THREE ESSAYS ON DEMAND ESTIMATION

By

Hee Kwon Kyung

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2019

ABSTRACT

THREE ESSAYS ON DEMAND ESTIMATION

By

Hee Kwon Kyung

Chapter 1: The Role of Reputation/Feedback Contents in NYC Airbnb Market: Evidence from Hedonic Price Regressions

Economists have found that reducing information asymmetry is crucial for online marketplaces to overcome market failure due to adverse selection. Reputation/feedback systems and multi-media web contents from sellers are known to be popular disclosure devices for this purpose. This paper employs hedonic price regressions to provide empirical evidence that the recent success of a sharing economy platform, Airbnb, also relies on such publicly available information on product quality. Machine learning selectors were employed to reduce high-dimensionality in the attribute space. To process consumer review texts and sellers' advertisement texts, word/phrase extraction and sentiment analysis were introduced. I propose a GMM estimation to produce more accurate implicit price estimates, that was designed to control for time-varying unobservables. 'Superhost' designation by the platform and consumer reviews showed greater impacts than seller side advertisement texts.

Chapter 2: Demand Estimation for NYC Airbnb Market: Value of Reputation/Feedback Contents and Voluntary Disclosures

The success of online marketplaces has often been attributed to reputation/feedback systems, in that they reduce adverse selection due to information asymmetry by disclosing enforced or verifiable ex-post information on product quality. This paper tries to quantify the value of such information content in NYC Airbnb market with a newly constructed dataset containing the actual 708,308 vacation rental reservations from Airbnb tourists. A three level nested logit model was employed to capture consumers' choice set formation behaviors during web search on the platform

using Google Maps API. High-dimensional attribute space due to extreme product heterogeneity necessitates variable selection using machine learning methods based on sparsity assumption. Though model selection procedures by LASSO and exact inference for post selection parameter estimates were proposed, structural modeling and endogeneity control turn out to be essential for successful identification. Text processing techniques were introduced to extract variables from sellers' advertisement texts and consumer reviews. The results confirm a key insight from information economics: enforced quality certifications and ex-post verified consumer reviews generate greater welfare impacts than non-verified seller side voluntary disclosures.

Chapter 3: Estimation for the Distribution of Random Coefficients with Heterogeneous Agent Types: Monte-Carlo Simulation

This paper is a simple Monte-Carlo extension for Fox, Kim, Ryan, and Bajari (2011), which gives a direct estimator for the distribution of random coefficients in diverse settings including logit demand models. The estimator is a simple inequality constrained least squares, and this study examines its behaviors given there are hundreds of consumer types, which could be an interesting case for various marketplaces. High-dimensional metrics are then introduced to reduce the dimensionality of design matrices the rank of which is the number of consumer types. The approximation performances to the cumulative distribution of random coefficients of such post lasso estimators are compared to those of baseline estimator.

ACKNOWLEDGEMENTS

I am deeply grateful for the guidance from my chair Professor Kyoo Il Kim and committee members, Professor Joseph A. Herriges, Distinguished Professor Peter Schmidt, and Professor Seunghyun Kim. Despite of my countless shortcomings, they have endowed me with their precious time and seasoned wisdom. Up to this date, I am not sure if I deserve such an honor. I also thank all the faculty members and staffs of the Economics Department at Michigan State University.

Born and raised in a perimeter of civilization at least in terms of knowledge, I enjoyed not only the PhD program but also my exposures to American culture, “the shower of enlightenment” I would say. The years spent here will remain as one of the best times I have ever had in my life.

Lastly, I thank my parents, sister, brother-in-law, and the little nephew for their love and trust.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

THE ROLE OF REPUTATION/FEEDBACK CONTENTS IN NYC AIRBNB MARKET: EVIDENCE FROM HEDONIC PRICE REGRESSIONS

1.1 Introduction

The explosive growth of Airbnb and other sharing economy platforms in the last decade begs a question; how could they build trust among total strangers over one-time transactions despite the theoretically expected market failure due to information asymmetry? (Akerlof (1970)) The P2P (Peer-to-Peer) platform accommodated more than 100 million parties of tourists and the market value topped at $31 billion as of 2017. One common insight on the success of Airbnb over various disciplines including tourism, marketing, and economics is that the reputation/feedback systems and information contents such as texts and photographs of rental units and hosts reduced information asymmetry, thus facilitating trust among market participants. (Guttentag (2015), Horton and Zeckhauser (2016), Ert, Fleischer, Magen (2016), Fradkin, Grewal, and Holtz (2018), and Liang, Schuckert, Law, and Chen (2017))

Indeed, it is one of the foundational ideas of classical information economics that a market provider or seller could partially contract on product quality by disclosing ex-post verifiable or enforced information such as warranties and insurances. (Grossman and Hart (1980), Grossman (1981), and Milgrom (1981)) Besides, repeated transactions and disclosure of reputation/feedback repository (history) to all potential buyers discipline sellers to act honestly in a future transaction with a total stranger, which is a particularly enlightening lesson for sharing economies and online retail outlets. (Kreps (1982, 1990), Tadelis (2016) and Milgrom, North, and Weingast (1990))

This paper argues that for NYC Airbnb market, disclosure of platform enforced and ex-post verified information contents on product quality also neutralizes the initial information asymmetry and prevents market failure. To test this hypothesis, I employ hedonic price regression and present empirical evidence that the quality certification 'Superhost' badge, host identity verification measures, and consumer review ratings and texts are more influential to the transaction prices than non-verified seller side advertisement texts. ('Cheap Talk')

The drive behind pursuing a question verified in various online marketplaces is twofold. First, unlike already successful online P2P markets for material goods, Airbnb is an intermediary for service products which involves a different type of risks over monetary losses. In fact, there have been many unfortunate incidents for Airbnb customers: infringements of privacy by hidden cameras, physical attacks by hosts, and loss of time and pleasure due to deceptive web listings. It is worth investigating how Airbnb could be successful with such a risk of adverse selection. Second, this paper offers several methodological tools to deal with a set of identification challenges modern online platform data presents.

