$\alpha = 1, \lambda = 0.001$	53.74	55.03
z_6	$\alpha = 1, \lambda = 0.001$	22.43	23.1	$\alpha = 1, \lambda = 0.001$	22.43	22.97
z_7	$\alpha = 1, \lambda = 0.001$	99.17	101.66	$\alpha = 1, \lambda = 0.001$	99.17	100.69
z_8	$\alpha = 1, \lambda = 0.001$	227.81	234.54	$\alpha = 1, \lambda = 0.001$	227.81	233.22
z_9	$\alpha = 1, \lambda = 0.001$	403.17	413.96	$\alpha = 1, \lambda = 0.001$	403.17	412.51
z_{10}	$\alpha = 1, \lambda = 0.001$	631.4	646.58	$\alpha = 1, \lambda = 0.001$	631.4	646.24
Panel B : Elastic Net Penalty (Without λ search)						
q	$\alpha = 0.1, \lambda = 0.001$	0.87	0.9	$\alpha = 0.01, \lambda = 0.001$	0.86	0.89
$ln(p)$	$\alpha = 0.01, \lambda = 0.001$	0.03	0.03	$\alpha = 0.01, \lambda = 0.001$	0.03	0.03
z_1	$\alpha = 0.01, \lambda = 0.001$	1.76	1.82	$\alpha = 0.01, \lambda = 0.001$	1.76	1.81
z_2	$\alpha = 0.01, \lambda = 0.001$	7.97	8.26	$\alpha = 0.01, \lambda = 0.001$	7.97	8.16
z_3	$\alpha = 0.01, \lambda = 0.001$	18.92	19.38	$\alpha = 0.1, \lambda = 0.001$	18.87	19.31
z_4	$\alpha = 0.01, \lambda = 0.001$	33.81	34.73	$\alpha = 0.01, \lambda = 0.001$	33.81	34.47
z_5	$\alpha = 0.01, \lambda = 0.001$	52.47	53.85	$\alpha = 0.01, \lambda = 0.001$	52.47	53.3
z_6	$\alpha = 0.75, \lambda = 0.001$	22.43	23.01	$\alpha = 0.1, \lambda = 0.001$	22.47	22.93
z_7	$\alpha = 0.25, \lambda = 0.001$	98.36	100.92	$\alpha = 0.25, \lambda = 0.001$	98.36	100.25
z_8	$\alpha = 0.01, \lambda = 0.001$	229.22	232.27	$\alpha = 0.5, \lambda = 0.001$	227.89	232.2
z_9	$\alpha = 1, \lambda = 0.001$	403.17	413.49	$\alpha = 0.01, \lambda = 0.001$	406.58	411.53
z_{10}	$\alpha = 0.5, \lambda = 0.001$	627.36	644.38	$\alpha = 0.1, \lambda = 0.001$	630.16	639.19
Panel C : $\ell 1$ Penalty or Lasso (With λ search)						
q	$\alpha = 1$	0.92	0.93	$\alpha = 1$	0.92	0.92
$ln(p)$	$\alpha = 1$	0.05	0.06	$\alpha = 1$	0.05	0.05
z_1	$\alpha = 1$	1.83	1.95	$\alpha = 1$	1.83	1.89
z_2	$\alpha = 1$	8.61	9.29	$\alpha = 1$	8.61	8.94
z_3	$\alpha = 1$	20.82	22.71	$\alpha = 1$	20.82	21.7
z_4	$\alpha = 1$	38.32	42.21	$\alpha = 1$	38.32	40.07
z_5	$\alpha = 1$	60.69	67.5	$\alpha = 1$	60.69	63.54
z_6	$\alpha = 1$	26.04	27.17	$\alpha = 1$	26.04	26.45
z_7	$\alpha = 1$	123.08	129.95	$\alpha = 1$	123.08	125.75
z_8	$\alpha = 1$	298.49	317.98	$\alpha = 1$	298.49	306.15
z_9	$\alpha = 1$	555.53	596.84	$\alpha = 1$	555.53	572.5
z_{10}	$\alpha = 1$	894.93	968.71	$\alpha = 1$	894.93	925.19
Panel D : Elastic Net Penalty (With λ search)						
q	$\alpha = 0.99$	0.92	0.93	$\alpha = 0.9$	0.92	0.92
$ln(p)$	$\alpha = 0.99$	0.05	0.06	$\alpha = 0.9$	0.05	0.05
z_1	$\alpha = 0.99$	1.83	1.95	$\alpha = 0.9$	1.83	1.89
z_2	$\alpha = 0.99$	8.61	9.31	$\alpha = 0.9$	8.61	8.94
z_3	$\alpha = 0.99$	20.82	22.72	$\alpha = 0.9$	20.82	21.67
z_4	$\alpha = 0.99$	38.32	42.25	$\alpha = 0.9$	38.32	40.02
z_5	$\alpha = 0.99$	60.69	67.55	$\alpha = 0.9$	60.69	63.48
z_6	$\alpha = 0.99$	26.04	27.16	$\alpha = 0.9$	26.04	26.45
z_7	$\alpha = 0.99$	123.08	130.03	$\alpha = 0.9$	123.08	125.73
z_8	$\alpha = 0.99$	298.49	318.17	$\alpha = 0.9$	298.31	306.03
z_9	$\alpha = 0.99$	555.53	597.45	$\alpha = 0.9$	554.89	571.86
z_{10}	$\alpha = 0.99$	894.93	970.34	$\alpha = 0.9$	893.12	923.66

Notes : q denotes the quantity as the count of persons. $ln(p)$ is the natural log of price per person. $z_1 - z_{10}$ are the instruments constructed from the past hosting behavior of each rental.

Table 2.4 – Demand Results

ML Method	2-fold		5-fold	
	DML 1	**DML 2**	**DML 1**	**DML 2**
$\ell 1$ Penalty or Lasso	-2.376^{***}	-2.696^{***}	-2.537^{***}	-2.667^{***}
(Without λ search)	(0.268)	(0.284)	(0.298)	(0.312)
Elastic Net Penalty	-1.358^{***}	-1.249^{***}	-1.249^{***}	-1.319^{***}
(Without λ search)	(0.265)	(0.254)	(0.253)	(0.255)
$\ell 1$ Penalty or Lasso	-12.873^{***}	-13.59^{***}	-11.139^{***}	-15.754^{***}
(With λ search)	(1.121)	(1.14)	(1.247)	(1.48)
Elastic Net Penalty	-12.834^{***}	-14.48^{***}	-8.665^{***}	-14.364^{***}
(With λ search)	(1.046)	(1.252)	(0.734)	(1.283)

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Notes : $N = 25,154,404$

gradually decreases regularization strength releasing the coefficients. As complexity increases, the training, as well as out-of-sample MSE, falls until the model starts to over-fit and when the cross-validated MSE begins to rise again. The performance results show that the training error of the best selected model now is higher compared to the case without the λ-search. Notice that, when I compute the model with Elastic Net penalty through rigorous λ-search, the models with a higher weight on $\ell 1$- penalty are selected for each low-dimensional target variables. These evidence suggest that the selected models are actually sparse, which is in line with our intuition.

2.4.2 Demand Estimates

After selecting the best combination of hyperparameter values for each model setting (i.e., Lasso and Elastic Net, λ-search and No λ-search, and 2-fold and 5-fold cross-validation) using the auxiliary sample, we use the best cross-validated prediction model to obtain out-of-sample residuals on the main sample. These residual for each target variable are obtained after partialling out the effect of high-dimensional control variables from these variables. Table 2.4 reports the estimates of the second stage GMM or 2-SLS estimation procedure for separate cross-fit samples (DML1) and pooled cross-fit samples (DML2).

The estimates are small in magnitude for residuals obtained from over-fitted regularized

regression, i.e., without λ-search as shown in the top panel of Table 2.4. The estimated demand parameter reflects that the demand is highly inelastic, which seems implausible in a market where there are thousands of rentals in the market at any given time. Notice that pooled GMM (DML2) yields more stable estimates across both 2-fold and 5-fold cross-fitting. Next, in the bottom panel of Table 2.4 where residual have been obtained from systematic λ-search algorithm. The result displays a notable difference in the demand results compared to the top panel, suggesting the importance of preventing over-fitting in the first stage of the DML procedure. Again, pooled GMM (DML2) appears to be more stable across different folds of cross-fitting samples. The estimates suggest an elastic demand for rentals, which is more plausible for Airbnb market of NYC.

2.4.3 Welfare

Using the estimates of demand obtained from the DML procedure, I recompute the welfare of consumers and rental owners. Under the assumption that the market conduct is Monopolistic Competition, I recover the markup and the marginal cost of renting on Airbnb. On average, the costs are now higher by about 95% (from 29$ to 56$ per person per night) than those recovered from demand estimated without taking the reviews into account in Chapter 1. The markups are also smaller by about 44% than those obtained previously. The recovered costs and markup indicate the transfer of biases from demand to supply side. Next, I compute the consumer and producer surplus using the new demand function estimated, which accounts for time-varying quality changes captured in reviews.