One distinctive feature of P2P online platform data is extreme product heterogeneity. It does not refer to the fact that consumers now have access to many products once bought and sold in traditional offline shops, thanks to online retail giants like Amazon. The extreme product heterogeneity of particular interest in this paper is uniqueness in each of the numerous products that have never been on markets. For example, in Etsy.com buyers can shop custom made apparel, crafts, toys, and 3d printer blueprints without any big brand names printed on. Taskers in Taskrabbit.com have no idea what kind of problems they are going to solve until the customers specify them. Likewise, potential Airbnb guests are staying in a house of someone they never met before.

The first identification challenge due to such extreme product heterogeneity is the high-dimensional characteristic (attribute) space. The platforms need to define and differentiate each product so that they can match and sell it to a consumer with specific preferences. Airbnb rental units consist of various types of personal properties that have never been publicly offered as travel accommodations, including cabins, castles, farm barns, boats, and tree houses. Customers have a choice over a new set of amenity and service features such as baby beds, children's dinnerware, EV chargers, and video game consoles. As a result, Airbnb data attaches more than 150 binary indicators to each rental unit for consumers to satisfy their heterogeneous preferences.

A high-dimensional dataset with numerous binary indicators poses two serious problems. Some attributes are common and some are scarce, causing multicollinearity. There are irrelevant attributes that do not affect consumers' purchase decisions significantly either in an economic or statistical sense. The hedonic price regression could thus suffer biases in implicit price estimates or misspecifications. A variable selection procedure for efficiently reducing dimensionality is necessary.

This paper proposes to adopt sparsity assumption and use variable (model) selection based on machine learning methods for choosing a subset of attributes that explain the variation in transaction prices the best. In fact, economists have been resorting to machine learning techniques to cope with high-dimensional data plagued by numerous collinear regressors, nuisance variables, and instruments over various research areas: demand estimation, program/policy evaluation, treatment effects, and general linear models. (Bajari, Nekipelov, Ryan, and Yang (2015), Belloni, Chernozhukov, Fernandez-Val, and Hansen (2017), Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), Chernozhukov, Hansen, and Spindler (2015))

The second identification challenge from product heterogeneity is how to process product information contained in unstructured formats. Airbnb hosts voluntarily disclose information on product quality in texts and images. Through texts, a host extols the merits of a rental unit and neighborhood: nearby tourist attractions, transportation logistics, restaurant recommendations, and house rules guests should abide by. They characterize the identity of each unique product that cannot be transmitted via numerical variables or search filters (binary indicators), letting sellers and platforms better differentiate each good/service from another. Buyers can also voluntarily disclose information by review texts. Reviews often reveal product information from the perspective of past customers, giving more individuality to products. Airbnb lets only the actual guests write reviews on the rental units they visited, so review texts are considered to be ex-post verified information. Photographs of rental units and textual advertisements from sellers can be deemed as non-verified.

Such information contents in unstructured formats should be incorporated into hedonic price regression model because they affect consumers' valuation and choices. This paper proposes to generate numerical variables from text data using widely accepted processing techniques: extracting keywords/phrases and sentiment analysis on consumer reviews. To be more specific, the frequency of appearance of certain words/phrases and the number of reviews that were classified as negative by a supervised machine learning will be used as additional product attributes. Image processing is beyond the scope of this paper, but represents another source of product information for consumers.

Tourism and hospitality researchers have been adopting textual analysis on guest reviews to capture consumer sentiments. The two most dominant approaches are tokenization of a review text into words/phrases and machine learning classification or prediction for emotional polarity of a review.[1] Economists also rely on textual analysis in hedonic price studies on various online marketplaces such as eBay Motors and real estate. (Lewis (2011) and Nowak and Smith (2017))

The third and final identification issue is endogeneity due to unobserved/omitted variables, which is a pervasive problem for hedonic price regressions. For Airbnb data, it is often hard to explain why a consumer chose a specific rental property among thousands of others even with high-dimensional attributes and unstructured format descriptions. There could be time-fixed unobservables such as seasonality in travel accommodation demand/supply and neighborhood traits like crime rates and education levels. Endogeneity can easily be dealt with fixed effects estimation if there are only time-fixed unobservables.

However, heterogeneity in consumer tastes and diverse individual schedules/itineraries are also candidates for unobservables that are time varying and correlated with observed attributes. This paper finds a strong evidence for omitted variable bias in preliminary OLS hedonic regressions and employs a fixed effects estimation to control for time-fixed unobservables. Then with an additional identifying assumption for time-varying correlated unobservables, I employ a GMM estimation. For all error structures and estimation methods, estimation results confirm our hypothesis that enforced and ex-post verified reputation/feedback signals are more influential than non-verified seller side information contents.

This paper adds value to the existing simple OLS hedonic price research for Airbnb by dealing with the identification challenges listed above. Wang and Nicolau (2017), Chen and Xie (2017), Teubner, Hawlitschek, and Dann (2017), Gibbs, Guttentag, Gretzel, Morton, and Goodwill (2018)

[1] See Alaei, Becken, and Stantic (2017) for a comprehensive review of methods and literature up to date.

investigated 33 cities across U.S, Austin in Texas, 86 cities in Germany, and five metropolitan areas of Canada, respectively. For hotel booking websites, Ye, Law, Gu (2009), Ye, Law, Gu, and Chen (2011), Ogut and Tas (2012), and Xie, Zhang, and Zhang (2014) regress the number of room nights sold on review ratings.

My research extends previous literature in economics and marketing/management on the role of reputation/feedback mechanisms and voluntary disclosures in reducing information asymmetry to NYC Airbnb case. Economists found that eBay's 'Superseller', seller rating scores, and text/photo descriptions have causal relationships with transaction prices. (Resnick, Zeckhauser, Swanson, and Lockwood (2006), Houser and Wooders (2006), Jin and Kato (2006), and Lewis (2011)) Marketing/management researchers investigated the role of online consumer reviews in business outcomes in various markets for books, movies, music, retail, video games, and electronics. (Chavalier and Mayzlin (2006), Liu (2006), Dellarocas, Zhang, and Awad (2007), Duan, Gu, Whinston (2008), Chintagunta, Gopinath, and Venkataraman (2010), Dhar and Chang (2009), Floyd et al. (2014), Cui, Lui, and Guo (2012), Ghose and Ipeirotis (2011))

The rest of this paper is organized as follows. Section 1.2 revisits Airbnb in terms of information asymmetry, presents data and explains processing details. Section 1.3 introduces identifying assumptions, preliminary OLS analyses, and estimation methods to control for unobserved variables. Section 1.4 reports and discusses estimation results. Section 1.5 concludes.

1.2 Airbnb and Data

1.2.1 Trade among Anonymous Sellers and Buyers

1.2.1.1 NYC Airbnb Market

Since 2016, more than 60 million tourists visit NYC annually including 20% of international arrivals. Average length of stays is about four days and the number of potential accommodation reservations totals 15 million. Without any possession of commercial real estates, Airbnb received more than a million reservations in 2017.[2] A survey on 4,000 people asking previous experiences with and intentions to use Airbnb rental services found that as of November, 2015 the market share of Airbnb is occupying 12% of leisure and business travelers. This share was expected to rise to 16-18% in 2016.[3] Another report states that the room nights share amounts to 8% compared to hotels as of August, 2015.[4]

Such successful entry and robust trading volumes seem quite surprising given that potential Airbnb guests choose to stay at a total stranger's housing unit relying solely on the information showing on computer screens. The web page content of each rental property then can be considered as the contract between buyers and sellers. Potential guests take most of the product information on web listings at face value. This paper focuses on how the platform generates such consumer trust and which information categories would look trustworthy and how much can be trusted in the eyes of consumers.

1.2.1.2 Reputation/Feedback Repository

According to the disclosure model of Grossman and Hart (1980), Grossman (1981), and Milgrom (1981), buyers should update their expectations for product quality based only on enforced or verifiable ex-post information contents. For Airbnb case, the quality certification 'Superhost' badge and actual guests' review ratings and texts apply to the categories. Both could be called as credible reputation/feedback repositories in that only actual guests can leave review ratings and texts and Airbnb enforces strict service quality criteria based on past performances in designating 'Superhost'. A host must have accommodated more than ten parties of guests, maintained a 90% response rate to booking requests or higher, received a five star review - review scores higher than 80 out of 100 - at least 80% of the time, and completed each of confirmed reservations without canceling.

[2] The visitor poll is from NYC & Company. Average length of stays was calculated from the actual booking records for hotels and Airbnb obtained from Expedia.com and Airdna, a data consulting branch firm of Airbnb, respectively.

[3] Who Will Airbnb Hurt More - Hotels or OTAs (Online Travel Agency)?, JP-Morgan's Global Insight

[4] Airbnb and Impacts on the New York City Lodging Market and Economy, Hospitality Valuation Services

The existence of these well-functioning public reputation/feedback repositories forces Airbnb hosts to provide present and future guests with accommodation services as specified by the web listings, even if the hosts will never meet them again. It is because positive feedback from customers of the past would reward the hosts in the future businesses with anonymous tourists to come and negative feedbacks would do the opposite. (Kreps (1982, 1990) and Tadelis (2016)) One clear empirical implication of this powerful insight would be that 'Superhost' badge, higher rating scores, and textual variables with positive sentiments from Airbnb tourists will attract more future guests and enable hosts to obtain price premiums.

1.2.1.3 Basic Contract Enforceability and Voluntary Disclosure from Sellers

A set of ground rules and coordination schemes for safe transactions are fundamental prerequisites for offline/online marketplaces since at least medieval trade fairs in Europe. (Greif (2006) and Milgrom, North, and Weingast (1990)) Reputation/feedback systems of Airbnb also rely on basic contract enforceability and consumer protection measures, including government issued identifications for both sellers and buyers, payment holdings by escrow during the first 24 hours after check-in, full/partial refunds, and dispute resolutions including providing alternative accommodations.

Together with such trust-enhancing apparatus, the idea that sellers have incentives to differentiate themselves from others now makes it possible for consumers to believe voluntary disclosure such as rental unit descriptions in text and image formats, not whole-heartedly but certainly at some level. For example, positive adjectives every hosts could say such as “nice”, “comfy”, and “best in New York City” would not appeal that much to consumers, whereas texts explaining locational merits like “5 min walk from Central Park” would appeal to consumers given that Google Maps API on each listing webpage allows potential guests to check the validity of such statements almost immediately.

Accommodation capacities such as the number of default guests, accompanying guests, bedrooms, bathrooms, and beds and binary filters indicating various amenitiy and service features are presented in standardized visualized items on each listing webpages or, search filter menus. They are comparable to hotel booking portals making it natural to include as product attributes in the hedonic price regression models.