Table 2.5 presents recomputed consumer surplus using fresh estimates of demand. I find that the actual consumer welfare is now increased to $452M$\$ from $391M$\$ calculated in section 1.4.1. This increase suggests how biased estimates of demand can influence the welfare results. These biases also appear in surplus distribution across space-type, property-type, and location. The entire home and apartment space type generate a much higher surplus (22$) compared to Private Rooms (6.6$) and Shared Rooms(6.9$). Across location, there is a small dispersion of average surplus than those calculated in Chapter 1. The consumer welfare

Table 2.5 – Consumer Surplus Across Rentals

	(1)	(2)	(3)	(4)	(5)	(6)
	Average CS ($)				Total CS (1000$)	
	All Rentals		Purged Rentals		All Rentals	Purged Rentals
VARIABLES	mean	sd	mean	sd	sum	sum
All rentals	14.67	37.67	30.77	56.33	452,489	45,456
Space-Type						
Entire Home/Apartment	22.22	49.43	30.77	56.33	355,564	45,456
Private Room	6.623	14.46			89,374	
Shared Room	6.938	11.98			7,550	
Property-Type						
Apartment/Condominium/Loft/Townhouse	14.21	35.10	30.79	56.37	396,520	45,439
House	21.56	63.27			50,189	
Bungalow/Castle/Guesthouse/Vacation home/Villa	13.53	30.73			408.2	
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	6.806	13.73			232.1	
Location						
Bronx	10.70	19.39	31.74	54.73	4,986	183.1
Brooklyn	14.07	36.37	27.12	44.12	158,775	8,434
Manhattan	15.50	39.74	32.56	60.53	249,939	34,942
Queens	12.92	32.20	21.47	36.99	35,071	1,714
All Staten Island	14.41	34.12	17.26	19.21	3,707	182.8

Notes: The consumer surplus has been first averaged within rentals before reporting the summary statistics. The total consumer surplus is calculated by aggregating over all rental-time combination.

has reduced in Manhattan while increased in all the other locations from previous estimates.

Table 2.6 shows a new set of rental owner's profits. Not surprisingly, as a result of higher recovered marginal costs, the rental profits are much smaller than those calculated in Chapter 1 without controlling for reviews. The total producer surplus is $302M$\$ compared to $566M$\$ calculated previously. The ability to accommodate a larger number of people, Entire home/apartments on average, make larger profits per rental per night, which is about 15\$. After controlling for reviews, it also appears that the distribution of profits on average does not vary a lot across the location. Higher marginal costs of renting offset higher prices in prime locations such as Manhattan. Also, spatially varying demand is offset by flexible supply.

Next in Table 2.7, notice that the loss in consumer surplus due to price change has

Table 2.6 – Profit and Producer Surplus Across Rentals

	(1)	(2)	(3)	(4)	(5)	(6)
	Average Profits ($)				Total Profits (1000$)	
	All Rentals		Purged Rentals		All Rentals	Purged Rentals
VARIABLES	mean	sd	mean	sd	sum	sum
All rentals	9.514	27.18	21.40	41.29	302,243	31,787
Space-Type						
Entire Home/Apartment	14.84	35.75	21.40	41.29	243,718	31,787
Private Room	3.766	10.20			53,261	
Shared Room	4.762	9.465			5,263	
Property-Type						
Apartment/Condominium/Loft/Townhouse	9.116	25.72	21.41	41.32	262,930	31,772
House	15.33	42.77			35,471	
Bungalow/Castle/Guesthouse/Vacation home/Villa	8.776	23.32			264.7	
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	4.554	9.914			165.0	
Location						
Bronx	8.016	17.15	25.57	45.75	3,862	141.3
Brooklyn	9.428	25.89	19.71	32.49	109,575	6,192
Manhattan	9.679	28.77	22.23	44.16	161,072	23,971
Queens	9.097	23.78	16.78	30.66	25,077	1,353
Staten Island	10.03	26.93	11.60	17.26	2,648	129.2

Notes: The producer surplus has been first averaged within rentals before reporting the summary statistics. The total producer surplus is calculated by aggregating over all rental-time combination.

decreased to $1.13M\$$ from $2.33M\$$ calculated previously. As in consumer surplus, an increase in seller's profits due to price increase is now smaller. This is because while constructing the counterfactual demand for purged rental, I assume that the reviews of purged rentals seize to grow after the date they were purged. Such an assumption is necessary because of our inability to know the evolution of reviews of every purged rental during the post-purge time frame. Nevertheless, these results are in line with our intuition about the competitive nature of NYC Airbnb Market. For further details refer to Appendix C

2.5 Conclusion

In this paper, we take a new step to address the age-old problem of residual demand estimation in a market so vast and diverse and with data so rich that it becomes a computational

Table 2.7 – Welfare Summary

	(1)	(2)	(3)
Surplus	Market Presence	Setting	Welfare ($1M$\$)
Counsumer	Existing	Actual	452.49
Counsumer	Existing	Counter-factual	453.62
Counsumer	Purged	Counter-factual	29.67
Producer	Existing	Actual	302.24
Producer	Existing	Counter-factual	297.41
Producer	Purged	Counter-factual	32.47

Notes: The Existing rentals are those which were not purged. These are aggregated figures summed up over all rental-time observation in each setting.

nuisance. The paper exploits the advances in Econometrics literature that leverages Machine Learning techniques to select predictive models and perform valid estimation and inference. In particular, we apply the Double (Debiased) Machine Learning estimation approach (Chernozhukov et al. (2018) and Chernozhukov et al. (2017)) in Airbnb setting where low dimensional demand parameter is confounded by a very rich set of features and characteristics observed by both consumers and rental owners on the rental website. These features are also observed by an econometrician, given the richness of modern data. However, in the presence of rich data, it remains unknown as to which variables would enter the model and how.

The paper uses text in the form of rental amenities, descriptions, and reviews as data and considers it to control for changing the quality of the rentals in estimation. I perform text preprocessing to obtain meaningful predictive features constructed using words and phrases such that they appropriately reflect the perception of consumers and sellers about the rental. While building the features, many design choices were made based on the trade-off between computational feasibility and marginal gains from these choices. The final data thus constructed contains $25M$ observations and thousands of features built from the rich data on Airbnb rental characteristics, amenities, description, and reviews that confound the true demand relationship. These high-dimensional features are required to be partialled out from the outcome variable as well as the variable of interest without falling into the omitted

variable trap.

The DML procedure provides a data-driven way to select variables and models for prediction using ML methods and yet enable valid inference. I execute this procedure on the entire sample using penalized linear regression methods in the first stage ML estimation. The estimated demand parameter in the second stage is -14.364 with the standard error of 1.283 when we perform a systematic search for hyperparameters and use the Elastic Net penalty to obtain the first stage predictions. Lasso also yields similar results on the full sample. This is a significant improvement in the demand estimates, which did not account for changes in the quality of rentals captured in review data. Though the 1% sample results (Appendix B) are not very informative, they do highlight the importance of the first stage model fit on the final DML estimates. Tuning these models and controlling their complexity is the primary task where I try to achieve the right balance between under-fitting and over-fitting.

Table 2.7 summarizes the welfare calculations resulting from DML estimates of firm facing demand function. This paper shows that not including reviews as controls in estimating the demand can cause significant biases. These biases transfer themselves to the supply side and in welfare calculation. In particular, the marginal cost of renting on Airbnb in NYC was highly underestimated, which generated unreasonably high rental profits.

On the one hand, this paper highlights the importance of using rich data and the DML procedure to estimate low dimensional relationships. It also points out the design and computational issues a researcher may face while implementing them in practice. Although this paper estimates a simple firm-facing demand curve, there are some valuable lessons learned about the computational nature of modern econometrics.

BIBLIOGRAPHY

Antweiler, Werner, and Murray Z Frank. 2004. "Is all that talk just noise? The information content of internet stock message boards." *The Journal of finance* 59 (3): 1259–1294.

Athey, Susan. 2018. "The impact of machine learning on economics." In *The economics of artificial intelligence: An agenda*. University of Chicago Press.

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu. 2019. "Ensemble Methods for Causal Effects in Panel Data Settings." In *AEA Papers and Proceedings*, 109:65–70.

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Athey, Susan, and Guido W Imbens. 2019. "Machine learning methods that economists should know about." *Annual Review of Economics* 11.

— . 2017. "The state of applied econometrics: Causality and policy evaluation." *Journal of Economic Perspectives* 31 (2): 3–32.

Athey, Susan, Julie Tibshirani, Stefan Wager, et al. 2019. "Generalized random forests." *The Annals of Statistics* 47 (2): 1148–1178.

Baker, Jonathan B, and Timothy F Bresnahan. 1988. "Estimating the residual demand curve facing a single firm." *International Journal of Industrial Organization* 6 (3): 283–300.