1.2.2 Data

1.2.2.1 Summary Statistics

The data source is InsideAirbnb.com, a public repository of rental unit prices, attributes, and review texts run by Airbnb. The panel dataset of this paper consists of two time periods. For each time period, three cross sections of NYC Airbnb rental unit data were stacked together: June, August, and December recorded each in 2016 and 2017, respectively. Summers and Decembers are the peak time periods for both demand/supply for NYC vacation rental businesses. A year gap is used to control for seasonality in fixed effects and GMM estimation with time-varying correlated unobservables. OLS estimation is conducted on pooled samples stacking cross sections together. Rental units with extreme prices at both 0.1% outer margins and units without any reviews were truncated.

Table 1.2 reports summary statistics for the cross section recorded in December 2017. Prices are pre-tax transaction rental prices which mean listing prices per night plus cleaning fees. All variables in Table 1.2 except for some review score categories (due to the rating inflation and collinear relationships, as will be explained shortly) were in fact selected by a lasso variant (Belloni and Chernozhukov (2013)) designed to achieve a successful asymptotic approximation to the objective (prices) with only a subset of all 180 variables. It is to efficiently reduce high-dimensionality in attribute space due to extreme product heterogeneity.[5] There exist clear differences between 'Superhost' units and others. The mean price is higher for 'Superhost' units by about $10 to $12. 'Verification Accounts' mean the number of contact methods a host maintains, for example phone, email, facebook, google, and other social media accounts. The fact that a host has multiple accounts

[5] Section 1.3 explains the methodology for the lasso variable (model) selection, performance of post selection OLS estimates, and exact inference for them. The lasso selection was conducted on the OLS dataset, containing all the cross sections recorded at June, August, and December for both 2016 and 2017.

implies that Airbnb has verified GPS coordinates for the location, government issued identification, and photographs of the host on the listing webpage. 'Verification Accounts' can thus be considered as a proxy for the degree of contract enforceability a host represents.

For all review score categories, 'Superhost' units have higher averages. However, review scores are extremely skewed toward left. Review score inflation in Airbnb market in fact has been extensively investigated by Zervas, Proserpio, and Byers (2015), Proserpio, Xu, and Zervas (2016), and Fradkin, Grewal, and Holtz (2018). Compared to other vacation rental portals, rating scores of Airbnb tend to be higher. For example, the average 5 star rating ('Overall Rating') for TripAdvisor.com vacation rentals is 3.8/5 and for Airbnb, 4.75/5. Reciprocity due to bilateral review policy and sellers' strategic manipulation were proposed to be the main causes for rating inflation. The empirical implication is that price premiums due to a unit increase in each rating scores would be small. Also, given that every rating scores are near perfection, it is important to catch which categories would appeal to consumers the most. The lasso procedure chose 'Cleanliness', 'Location', and 'Value'.

Table 1.1: Definitions for Review Score Categories

Category	Questions Asked in Reviewing Process
Rating	Overall experience
Accuracy	How accurately did the photos and description represent the actual space?
Cleanliness	Did the cleanliness match your expectations of the space?
Check-in	How smooth was the check-in process, within control of the host?
Communication	How responsive and accessible was the host before and during your stay?
Location	How appealing is the neighborhood (safety, convenience, desirability)?
Value	How would you rate the value of the listing?

The number of negative reviews is about twofold for 'Superhost' units, compared to normal hosts' units but the difference is due to the fact that, on average, 'Superhost' units received two times as many consumer reviews. Reviews on 'Superhost' units also contain more positive phrases expressing a strong satisfaction enough to write recommendations to future guests or intentions to come back.[6]

[6] More details on text processing is provided in the next sub-subsection 1.2.2.2

Table 1.2: Summary Statistics

(Cross Section: 201712)	Mean	S.D.	Superhost	Normalhost	Min	Max
Number of Obs	13,364		2,532	10,832		
Price ($)	185.4537	119.439	195.2812	183.1565	34	1000
QUALITY CERTIFICATION						
Superhost Indicator	0.1895		1	0	0	1
Verification Accounts	4.5934	1.3273	4.8175	4.5410	1	9
REVIEW SCORES						
Overall Rating	93.1792	7.2012	96.4680	92.4105	20	100
Accuracy	9.5657	0.7368	9.8776	9.4928		
Check-in/out	9.2114	1.0106	9.7149	9.0937		
Cleanliness	9.7456	0.6018	9.9645	9.6944	2	10
Communication	9.7722	0.5659	9.9664	9.7268		
Location	9.4192	0.7637	9.5170	9.3963		
Value	9.3294	0.7874	9.6584	9.2525		
REVIEW TEXT						
Negative Reviews	4.9400	6.4968	9.1078	3.9658	0	76
Positive Phrases	3.4892	4.9246	6.5687	2.7693	0	48
SELLER TEXT						
Positive Adjectives	6.5965	4.6933	7.4214	6.4036	0	39
Location Phrases	10.0342	7.0098	11.3468	9.7274	0	62
ACCOMMODATION CAPACITES						
Default Guests	2.9461	1.8063	3.0865	2.9133	1	16
Bedrooms	1.1675	0.6963	1.2014	1.1595	0	4.5
Bathrooms	1.1143	0.3626	1.1145	1.1142	0	9
Beds	1.6051	1.0757	1.6904	1.5852	1	16
Guests Included	1.6192	1.1364	1.7749	1.5828	1	14
AMENITY AND SERVICES						
Air Conditioning	0.8836		0.9356	0.8714		
Buzzer Wireless Intercom	0.5186		0.4897	0.5254		
Cable TV	0.3631		0.4313	0.3471		
Free Parking	0.1076		0.1445	0.0990		
Indoor Fire Place	0.0387		0.0474	0.0367		
Lock on Bedroom Door	0.1408		0.1829	0.1310	0	1
Cats Allowed	0.0648		0.0746	0.0625		
Internet	0.7710		0.7670	0.7720		
Shampoo	0.6207		0.7753	0.5846		
(Room Type)						
Entire Home/Apt	0.5622		0.5391	0.5676		
Shared Room	0.0137		0.0138	0.0137		

Seller texts include rental unit titles, sub-titles, descriptions on various aspects such as neighborhoods, transportation, pros and cons, and etc. 'Superhost' rentals in fact contain fewer 'Positive Adjectives' and 'Location Phrases' than normal host units. The selected accommodation capacities conform to the empirical studies on price determinants of hotels and Airbnb.[7] Amenity and service features were cross-selected by additional data-driven machine learning methods other than Belloni and Chernozhukov (2013).

1.2.2.2 Text Processing

This paper employs n-gram word/phrase extraction (bag of words) and sentiment analysis (classification) using a supervised machine learning method to process seller and buyer texts. N-gram bag of words means extracting words/phrases purely according to the frequency of occurrences and use them as regressors. As shown in tables, selected features are often reduced and categorized at a researcher's discretion.

Table 1.3: Selected Words/Phrases from Airbnb Hosts' Advertisement Texts

Category	Positive Adjectives	Location Words
Unigram	Amazing, Beautiful, Cozy, Friendly, Spacious, ...	Broadway, Manhattan, Soho Brooklyn, Chelsea, ...
Bigram		Central Park, Columbia University, Hell's Kitchen, Brooklyn Bridge, Times Square, Union Square, Walking Distance, Rockefeller Center, ...
Trigram		Empire State Building, The G train, Major subway lines, Grand Central Station, The Hudson River, ...
Quadrigram		Metropolitan Museum of Art Museum of Natural History, ...

This paper extracts 36 words/phrases out of the 3,000 most frequently appearing ones among more than 120,000 seller advertisement texts. The counts for each rental unit were summed over the regarding two categories: 'Positive Adjectives' and 'Location Words' (Table 1.3). Similarly,

[7] See Wang and Nicolau (2017) for a comprehensive review up to date.

38 'Positive Phrases' expressing recommendations for future guests and intentions to revisit were selected out of 7,000 most frequently appearing words/phrases among more than 850,000 Airbnb guest review texts (Table 1.4).

Table 1.4: Selected Words/Phrases from Airbnb Guests' Review Texts

Recommendations	Intentions to Revisit
can highly recommend	cant wait to come back
can recommend	cant wait to go back
i definitely recommend this	hope to be back soon
i really recommend this place	hope to come back
i recommend	hope to see you again
id recommend	hope to stay here again
we recommend	hope to stay there again
will recommend	id definitely stay here again
would absolutely recommend	...
would definitely recommend	would definitely come back
would highly recommend	would definitely consider staying here again
would not hesitate to recommend	would stay there again
would recommend	wouldnt hesitate to stay here again

Supervised machine learning means fitting a function that maps an input to an output based on an example input-output pair dataset. For sentiment classification of review texts it includes the following procedures; a researcher conducts a pre-processing such as removing non-alphabetical components, (arabic numbers, commas, punctuations, and etc) trimming white spaces, and converting to lower case letters. A classification machine is then trained on sample reviews, with emotional polarity as outputs and words/phrases as inputs. The choice on word/phrase regressors could either rely on pre-established dictionaries (lexicons) or n-gram words/phrases from sample reviews whichever yields the best in-sample prediction performances. The trained machine is then scaled up on the whole review text corpus.

This paper trains a classification machine on a set of 1,000 sample reviews collected from four major U.S cities other than NYC: Ashevill (NC), Austin (TX), Denver (CO), and Washington D.C. Using 3,500 n-gram bag of words/phrases, multiple supervised machine learning models were constructed and the highest in-sample prediction rate (87%) was achieved with classification tree in 'Caret' R package over Naive Bayes and Support Vector Machine. The classification was then applied to the whole 850,000 NYC Airbnb review texts. Table 1.5 and Table 1.6 provide a conceptual example.

Table 1.5: Example Guest Reviews and Sentiment Labels

Reviews	Raw Texts
Ex.1	This is a ***dirty*** frat house.
(Negative)	No locks other than main building door.
	Dirty toilets. No host present. ***Rotting*** food in the fridge.
Ex. 2	My room at the BPS Hostel was ***clean*** and ***cool***.
(Nonnegative:	The staff and fellow guests were friendly and ***helpful***.
Neutral)	The location is very convenient for local eateries, coffee shops, pubs and deli's.
	However, I do not feel it was good value for money at $72 per day.
	There was no room service,
	I shared a bathroom with upto 8 others and the ***breakfast*** was weak.
Ex. 3	***Great*** location just outside of downtown Asheville.
(Nonnegative:	I stayed here with three other people. ***Plenty*** of space.
Positive)	Mike was very easy to work with, and made sure we had everything we needed.