Barron, Kyle, Edward Kung, and Davide Proserpio. 2018. "The sharing economy and housing affordability: Evidence from Airbnb."

Belloni, Alexandre, and Victor Chernozhukov. 2011. "High dimensional sparse econometric models: An introduction." In *Inverse Problems and High-Dimensional Estimation*, 121–156. Springer.

Belloni, Alexandre, Victor Chernozhukov, et al. 2013. "Least squares after model selection in high-dimensional sparse models." *Bernoulli* 19 (2): 521–547.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-dimensional methods and inference on structural and treatment effects." *Journal of Economic Perspectives* 28 (2): 29–50.

Belloni, Alexandre, Victor Chernozhukov, and Ying Wei. 2016. "Post-selection inference for generalized linear models with many controls." *Journal of Business & Economic Statistics* 34 (4): 606–619.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting poverty and wealth from mobile phone metadata." *Science* 350 (6264): 1073–1076.

Breiman, Leo. 2017. *Classification and regression trees*. Routledge.

Chen, M Keith, and Michael Sheldon. 2016. "Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the Uber Platform." In *Ec*, 455.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21, no. 1 (): C1–C68. eprint: http://oup.prod.sis.lan/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf.

Chernozhukov, Victor, Matt Goldman, Vira Semenova, and Matt Taddy. 2017. "Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels." *arXiv preprint arXiv:1712.09988*.

Cohen, Molly, and Arun Sundararajan. 2015. "Self-regulation and innovation in the peer-to-peer sharing economy." *U. Chi. L. Rev. Dialogue* 82:116.

Correia, Sergio. 2016. *Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator*. Tech. rep. Working Paper.

Cramer, Judd, and Alan B Krueger. 2016. "Disruptive change in the taxi business: The case of Uber." *American Economic Review* 106 (5): 177–82.

Cullen, Zoë, and Chiara Farronato. 2014. "Outsourcing tasks online: Matching supply and demand on peer-to-peer internet platforms." *Job Market Paper*.

Donaldson, Dave, and Adam Storeygard. 2016. "The view from above: Applications of satellite data in economics." *Journal of Economic Perspectives* 30 (4): 171–98.

Edelman, Benjamin G, and Damien Geradin. 2015. "Efficiencies and regulatory shortcuts: How should we regulate companies like Airbnb and Uber." *Stan. Tech. L. Rev.* 19:293.

Edelman, Benjamin G, and Michael Luca. 2014. "Digital discrimination: The case of Airbnb. com."

Einav, Liran, Chiara Farronato, and Jonathan Levin. 2016. "Peer-to-peer markets." *Annual Review of Economics* 8:615–635.

Farronato, Chiara, and Andrey Fradkin. 2018. *The welfare effects of peer entry in the accommodation market: The case of airbnb*. Tech. rep. National Bureau of Economic Research.

Filippas, Apostolos, and John J Horton. 2017. "The tragedy of your upstairs neighbors: When is the home-sharing externality internalized?" *Available at SSRN 2443343*.

Fradkin, Andrey. 2015. "Search frictions and the design of online marketplaces." *Work. Pap., Mass. Inst. Technol*.

Fradkin, Andrey, Elena Grewal, and David Holtz. 2018. "The determinants of online review informativeness: Evidence from field experiments on Airbnb."

Fraiberger, Samuel P, Arun Sundararajan, et al. 2015. "Peer-to-peer rental markets in the sharing economy." *NYU Stern School of Business research paper* 6.

Friedman, Jerome H. 2002. "Stochastic gradient boosting." *Computational statistics & data analysis* 38 (4): 367–378.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Gentzkow, Matthew, Bryan T Kelly, and Matt Taddy. 2017. *Text as data*. Tech. rep. National Bureau of Economic Research.

Gentzkow, Matthew, and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78 (1): 35–71.

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy. 2016. *Measuring group differences in high-dimensional choices: Method and application to Congressional speech*. Tech. rep. National Bureau of Economic Research.

Glaeser, Edward L, Scott Duke Kominers, Michael Luca, and Nikhil Naik. 2018. "Big data and big cities: The promises and limitations of improved measures of urban life." *Economic Inquiry* 56 (1): 114–137.

Gong, Jing, Brad N Greenwood, and Yiping Song. 2017. "Uber might buy me a mercedes benz: An empirical investigation of the sharing economy and durable goods purchase."

Guttentag, Daniel A, and Stephen LJ Smith. 2017. "Assessing Airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations." *International Journal of Hospitality Management* 64:1–10.

Hall, Jonathan V, John J Horton, and Daniel T Knoepfle. 2018. *Pricing efficiently in designed markets: Evidence from ride-sharing*. Tech. rep. Working Paper), New York University.

Hall, Jonathan, Cory Kendrick, and Chris Nosko. 2015. "The effects of Uber's surge pricing: A case study." *The University of Chicago Booth School of Business*.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, and James Franklin. 2005. "The elements of statistical learning: data mining, inference and prediction." *The Mathematical Intelligencer* 27 (2): 83–85.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical learning with sparsity: the lasso and generalizations*. Chapman / Hall/CRC.

Hausman, Jerry A. 1996. "Valuation of new goods under perfect and imperfect competition." In *The economics of new goods*, 207–248. University of Chicago Press.

Henderson, J Vernon, Adam Storeygard, and David N Weil. 2012. "Measuring economic growth from outer space." *American economic review* 102 (2): 994–1028.

Hoberg, Gerard, and Gordon Phillips. 2016. "Text-based network industries and endogenous product differentiation." *Journal of Political Economy* 124 (5): 1423–1465.

Hoerl, Arthur E, and Robert W Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics* 12 (1): 55–67.

Horn, Keren, and Mark Merante. 2017. "Is home sharing driving up rents? Evidence from Airbnb in Boston." *Journal of Housing Economics* 38:14–24.

Horton, John J, and Richard J Zeckhauser. 2016. *Owning, using and renting: some simple economics of the" sharing economy"*. Tech. rep. National Bureau of Economic Research.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

Kang, Jun Seok, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. "Where not to eat? Improving public policy by predicting hygiene inspections using online reviews." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1443–1448.

Kaplan, Roberta A, and Michael L Nadler. 2015. "Airbnb: A case study in occupancy regulation and taxation." *U. Chi. L. Rev. Dialogue* 82:103.

Knaus, Michael C. 2018. "A Double Machine Learning Approach to Estimate the Effects of Musical Practice on Student's Skills." *arXiv preprint arXiv:1805.10300*.

Kogan, Shimon, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. "Predicting risk from financial reports with regression." In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 272–280. Association for Computational Linguistics.

Koopman, Christopher, Matthew Mitchell, and Adam Thierer. 2014. "The sharing economy and consumer protection regulation: The case for policy change." *J. Bus. Entrepreneurship & L.* 8:529.

Lee, Dayne. 2016. "How Airbnb short-term rentals exacerbate Los Angeles's affordable housing crisis: Analysis and policy recommendations." *Harv. L. & Pol'y Rev.* 10:229.

Leeb, Hannes, and Benedikt M Pötscher. 2008. "Can one estimate the unconditional distribution of post-model-selection estimators?" *Econometric Theory* 24 (2): 338–376.

Leeb, Hannes, Benedikt M Pötscher, et al. 2006. "Can one estimate the conditional distribution of post-model-selection estimators?" *The Annals of Statistics* 34 (5): 2554–2591.

Lobell, David B. 2013. "The use of satellite data for crop yield gap analysis." *Field Crops Research* 143:56–64.

Muehlenbachs, Lucija, Elisheba Spiller, and Christopher Timmins. 2015. "The housing market impacts of shale gas development." *American Economic Review* 105 (12): 3633–59.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Peltzman, Sam. 1976. "Toward a more general theory of regulation." *The Journal of Law and Economics* 19 (2): 211–240.

Proserpio, Davide, and Gerard J Tellis. 2017. "Baring the Sharing Economy: Concepts, Classification, Findings, and Future Directions." *Classification, Findings, and Future Directions (December 28, 2017)*.

Proserpio, Davide, Wendy Xu, and Georgios Zervas. 2018. "You get what you give: theory and evidence of reciprocity in the sharing economy." *Quantitative Marketing and Economics* 16 (4): 371–407.

Quattrone, Giovanni, Davide Proserpio, Daniele Quercia, Licia Capra, and Mirco Musolesi. 2016. "Who benefits from the sharing economy of Airbnb?" In *Proceedings of the 25th international conference on world wide web*, 1385–1394. International World Wide Web Conferences Steering Committee.

Rauch, Daniel E, and David Schleicher. 2015. "Like Uber, but for local government law: the future of local regulation of the sharing economy." *Ohio St. LJ* 76:901.

Rogers, Brishen. 2015. "The social costs of Uber." *U. Chi. L. Rev. Dialogue* 82:85.

Sheppard, Stephen, Andrew Udell, et al. 2016. "Do Airbnb properties affect house prices." *Williams College Department of Economics Working Papers* 3.