Table 1.6: Bag of Words Matrix for Example Guest Reviews

Label		breakfast	clean	cool	dirty	great	helpful	however	plenty	rot	
Ex.1		0	0	0	2	0	0	0	0	1	
Ex.2	...	1	1	1	0	0	1	1	0	0	...
Ex.3		0	0	0	0	1	0	0	1	0	

N-gram bag of words and sentiment analysis by supervised machine learning have been the dominant processing techniques in hospitality research on the impacts of consumer reviews on prices and business performances, over many platforms such as Ctrip.com, TripAdvisor.com, Booking.com, Expedia.com, Travel.yahoo.com, and Yelp.com. Also, supervised machine learning showed better prediction results than lexicon based methods. Existing lexicons do not share many of the words/phrases on a specific web portal of interest.[8]

Following the conventional approach of using counts of words/phrases and negative reviews as covariates instead of proportionate variables with scaling purposes was done for two reasons. First, review texts are time-cumulative and superimposed so it is unlikely for consumers to read all

[8] See Alei, Becken, and Stantic (2017) for a comprehensive review on sentiment analysis methods listed above and performances over online hospitality research in the last decade.

the reviews. If one uses the proportion of negative reviews (from total number of reviews) instead of the counts, it suffers the risk of downward (upward) bias of the coefficient for rental units that received a great (small) number of reviews already. Second, one can imagine the proportion of ’Positive Adjectives’ or ’Location Words’ among total number of words sellers’ advertisement texts include. But high ratio does not necessarily mean more value. Individual specific perceptions and expectations on product quality induced from such texts could involve further considerations on various unobserved factors.

1.3 Model and Identifying Assumption

1.3.1 High-Dimensional Metrics

1.3.1.1 Sparsity and Variable Selection

Following Rosen (1974), equation (1.1) presents the baseline hedonic regression model, which expresses prices as a function of observed attributes and errors (unobserved variables).

$$log(p_{it}) = \alpha + \beta X_{it} + \epsilon_{it} \tag{1.1}$$

p_{it} is per night rental prices of unit $i = 1, ..., n$ at time period t, $X_{it} \in R^p$ represents rental unit attributes, and ϵ_{it} is the error term. The first identification challenge with NYC Airbnb data is high-dimensional attribute space plagued by multicollinearity and irrelevant variables. This paper hence adopts sparsity that is frequently assumed in high-dimensional metrics i.e., that there exist $s = o(n) \ll p$ attributes that asymptotically capture most of the impacts of all p regressors.

A practical implication of sparsity for general linear regression models with Gaussian or heteroskedastic errors is that an econometrician first chooses a set of s variables that affects prices the most by lasso and then conducts an OLS only with the s variables. Such OLS post lasso, with theoretically suggested conditioning parameters for the first step lasso selector achieves a successful asymptotic approximation to the ’true’ $log(p_{it})$ objective function. (Belloni and

Chernozhukov (2013), Chernozhukov, Hansen, and Spindler (2015), Belloni, Chernozhukov, and Wang (2014))

More specifically, the typical risk minimization problem of balancing bias and variance for hedonic price estimation with sparsity assumption can be stated as the following.

$$\min\ c_s^2 + \sigma^2 \frac{s}{n} \tag{1.2}$$
$$c_s^2 = \min_{dim(\beta)\leq s} E[(log(p_{it}) - \beta X_{it})^2]$$

$c_s^2 + \sigma^2 \frac{s}{n}$ is the upper bound of the risk for the best log price estimator using only $s \ll p$ covariates. (the 'oracle risk'), which is achieved if the first stage lasso selector chose the correct s variables which by sparsity assumption that captures the most of the impacts of all p regressors. Then the resulting 'oracle rate' of error convergence rate is given by $\sqrt{s/n}$.

One important appeal of OLS post lasso is that even if lasso selector gives only a subset of s covariates, post selection OLS estimator still achieves the 'near oracle rate' of $\sqrt{s * log(p)/n}$. In other words, $\sqrt{E[(log(p_{it}) - \beta^{\hat{M}} X^{\hat{M}})^2]} = O_p(c_s + \sigma\sqrt{s * log(p)/n})$, where $X^{\hat{M}}$ and $\beta^{\hat{M}}$ represent the vector of attributes chosen by lasso (the observed selected model $\hat{M}$) and the corresponding post selection OLS coefficients.

1.3.1.2 Preliminary Analysis: OLS post Lasso and Post Selection Inference

To achieve the 'near oracle property' of OLS post lasso, lasso procedures need to use theoretically imposed conditioning parameters. Borrowing notations from Belloni and Chernozhukov (2013), the lasso selector based on sparsity assumption chooses variables with non-zero coefficients in solving the following penalized regression problem;

$$\hat{\beta} = argmin_{\beta \in R^p}\, \hat{Q}(\beta) + \frac{\lambda}{n}||\hat{\Psi}\beta||_1 \tag{1.3}$$
$$\hat{Q}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(log(p_{it}) - \beta X_{it})^2$$

where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ and $\hat{\Psi} = diag(\hat{\psi}_1, ..., \hat{\psi}_p)$. The theoretically suggested penalty loadings $\hat{\Psi}$ and penalty level λ for heteroskedastic errors are;

$$\hat{\psi}_j = \sqrt{\frac{1}{n}\sum_n (x_{ij}^2 \hat{\epsilon}_i^2)} \tag{1.4}$$

$$\lambda = 2c\sqrt{n}\Phi^{-1}(1 - \gamma/(2p))$$

where Φ denotes the cumulative standard normal distribution and $\hat{\epsilon}$ is an empirical estimate of errors (residuals). The suggested preset values for c and γ are 1.1 and 0.1. $\hat{\Psi}$ and λ for homoskedastic errors result in similar variable selection results.

If one proceeds to OLS with the selected (observed) model $\hat{M}$ with variables of non-zero coefficients from equation (1.3) however, then classical inferences (confidence intervals and p-values) on $\hat{\beta}^{\hat{M}}$ are no longer valid. It is because of the non-selected (omitted) variables, making the post selection OLS only with the attributes in $X^{\hat{M}}$ biased. Though the asymptotic distribution of lasso coefficients for our case of $n \gg p$ is available, (Knight and Fu (2000)) an exact post selection inference for OLS post lasso is the primary target of interest.