Stephens-Davidowitz, Seth. 2014. "The cost of racial animus on a black candidate: Evidence using Google search data." *Journal of Public Economics* 118:26–40.

Stigler, George J. 1971. "The theory of economic regulation." *The Bell journal of economics and management science*: 3–21.

Tetlock, Paul C. 2007. "Giving content to investor sentiment: The role of media in the stock market." *The Journal of finance* 62 (3): 1139–1168.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.

Varian, Hal R. 2014. "Big data: New tricks for econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.

Zervas, Georgios, Davide Proserpio, and John Byers. 2015. "A first look at online reputation on Airbnb, where every stay is above average." *Where Every Stay is Above Average (January 28, 2015)*.

Zervas, Georgios, Davide Proserpio, and John W Byers. 2017. "The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry." *Journal of Marketing Research* 54 (5): 687–705.

Zhou, Zhi-Hua. 2012. *Ensemble methods: foundations and algorithms*. Chapman / Hall/CRC.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the royal statistical society: series B (statistical methodology)* 67 (2): 301–320.

Appendix A

ML METHODS

In this section, I describe the ML methods and issues related to their hyperparameters tuning. These methods are a part of the first stage of the DML procedure and are applied to the low-dimensional variables (targets) to partial out the effect of high-dimensional confounders. The performance of these methods is crucial to the successful inference of the low dimensional parameter in the second stage.

I use different Supervised Regression-based ML techniques for all low-dimensional variables. However, the design of my code is equipped to adapt to binary targets also. The ML models I use belong to four different classes of supervised ML methods, namely Generalized Linear Models, Random Forest, Boosting Machines, and Neural Network. The Ensemble method trains a model using out of sample predictions of above conceptually same or different algorithms. Athey et al. (2019) shows a better performance of ensemble methods over individual algorithms. I use off-the-shelf computational tools and packages from H2O to implement these algorithms using optimization methods such as Stochastic Gradient Descent (Friedman (2002)).

To keep things simple in this section, I denote first stage outcome/target/dependent variable by y_{jt} which can be one of the following low-dimensional variable: $\{q_{jt}, ln(p_{jt}), \mathbf{z}_{jt}\}$. I denote the predictors/features/independent variables $\mathbf{x}_{jt} = [x_{1,jt}, \ldots, x_{K,jt}]'$ that includes high-dimensional covariates from data (text and otherwise) plus the information set.

A.1 Random Forests

I begin with Tree-based methods (see Breiman (2017) for more details) to get prediction of the target variables y_{jt}. In particular, I use extensions (ensembles) of Decision Trees namely Random Forest and Boosting Machines (see James et al. (2013) and Hastie et al. (2005) for more details). Decision Tree is a technique where a decent prediction of the outcome is the mean of that outcome on many non-overlapping sub-samples of the data. These sub-samples are created by a recursive binary splitting technique in that a sample splits on a particular value of a single predictor from a set of candidate predictors. For every split, the method

selects the split variable ($x_{k,jt} \in \mathbf{x}_{jt}$) and the split point (c) to minimizes the sum of squared residuals (RSS) computed over samples resulting from the split when the predicted value is the mean of outcome y_{jt} in each of the two split samples. Specifically, for every sample split, (k, c) solves :

$$\underset{(k,c)}{\text{minimize}}\left[\sum_{jt:x_{k,jt}<c}\left(y_{jt}-\frac{\sum_{jt:x_{k,jt}<c} y_{jt}}{\sum_{jt:x_{k,jt}<c} 1}\right)^2+\sum_{jt:x_{k,jt}\geq c}\left(y_{jt}-\frac{\sum_{jt:x_{k,jt}\geq c} y_{jt}}{\sum_{jt:x_{k,jt}\geq c} 1}\right)^2\right] \tag{A.1}$$

In order to split samples quickly, the algorithm discretizes the space of the variable into bins. The number of bins is fed to the algorithm as a hyperparameter. The tree is grown by splitting data several times using the above rule (A.1) until some kind of restriction is reached. These restrictions or limits can be on the minimum number of observations remaining in a sample to allow further split or the maximum number of times splitting should happen, among others. These restrictions also act as hyperparameters of tree-based algorithms. In regression problems, the mean outcome in each of the final non-overlapping data partitions (leaf of a tree) is used as the predictor.

In random forest, instead of splitting one sample (growing a single tree), the method splits several bootstrapped samples (grow multiple trees or a forest). In addition, for every split in the trees grown, the algorithm considers a random subset of predictors, $\mathbf{x}_{jt}$, say M out of total K predictors. The number of trees (bootstrapped samples) is also provided as a hyperparameter to the random forest algorithm. The final prediction is based on the bootstrapped average of predictions from each of the weak learners (the trees).

A.2 Boosting Machines

Like Bagging (bootstrapping) and Random Forest, Boosting also performs tree ensembling, but instead of building trees independently on the bootstrapped sample, it grows trees sequentially (see James et al. (2013) and Hastie et al. (2005) for more details). The idea of boosting is to learn from the residuals of the previously grown trees. Specifically, the algorithm to implement boosted tress is as follows:

1. Start with Null decision tree : $\hat{f}(\mathbf{x}_{jt}) = 0$, and residuals $\hat{r}_{jt} = y_{jt}$

2. For each tree $b \in \{1, \dots B\}$, where B is the number of trees to be grown.

(a) Build a tree ($\hat{f}^b(\mathbf{x}_{jt})$) by splitting the data d times according to the rule A.1 on the residuals $\hat{r}_{jt}$.

(b) Update the residuals: $\hat{r}_{jt} \leftarrow \hat{r}_{jt} - \epsilon \hat{f}^b(\mathbf{x}_{jt})$, where $\epsilon \in (0, 1)$ is the shrinkage parameter or learning-rate.

(c) Update the predictor: $\hat{f}(\mathbf{x}_{jt}) \leftarrow \hat{f}(\mathbf{x}_{jt}) + \epsilon \hat{f}^b(\mathbf{x}_{jt})$.

3. Obtain final estimator: $\hat{f}(\mathbf{x}_{jt})$.

Similar to RF, in the boosted tree algorithm, we supply the number of trees B, maximum depth of each tree d, minimum number of observation necessary to split and number of bins as hyperparameters. In addition, we also provide learn rate ϵ which determines the fraction of the stage predictor's strength to carry over to the next stage. Slow learning (lower values of ϵ) usually performs well as it prevents over-fitting.

A.3 Neural Networks

Neural Networks is a technique where the estimated regression function is built with a combination of several transformed linear functions with numerous unknown parameters. For example, consider the following complex function that links predictors $\mathbf{x}_{jt}$ to outcome y_{jt}:

$$y_{jt} = \sum_{k_L=1}^{K_L} \beta_{k_L}^{(L)} \underbrace{g\Big(\sum_{k_{L-1}=1}^{K_{L-1}} \beta_{k_L,k_{L-1}}^{(L-1)} \dots \underbrace{g\Big(\sum_{k_1=1}^{K_1} \beta_{k_2,k_1}^{(1)} \underbrace{g\Big(\sum_{k_0=1}^{K_0} \beta_{k_1,k_0}^{(0)} x_{k_0,jt} \Big)}_{\text{Layer 1}} \Big)}_{\text{Layer 2}} \dots \Big)}_{\text{Layer L}} + \epsilon_{jt} \tag{A.2}$$

where L is the number of layers and K_l are the number of neurons in layer l. At each layer l and neuron k_l, a linear function is created from the outputs of the previous layer k_{l-1} and transformed using a non linear function g such as hyperbolic tan, maxout, and rectifier function. $\beta_{k_1,k_0}^{(0)}$ is the weight on k_0th input predictor $x_{k_0,jt}$ that forms input to neuron k_1 of next layer i.e. layer 1. $\beta_{k_{l+1},k_l}^{(l)}$ is the weight on k_l neuron of layer $l \in \{1 \dots L\}$ that forms input to neuron k_{l+1} of next layer $l+1$. In order to fix a model, the number of layers L, and neurons in each hidden layer $\{K_1 \dots K_L\}$ are supplied as hyper parameters. The complexity of a Neural Network model depends on these tuning parameters.

The objective of a Neural Network Algorithm is to estimate the set of weights $\mathbf{B} = \{\beta^{(0)}, \beta^{(1)} \dots \beta^{(L)}\}$ that minimizes a loss function subject to regularization:

$$\underset{\mathbf{B}}{\text{minimize}} \left\{ \frac{1}{2} \|\mathbb{L}(\mathbf{B}; y_{jt}, \mathbf{x}_{jt})\|_2^2 + \left[\lambda_1 \|\mathbf{B}\|_1 + \frac{1}{2} \lambda_2 \|\mathbf{B}\|_2^2 \right] \right\} \tag{A.3}$$

where λ_1 and λ_2 are the strengths of $\ell 1$ and $\ell 2$ regularization penalties on the weights. These penalties on weights control the complexity of the model and prevent over fitting. For example, $\ell 1$-penalty can force unimportant nodes to zero, which can sometimes greatly simplify a deep complicated network. I provide these as tuning parameters to the algorithm.