Such 'post selection inference' after variable selection with machine learning is a relatively new and still developing area. This paper follows Lee, L. Sun, Sun, and Taylor (2016) which provides an exact distribution of post selection OLS estimates and hence, exact confidence intervals, p-values, and tail areas. The idea is that given a response $y \sim N(\mu, \sigma^2 I_n)$, the model selection event $\{\hat{M} = M\}$ by lasso can be expressed as a form of polyhedron $\{Ay \leq b\}$. Then $\{Ay \leq b\}$ once again can be transformed into an interval with low and upper endpoints being functions $\nu^-(z_j)$ and $\nu^+(z_j)$ of residuals z_j of y in the direction of x_j, $\{\nu^-(z) \leq y \leq \nu^+(z)\}$. Due to the independence between y and z_j, the distribution of an individual coefficient $\hat{\beta}_j^{\hat{M}}$ (a simple linear transformation of y) from OLS conditional on the model selection results $\hat{M}$ is a truncated normal.

One advantageous fact about Lee et al. (2016) is that a practitioner can produce exact confidence intervals and p-values with a fixed penalty parameter λ'. To be more specific, the lasso formulation for the exact post selection inference is the original lasso by Tibshirani (1996).

$$\hat{\beta} = argmin_{\beta \in R^p}\ \hat{Q}(\beta) + \lambda' ||\beta||_1 \tag{1.5}$$

Therefore, with a range of values of λ' that produces the same model $\hat{M}$ including variables of non-zero coefficients from the penalized regression problem in equation (1.3), a practitioner can produce exact inference for OLS post lasso, preserving the 'oracle' property.

Table 1.7 and Table 1.8 (the first column) report the OLS post lasso estimation results on the pooled sample stacking cross sections recorded at 2016 and 2017. The lasso selection procedure for the 'oracle' rate (equation (1.3)) considers all variables in the dataset, to preliminary check this paper's idea: platform enforced quality certifications and consumer review contents are more influential to prices than sellers' disclosures. The variables pertaining to the information contents from the platform, buyers and sellers were indeed all selected.

The selected model is stable over a range of the penalty control parameter c, from 0.9 to 1.3 with 0.05 increments. Also, the data-driven lasso (equation (1.5)) selects the model with a range of λ' values and the exact inferences for post OLS estimates were produced. The confidence intervals for parameter estimates essentially reproduces those of OLS, but are slightly wider for most variables. This is due to the re-normalization of density resulting from the truncation. The margins are small given the large number of samples in the dataset.[9]

The first column of Table 1.7 reports OLS post lasso coefficients on information variables. 'Superhost' badge has 5.25% price impacts and host verification accounts have 1.73%, both of which are statistically significant at a 1% level. Among seven rating categories, 'Cleanliness', 'Location', and 'Value' were selected. 'Location' score has the greatest price impact of 13.68%. The negative coefficients for 'Value' scores come natural since they represent per dollar satisfaction. The following estimation (Subsection 1.3.2) controlling for unobservables proceeds with these three review scores.

Textual variables indeed turn out to show expected impacts on prices but the magnitudes are much smaller than those of quality certifications and review scores. One thing to note is that the magnitudes and statistical significance of coefficients are greater for review text variables than seller text variables; 'Positive Adjectives' from seller texts are insignificant, and 'Location Phrases' show

[9] See Appendix A.3 and B.1 for a detailed explanation for the exact post selection inference and the resulting confidence intervals and tail areas for p-values.

significant but much smaller implicit price (0.12%) estimate compared to 'Negative Reviews' and 'Positive Phrases' from reviews (-0.73% and 0.70%). This differential impacts are maintained in price elasticity measures ($\beta_k x_k$) on average values of attributes x_k's as expected.

The first column of Table 1.8 reports OLS post lasso coefficients for the chosen 11 amenity and service features out of the total 150. For robustness to the selection procedures, the successive four columns of Table 1.7 and Table 1.8 report OLS coefficients with model selection using other ML methods: data-driven lasso, ridge regression, elastic net, and gradient boosting with n-folds cross-validation and RMSE criterion for price approximation or prediction performances. The difference between the first column (OLS post lasso) and the second (data-driven lasso) is about the purpose of penalized regression problem. The former is for achieving 'oracle' rate of post OLS regression, and the latter is for minimizing the RMSE.

Data-driven lasso is as given in equation (1.5). Ridge regression (Hoerl and Kennard (1970)) uses the penalty with L2 norm, $||\beta||_2$. Elastic net (Zou and Hastie (2005)) defines the penalty with a linear combination of L1 and L2 norms. Gradient boosting (Friedman (2001)) is a variation of regression tree methods which means recursive partitioning of data space for classification or prediction purposes.[10]

Key information variables and accommodation capacities were selected by all five ML methods. Eight additional binary amenity and service features were cross selected by the four data-driven ML methods. This paper proceeds to estimation methods for controlling unobservables with variables in Table 1.7, and 11 unanimously selected amenity and service features in Table 1.8: 'Air Conditioning', 'Buzzer Wireless Intercom', 'Cable TV', 'Free Parking on Street', 'Indoor Fire Place', 'Lock on Bedroom Door', 'Cats Allowed', 'Internet', 'Shampoo', and room types of 'Entire Home/Apt' and 'Shared Room'. One reassuring fact is that whether the model includes binary features selected by ML methods does not systematically alter the main hypothesis of this paper on the superiority of enforced and ex-post verified information contents over non-verified seller side disclosures.

[10] See Appendix A.1 and A.2 for detailed explanation on the methodologies.