A.4 Ensemble Learning

Ensemble learning is a general technique that aggregates a diverse set of ML methods to obtain better predictions than individual methods (see Zhou (2012) for more details). The idea is to build and estimate an umbrella ML function (called the "meta learner") of predictions from the trained individual methods (called the "base learners"). The process involves training each base learner on the data, and use cross-validated prediction of each of the base learners to build and train the "meta learner" in the next stage. The umbrella function can be one of Regularized Regression, Random Forest, Boosting, or Neural Network. In this paper, I use the four best performing ML functions from each class of ML methods discussed earlier in this section as the base learners and a simple linear regression function as the meta learner.

Appendix B

SUB SAMPLE ESTIMATES

In this section, I revisit the simple demand estimation problem where high dimensional functions are well approximated by one of the following: Regularized Linear Regression, Random Forest, Boosted Trees, Neural Net, or Ensemble. I describe the results of the DML procedure on a random sample of 1000 rentals (about 1%). I perform a small sample exercise to test and compare the performance of ML methods other than Regularized Linear Regression. Due to computational constraints, running the sophisticated methods on the full sample of the data is entirely infeasible. The size of the sample data is $241,152$ observations. However, this time, I allow a much larger number of features to enter our specification compared to when I ran the process on the full sample, as described in the previous section 2.4.1. This was possible due to a smaller sample size. I enable 800 and 3000 features constructed from amenities and description data, respectively. I also use 5000 most frequent features from the reviews data. The remaining rental characteristics enter the model as earlier. Although, since there are fewer rentals than the original data, many of the categorical variables have only a limited number of categories and thus add less to the dimensionality of the data than the full sample.

In the first stage, I use ML methods described in Appendix A and the cross-fitted predictions to construct the residuals for each of our low-dimensional variables of interest. For every variable, I first find a relatively good model within each class of ML methods. Table B.1 presents the choice of hyperparameters used for training these models. I use a random search method to find a reasonably good set of tuning parameters where algorithm randomly selects 10 combinations of hyper-parameters from its user-defined space for each class of ML methods and train the models with them. Next, I use an Ensemble method to combine the best models ("usually called the base learners") within each class and build a simple linear umbrella predictor. The weights are calculated using least squares on predictions on the cross-validation data. Finally, I find the best of the five previously trained models, i.e., the best within each class and the ensemble, and use residuals for GMM estimation.

Table B.1 – ML Algorithm Hyper-Parameters

ML Method	Hyper-Parameter	Values
Regularized Linear Regression	λ	$\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$
	α	$\{0.01, 0.1, 0.5, 0.9, 1\}$
Random Forest	*ntrees*	$\{50, 100, 500\}$
	max_depth	$\{20, 40, 80\}$
	min_rows	$\{100, 1000, 2000\}$
	nbins	$\{20, 40, 80\}$
Boosted Trees	*learn_rate*	$\{0.1, 0.3, 0.7\}$
	ntrees	$\{50, 100, 500\}$
	max_depth	$\{20, 40, 80\}$
	min_rows	$\{100, 1000, 2000\}$
	nbins	$\{20, 40, 80\}$
Neural Network	*hidden*	$[10, 20, 10], [50, 20]$
	activation	{"Rectifier", "Maxout", "Tanh"}
	$\ell 1$	$\{0, 0.00001, 0.0001, 0.001, 0.01, 0.1\}$
	$\ell 2$	$\{0, 0.00001, 0.0001, 0.001, 0.01, 0.1\}$

B.1 Machine Learning Algorithms

We perform the DML procedure 2-fold and 5-fold cross fitting exercise. Table B.2 and table B.3 report the performance of the best ML method discovered through random search. We find models trained on 5-fold samples perform better and are more consistent compared to those trained on 2-fold CV samples. For example, we find training MSE for target $ln(p)$ in 2-fold CV is too small compared to that of 5-fold CV despite fitting a less complex model.

In regularized linear regression, we experiment with values of regularization strength, λ, and distribution of $\ell 1$-penalty and $\ell 2$-penalty, α. I find that for most of the targets, the fitted models are complex as the regularization strength is low. Moreover, there is a clear sign of over-fitting for target $ln(p)$ where training error (0.008) is too low in contrast to CV error (0.011). Although the model uses little regularization 5-fold CV estimates, the error rates are similar to other models, such as random forest and neural networks.

The random forest models show stable performance both in 2-fold and 5-fold cv samples. To tune random forest, I use the number of trees (*ntrees*), the number of splits in each tree (*max_depth*), the minimum number of observation to split (*min_rows*), and the number of bins to evaluate best split-point (*nbins*). Although I find the similar performance of all the models of random forest across 2-fold and 5-fold CV samples, it takes a larger number of trees (bootstrapped samples) to arrive at a stable model in 2-folds.

Table B.2 – ML Algorithm Performance (2-fold)

Target Variable	**Hyper-Parameters**	**MSE (Training)**	**MSE (CV)**
Panel A : Linear Model (Elastic Net Penalty)			
q	$\alpha = 1, \lambda = 0.01$	0.565	0.577
$ln(p)$	$\alpha = 1, \lambda = 1e-04$	0.008	0.011
z_1	$\alpha = 1, \lambda = 1e-04$	1.290	0.949
z_2	$\alpha = 1, \lambda = 0.001$	2.365	4.228
z_3	$\alpha = 0.1, \lambda = 0.01$	9.725	11.716
z_4	$\alpha = 0.1, \lambda = 0.01$	24.661	20.560
z_5	$\alpha = 1, \lambda = 0.01$	27.568	33.361
Panel B : Random Forest			
q	$ntrees = 100, max\_depth = 40, min\_rows = 100, nbins = 80$	0.675	0.733
$ln(p)$	$ntrees = 50, max\_depth = 80, min\_rows = 100, nbins = 40$	0.027	0.031
z_1	$ntrees = 113, max\_depth = 40, min\_rows = 100, nbins = 40$	1.179	1.366
z_2	$ntrees = 50, max\_depth = 80, min\_rows = 100, nbins = 40$	4.738	5.985
z_3	$ntrees = 57, max\_depth = 80, min\_rows = 100, nbins = 80$	10.607	13.125
z_4	$ntrees = 78, max\_depth = 40, min\_rows = 100, nbins = 80$	18.028	22.118
z_5	$ntrees = 11, max\_depth = 20, min\_rows = 1000, nbins = 20$	48.641	60.981
Panel C : Boosted Trees			
q	$learn\_rate = 0.1, ntrees = 44, max\_depth = 20, min\_rows = 100, nbins = 20$	0.097	0.788
$ln(p)$	$learn\_rate = 0.1, ntrees = 100, max\_depth = 20, min\_rows = 1000, nbins = 20$	0.012	0.033
z_1	$learn\_rate = 0.1, ntrees = 39, max\_depth = 20, min\_rows = 100, nbins = 40$	0.202	1.496
z_2	$learn\_rate = 0.7, ntrees = 26, max\_depth = 80, min\_rows = 100, nbins = 20$	0.077	7.436
z_3	$learn\_rate = 0.1, ntrees = 97, max\_depth = 40, min\_rows = 1000, nbins = 40$	4.972	19.563
z_4	$learn\_rate = 0.1, ntrees = 84, max\_depth = 40, min\_rows = 100, nbins = 40$	0.346	23.736
z_5	$learn\_rate = 0.1, ntrees = 100, max\_depth = 40, min\_rows = 100, nbins = 80$	0.344	34.613
Panel D : Neural Network			
q	$hidden = 10-20-10, activation = Rectifier, l1 = 0.001, l2 = 1e-04$	0.790	0.837
$ln(p)$	$hidden = 10-20-10, activation = Rectifier, l1 = 1e-04, l2 = 0.001$	0.020	0.025
z_1	$hidden = 50-20, activation = Maxout, l1 = 1e-05, l2 = 1e-05$	1.683	1.613
z_2	$hidden = 10-20-10, activation = Rectifier, l1 = 1e-05, l2 = 0.001$	6.715	7.426
z_3	$hidden = 50-20, activation = Tanh, l1 = 0, l2 = 1e-05$	17.225	17.167
z_4	$hidden = 10-20-10, activation = Tanh, l1 = 1e-04, l2 = 1e-05$	60.256	37.844
z_5	$hidden = 10-20-10, activation = Rectifier, l1 = 1e-04, l2 = 0$	51.668	49.855
Panel E : Ensemble			
q	NA	0.528	0.576
$ln(p)$	NA	0.008	0.011
z_1	NA	0.896	0.887
z_2	NA	1.403	3.753
z_3	NA	6.840	9.077
z_4	NA	9.830	15.685
z_5	NA	7.142	23.578
Panel F : Best			
q	$Best\_model = Ensemble$	0.528	0.576
$ln(p)$	$Best\_model = Ensemble$	0.008	0.011
z_1	$Best\_model = Ensemble$	0.896	0.887
z_2	$Best\_model = Ensemble$	1.403	3.753
z_3	$Best\_model = Ensemble$	6.840	9.077
z_4	$Best\_model = Ensemble$	9.830	15.685
z_5	$Best\_model = Ensemble$	7.142	23.578

Table B.3 – ML Algorithm Performance (5-fold)

Target Variable	**Hyper-Parameters**	**MSE (Training)**	**MSE (CV)**
Panel A : Linear Model (Elastic Net Penalty)			
q	$\alpha = 0.01, \lambda = 0.01$	0.640	0.660
$ln(p)$	$\alpha = 0.5, \lambda = 1e-04$	0.035	0.036
z_1	$\alpha = 0.9, \lambda = 1e-04$	1.739	1.797
z_2	$\alpha = 0.1, \lambda = 0.001$	8.055	8.338
z_3	$\alpha = 0.5, \lambda = 0.001$	19.310	20.069
z_4	$\alpha = 0.01, \lambda = 0.01$	34.010	35.222
z_5	$\alpha = 0.5, \lambda = 0.001$	54.268	56.350
Panel B : Random Forest			
q	$ntrees = 27, max\_depth = 40, min\_rows = 100, nbins = 80$	0.650	0.655
$ln(p)$	$ntrees = 45, max\_depth = 80, min\_rows = 100, nbins = 20$	0.027	0.036
z_1	$ntrees = 35, max\_depth = 40, min\_rows = 100, nbins = 80$	1.342	1.439
z_2	$ntrees = 30, max\_depth = 80, min\_rows = 100, nbins = 80$	5.659	6.324
z_3	$ntrees = 40, max\_depth = 20, min\_rows = 100, nbins = 80$	12.259	14.100
z_4	$ntrees = 26, max\_depth = 40, min\_rows = 100, nbins = 80$	20.372	23.988
z_5	$ntrees = 31, max\_depth = 80, min\_rows = 100, nbins = 80$	28.551	35.261
Panel C : Boosted Trees			
q	$learn\_rate = 0.1, ntrees = 50, max\_depth = 20, min\_rows = 2000, nbins = 80$	0.484	0.676
$ln(p)$	$learn\_rate = 0.1, ntrees = 293, max\_depth = 80, min\_rows = 1000, nbins = 40$	0.003	0.030
z_1	$learn\_rate = 0.1, ntrees = 77, max\_depth = 20, min\_rows = 1000, nbins = 40$	0.672	1.428
z_2	$learn\_rate = 0.1, ntrees = 100, max\_depth = 20, min\_rows = 100, nbins = 40$	0.303	5.712
z_3	$learn\_rate = 0.3, ntrees = 43, max\_depth = 20, min\_rows = 100, nbins = 20$	0.518	13.482
z_4	$learn\_rate = 0.1, ntrees = 244, max\_depth = 80, min\_rows = 1000, nbins = 80$	2.054	20.765
z_5	$learn\_rate = 0.1, ntrees = 170, max\_depth = 20, min\_rows = 100, nbins = 20$	0.213	26.307
Panel D : Neural Network			
q	$hidden = 10-20-10, activation = Maxout, l1 = 0.01, l2 = 1e-04$	1.053	0.733
$ln(p)$	$hidden = 10-20-10, activation = Maxout, l1 = 0, l2 = 0$	0.034	0.030
z_1	$hidden = 50-20, activation = Rectifier, l1 = 1e-04, l2 = 0.001$	1.631	1.790
z_2	$hidden = 10-20-10, activation = Rectifier, l1 = 0, l2 = 1e-05$	8.400	8.133
z_3	$hidden = 10-20-10, activation = Rectifier, l1 = 1e-04, l2 = 0.001$	15.721	17.801
z_4	$hidden = 10-20-10, activation = Rectifier, l1 = 0, l2 = 1e-04$	23.613	28.548
z_5	$hidden = 50-20, activation = Rectifier, l1 = 1e-04, l2 = 0.001$	40.941	43.365
Panel E : Ensemble			
q	NA	0.515	0.637
$ln(p)$	NA	0.011	0.025
z_1	NA	0.871	1.341
z_2	NA	1.161	5.355
z_3	NA	3.579	12.017
z_4	NA	5.258	19.585
z_5	NA	1.502	24.901
Panel F : Best			
q	$Best\_model = Ensemble$	0.515	0.637
$ln(p)$	$Best\_model = Ensemble$	0.011	0.025
z_1	$Best\_model = Ensemble$	0.871	1.341
z_2	$Best\_model = Ensemble$	1.161	5.355
z_3	$Best\_model = Ensemble$	3.579	12.017
z_4	$Best\_model = Ensemble$	5.258	19.585
z_5	$Best\_model = Ensemble$	1.502	24.901

Boosted trees have performed poorly across both 2-fold and 5-fold CV samples, which, besides, to learn rate ($learn\_rate$), uses the same set of hyperparameters as random forest. This is perhaps because of the poor choice of manually provided hyperparameter values, especially the number of splits (max_depth) and learn rate ($learn\_rate$). Growing dense trees combined with rapid learning often over-fits boosted trees. This is evident from extremely small training MSE and large cross-validated MSE in 2-fold, and 5-fold CV samples where the number of splits is large or the learn rate is high.

Next, I test a different set of hyperparameters on the Neural Network method. The first hyperparameter is the network itself, which can be arbitrarily complex, depending on the user. In this exercise, I work with the conservative complexity of the models and pick instead of small networks for training: the larger the network size and complexity, the more time and computational resources it consumes. I further limit complexity by using $\ell 1$ and $\ell 2$ penalties on the weights of the neural network. Since I am not using an extremely complex network, I choose modest weights. I train neural networks with different activation functions. Neural Networks in 2-fold CV show some minor instability, especially in estimating the nuisance for $ln(p)$, where the training MSE id very low in spite of a relatively less complicated model. The models fitted on 5 fold CV samples perform better out of sample and show similar performance as Random Forest. I also test the choice of the activation function, and I find the rectifier and maxout functions are preferred over complex tanh function.

The Ensemble uses Linear Regression as the meta learner of the above-discussed base learner. The Ensembles, like all other base models, show more stability in 5-fold cross-validated samples. However, their risk of over-fitting is also linked to the base learners. The low training mean squared error in 2-fold cross-validation is transferred from overfitted penalized regression and boosting. The quality of the Ensemble also improves with the quality of other ML predictions, as we see in 5-fold cross-validation samples. Not surprisingly, the best model is the Ensemble in both 2-fold and 5-fold cross-validation, where the best out of sample predictions, some of which coming from lucky over fitted models, dominate in their effect on the Ensemble.

B.2 Demand Estimates

Table B.4 reports the estimates of demand for a small sample when I use all the ML algorithms in the first stage of the procedure. The results are tightly linked to the model fit in the first stage. In the small sample, DML results with 2-fold cross-fitting are unstable and misleading,

Table B.4 – Demand Results (All Machine Learning Algorithms)

	Elastic Net	**Random Forest**	**Boosted Trees**	**Neural Network**	**Ensemble**	**Best**
	Panel A : 2-fold					
DML 1	0.015	−4.635***	−5.307***	−0.595	0.125	0.125
	(3.211)	(0.901)	(1.306)	(0.610)	(1.583)	(1.583)
DML 2	3.173	−4.686***	−5.140***	3.405***	1.809	1.809
	(2.790)	(0.811)	(1.008)	(1.297)	(1.653)	(1.653)
	Panel B : 5-fold					
DML 1	−4.958***	−6.120***	−8.741***	−5.203***	−3.135***	−3.135***
	(0.318)	(0.293)	(0.510)	(0.373)	(0.184)	(0.184)
DML 2	−5.083***	−6.232***	−8.859***	−3.356***	−3.123***	−3.123***
	(0.326)	(0.299)	(0.510)	(0.330)	(0.183)	(0.183)

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Notes : k denotes the minimum percentage values of a feature's document frequency, below which features have been trimmed out.

especially when the first stage models over-fit. 2-fold Elastic-Net where many target variables seem to over-fit yields positive though insignificant estimates. On the contrary, 2-fold, as well as 5-fold over-fitted, Boosted Trees, result in substantial negative demand estimates. Among 5-fold cross-fitted models provide more stable estimates of demand. In particular, Elastic-Net, Random Forest and Neural Net exhibit stable DML1 and DML2 estimates ranging from −3.356 to −6.232. The DML2 estimate of −3.123 is obtained from the best first-stage model, which also happens to be the ensemble method. Although these results are obtained in a tiny sample (1% of the total data), comparing these estimates suggest the importance of obtaining just the right model in the first stage.

Appendix C

SUPPLY SIDE AND WELFARE CALCULATIONS

C.1 Demand Side

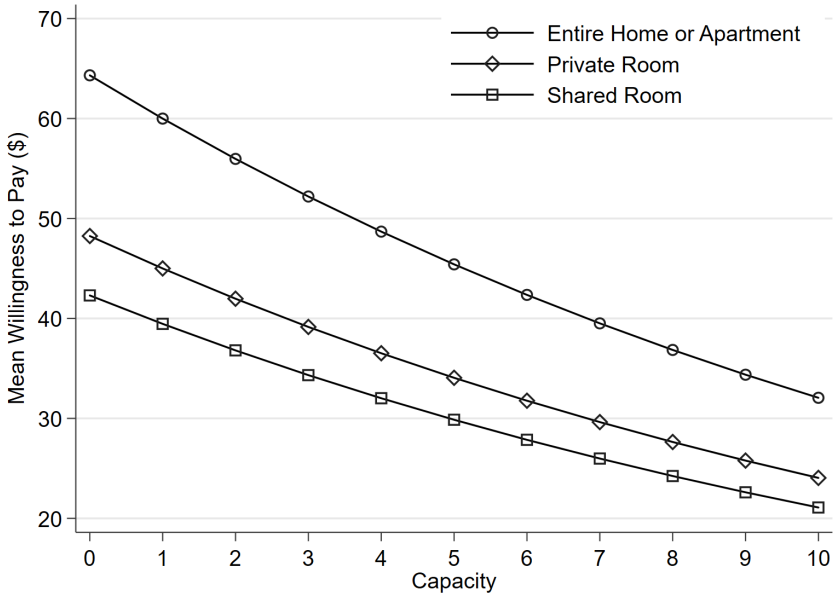


Figure C.1 – Willingness To Pay
Notes: The WTP has been averaged over all rentals and time within each rental type category.

Table C.1 – Willingness To Pay Across Rentals

	(1)	(2)	(3)	(4)
	WTP ($)			
	All Rentals		Purged Rentals	
VARIABLES	mean	sd	mean	sd
All rentals	47.70	26.61	53.38	25.63
Space-Type				
Entire Home/Apartment	54.13	26.62	53.38	25.63
Private Room	41.34	24.54		
Shared Room	34.93	26.06		
Property-Type				
Apartment/Condominium/Loft/Townhouse	48.52	25.99	53.44	25.59
House	35.99	23.16		
Bungalow/Castle/Guesthouse/Vacation home/Villa	40.29	31.47		
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	45.87	27.06		
Location				
Bronx	31.90	18.59	28.64	7.714
Brooklyn	40.04	20.23	42.11	19.98
Manhattan	55.43	27.99	58.11	26.06
Queens	36.06	27.95	36.18	15.62
All Staten Island	42.93	26.78	55.86	30.18

Notes: The average willingness to pay for the 2^{nd} unit of stay per night has been calculated for each rental and then summary statistics have been reported across rentals and by rental categories.

Table C.2 – Changes in Consumer Welfare

	(1)	(2)	(3)	(4)	(5)	(6)
	Number of rentals		CS_{actual}	CS_{cf}	$\Delta_{price}CS$	$\Delta_{total}CS$
	All	Purged	(1000$)	(1000$)	(1000$)	(1000$)
All rentals	108,212	4,127	452,489	483,290	-1,127	-30,801
Space-Type						
Entire Home/Apartment	55,790	4,127	355,564	386,102	-863.1	-30,537
Private Room	48,183	0	89,374	89,622	-248.0	-248.0
Shared Room	4,207	0	7,550	7,565	-15.35	-15.35
Property-Type						
Apartment/Condominium/Loft/Townhouse	99,782	4,122	396,520	427,278	-1,084	-30,759
House	7,025	0	50,189	50,221	-32.60	-32.60
Bungalow/Castle/Guesthouse/Vacation home/Villa	145	0	408.2	408.6	-0.403	-0.403
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	89	0	232.1	232.4	-0.296	-0.296
Location						
Bronx	1,570	32	4,986	4,986	-0.195	-0.195
Brooklyn	39,527	867	158,775	162,341	-145.5	-3,566
Manhattan	56,977	2,997	249,939	276,404	-961.6	-26,465
Queens	9,396	212	35,071	35,669	-9.446	-597.8
Staten Island	733	19	3,707	3,878	-9.802	-171.4

Notes: The consumer surplus and changes have been aggregated across all the rental-time pair by each rental category.

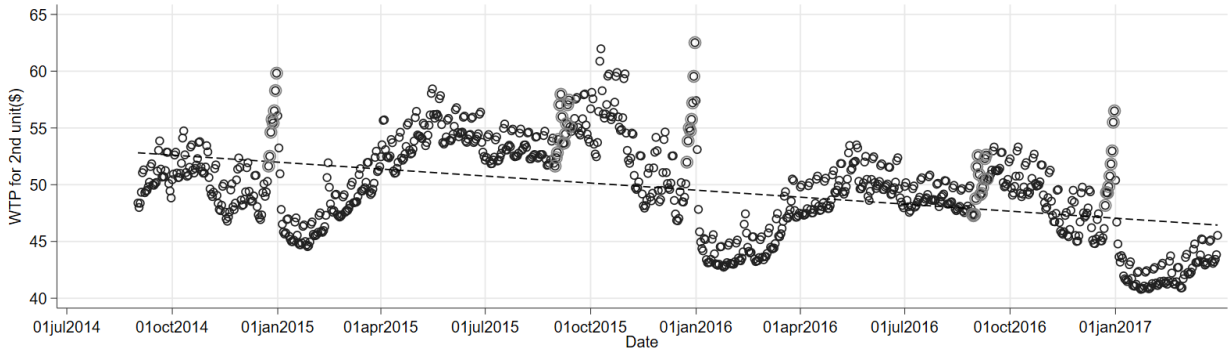
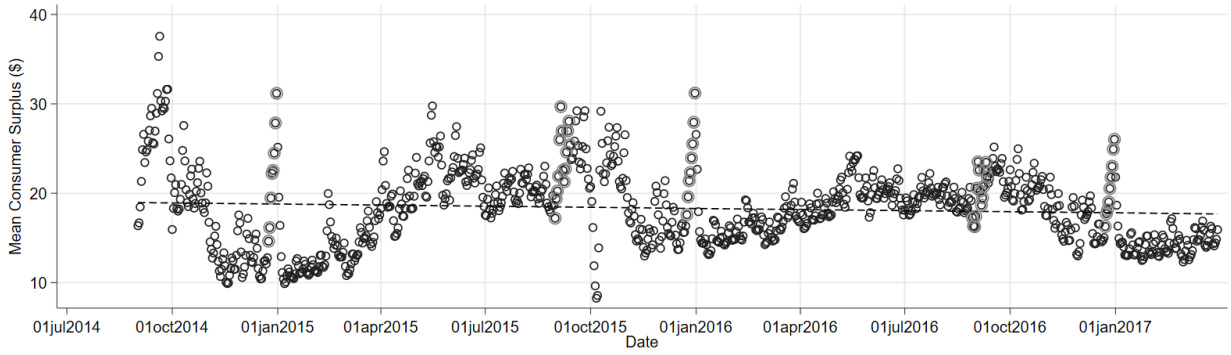
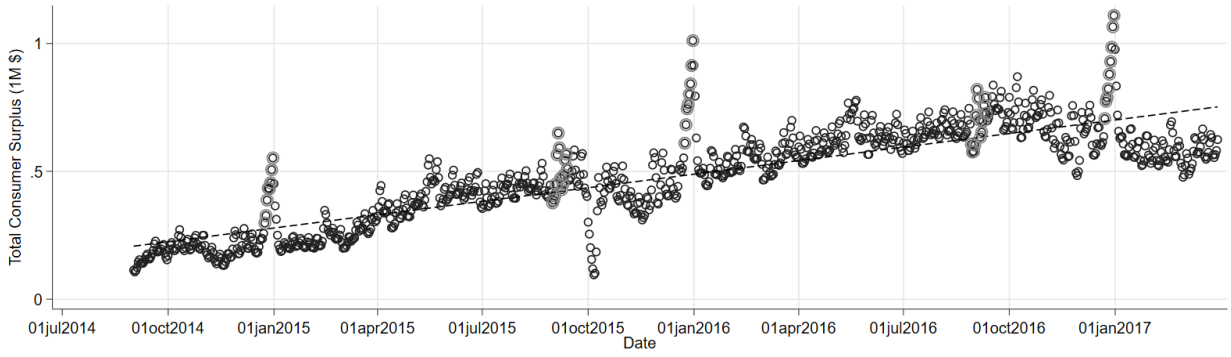


FIGURE C.2 – WILLINGNESS TO PAY ACROSS MARKETS

Notes: For each time/market, the willingness to pay for 2^{nd} unit of accommodation has been averaged and then plotted against time.



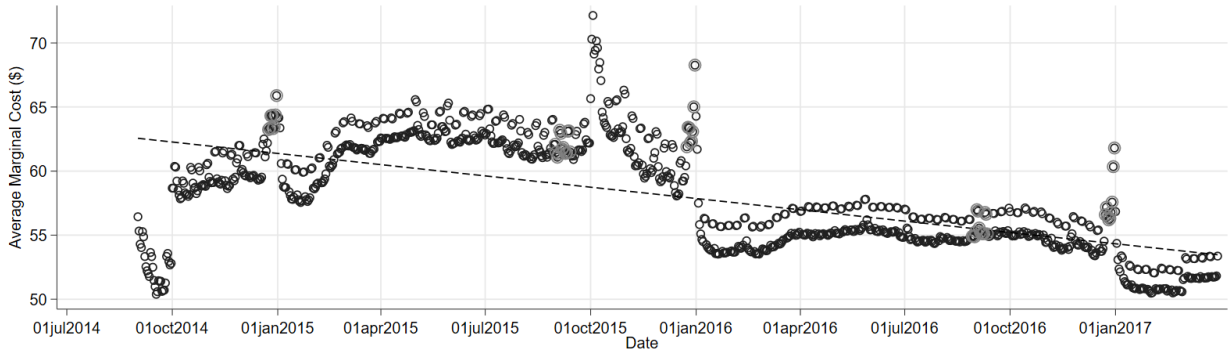
(A) Average Consumer Surplus



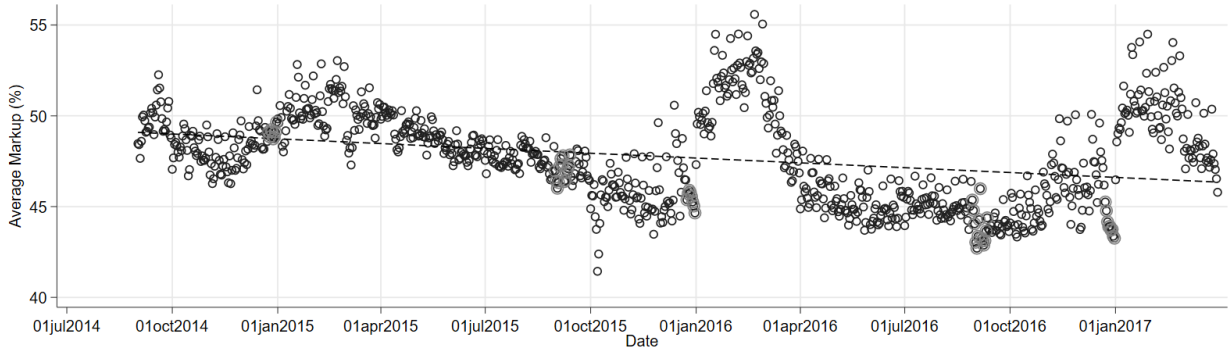
(B) Total Consumer Surplus

FIGURE C.3 – CONSUMER SURPLUS ACROSS MARKETS

C.2 Supply Side



(A) Average Marginal Cost



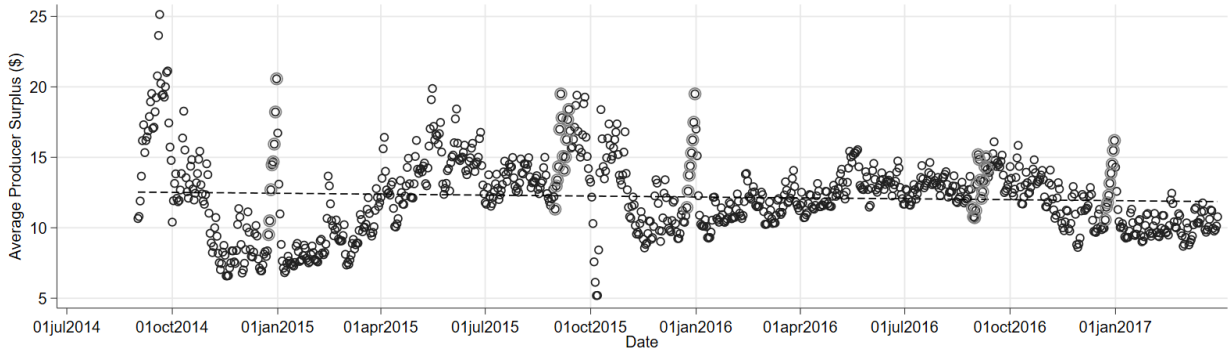
(B) Average (Sales Weighted) Markup

FIGURE C.4 – MARGINAL COST AND MARKUP ACROSS MARKETS
Notes: The markups are sales weighted where the weights are calculated for each market/time.

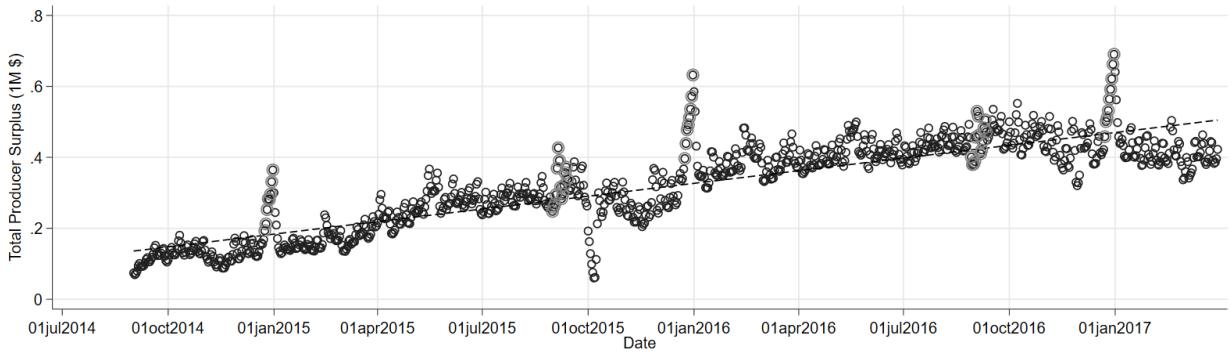
Table C.3 – Markup and Marginal Cost Across Rentals

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Marginal Cost ($)				Markup (%)			
	All Rentals		Purged Rentals		All Rentals		Purged Rentals	
VARIABLES	mean	sd	mean	sd	mean	sd	mean	sd
All rentals	55.74	59.46	58.79	52.69	16.21	18.97	28.56	23.13
Space-Type								
Entire Home/Apartment	61.74	57.42	58.79	52.69	19.50	21.14	28.56	23.13
Private Room	49.92	60.88			12.35	15.08		
Shared Room	42.09	58.15			17.00	20.16		
Property-Type								
Apartment/Condominium/Loft/Townhouse	56.93	57.34	58.85	52.69	15.74	18.56	28.52	23.11
House	36.99	45.61			22.34	22.31		
Bungalow/Castle/Guesthouse/Vacation home/Villa	47.83	63.79			22.59	21.53		
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	58.21	52.67			13.19	19.22		
Location								
Bronx	34.79	37.17	22.90	11.56	18.91	19.63	32.01	24.05
Brooklyn	44.02	43.95	42.36	35.13	17.11	18.81	30.07	23.02
Manhattan	67.14	62.68	65.44	56.77	15.24	18.91	28.09	23.21
Queens	39.96	81.76	36.64	28.19	17.70	19.49	29.07	22.45
Staten Island	49.28	45.19	66.96	49.90	18.48	19.58	22.37	19.04

Notes: The reported markups are sales weighted averages. But the sales weights are actually calculated within a rental.



(A) Average Profits



(B) Total Producer Surplus

FIGURE C.5 – PRODUCER SURPLUS ACROSS MARKETS

Table C.4 – Changes in Producer Welfare

	(1)	(2)	(3)	(4)	(5)	(6)
	Number of rentals		PS_{actual}	PS_{cf}	$\Delta_{price}PS$	$\Delta_{total}PS$
	All	Purged	(1000$)	(1000$)	(1000$)	(1000$)
All rentals	108,212	4,127	302,243	329,882	4,829	-27,639
Space-Type						
Entire Home/Apartment	55,790	4,127	243,718	272,217	3,970	-28,499
Private Room	48,183	0	53,261	52,458	803.0	803.0
Shared Room	4,207	0	5,263	5,207	56.66	56.66
Property-Type						
Apartment/Condominium/Loft/Townhouse	99,782	4,122	262,930	290,800	4,598	-27,870
House	7,025	0	35,471	35,282	189.4	189.4
Bungalow/Castle/Guesthouse/Vacation home/Villa	145	0	264.7	262.8	1.934	1.934
Boat/Cabin/Camper/RV/Hut/Lighthouse/Plane/Treehouse	89	0	165.0	163.8	1.210	1.210
Location						
Bronx	1,570	32	3,862	3,859	2.679	2.679
Brooklyn	39,527	867	109,575	112,756	734.4	-3,181
Manhattan	56,977	2,997	161,072	184,830	4,003	-23,758
Queens	9,396	212	25,077	25,703	48.82	-625.6
Staten Island	733	19	2,648	2,725	40.16	-77.51

Notes: The table presents aggregate producer surplus and the changes calculated for different categories of rentals.