Table 1.7: OLS post Lasso on the Pooled Sample (1)

obj : $log(p_{it})$	OLS post LASSO	Data Driven ML			
obs: 75,236		LASSO	RIDGE	GBM	ENET
Quality Certification					
Superhost Indicator	0.0525^a	0.0532^a	0.0509^a	0.0514^a	0.0528^a
	(0.0041)	(0.0041)	(0.0041)	(0.0041)	(0.0041)
Verification Accounts	0.0173^a	0.0176^a	0.0171^a	0.0171^a	0.0176^a
	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)
Review Scores					
Cleanliness	0.0490^a	0.0492^a	0.0487^a	0.0489^a	0.0491^a
	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)
Location	0.1368^a	0.1372^a	0.1372^a	0.1370^a	0.1373^a
	(0.0018)	(0.0018)	(0.0018)	(0.0018)	(0.0018)
Value	-0.0919^a	-0.0919^a	-0.0919^a	-0.0921^a	-0.0920^a
	(0.0022)	(0.0022)	(0.0022)	(0.0022)	(0.0022)
Review Text					
Negative Reviews	-0.0073^a	-0.0072^a	-0.0074^a	-0.0073^a	-0.0072^a
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Positive Phrases	0.0070^a	0.0071^a	0.0070^a	0.0070^a	0.0071^a
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Seller Text					
Positive Adjectives	0.0002		0.0002	0.0002	
	(0.0003)		(0.0003)	(0.0003)	
Location Phrases	0.0012^a		0.0012^a	0.0012^a	
	(0.0002)		(0.0002)	(0.0002)	
Accommomdation Capacities					
Default Guests	0.0563^a	0.0565^a	0.0555^a	0.0555^a	0.0565^a
	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)
Bedrooms	0.1037^a	0.1035^a	0.1035^a	0.1037^a	0.1031^a
	(0.0042)	(0.0042)	(0.0042)	(0.0042)	(0.0042)
Bathrooms	0.1105^a	0.1106^a	0.1104^a	0.1104^a	0.1104^a
	(0.0028)	(0.0028)	(0.0028)	(0.0028)	(0.0028)
Beds	-0.0192^a	-0.0196^a	-0.0199^a	-0.0198^a	-0.0197^a
	(0.0024)	(0.0024)	(0.0024)	(0.0024)	(0.0024)
Guests Included	0.0242^a	0.0244^a	0.0238^a	0.0239^a	0.0243^a
	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)
Constant	3.0792^a	3.0824^a	3.0874^a	3.0878^a	3.0836^a
	(0.0206)	(0.0206)	(0.0207)	(0.0207)	(0.0206)

^{a} : 1% significant,^{b}: 5%, ^{c} : 10%, standard errors in parentheses

Table 1.8: OLS post Lasso on the Pooled Sample (2): Amenity Feature Selection

$obj : log(p_{it})$	OLS post LASSO	Data Driven ML			
obs: 75,236		LASSO	RIDGE	GBM	ENET
UNANIMOUS CHOICE					
Air Conditioner	0.1263^a	0.1271^a	0.1262^a	0.1263^a	0.1270^a
	(0.0042)	(0.0042)	(0.0042)	(0.0042)	(0.0042)
Buzzer Wireless Intercom	0.0878^a	0.0888^a	0.0880^a	0.0879^a	0.0890^a
	(0.0027)	(0.0027)	(0.0027)	(0.0027)	(0.0027)
Cable TV	0.1000^a	0.0995^a	0.0988^a	0.0992^a	0.0992^a
	(0.0028)	(0.0028)	(0.0028)	(0.0028)	(0.0028)
Free Parking on Street	-0.1209^a	-0.1206^a	-0.1227^a	-0.1224^a	-0.1210^a
	(0.0043)	(0.0043)	(0.0043)	(0.0043)	(0.0043)
Indoor Fire Place	0.1223^a	0.1219^a	0.1204^a	0.1208^a	0.1212^a
	(0.0068)	(0.0068)	(0.0068)	(0.0068)	(0.0068)
Lock on Bedroom Door	-0.0677^a	-0.0677^a	-0.0704^a	-0.0704^a	-0.0684^a
	(0.0046)	(0.0046)	(0.0047)	(0.0047)	(0.0046)
Cats Allowed	-0.0846^a	-0.0844^a	-0.0838^a	-0.0841^a	-0.0839^a
	(0.0053)	(0.0053)	(0.0053)	(0.0053)	(0.0053)
Internet	0.0259^a	0.0267^a	0.0248^a	0.0251^a	0.0266^a
	(0.0035)	(0.0035)	(0.0035)	(0.0035)	(0.0035)
Shampoo	0.0392^a	0.0400^a	0.0366^a	0.0371^a	0.0395^a
	(0.0028)	(0.0028)	(0.0029)	(0.0029)	(0.0028)
Room Type					
Entire Home/Apt	0.5901^a	0.5901^a	0.5881^a	0.5880^a	0.5901^a
	(0.0034)	(0.0034)	(0.0034)	(0.0034)	(0.0034)
Shared Room	-0.1640^a	-0.1655^a	-0.1637^a	-0.1638^a	-0.1653^a
	(0.0112)	(0.0112)	(0.0112)	(0.0112)	(0.0112)
CROSS SELECTED					
Fire Extinguisher			0.0025	0.0035	0.0052
			(0.0031)	(0.0031)	(0.0029)
Other Pets Allowed			-0.0500^b	-0.0498^b	-0.0479^b
			(0.0213)	(0.0213)	(0.0213)
Family/Kid Friendly			0.0107^a	0.0110^a	
			(0.0030)	(0.0030)	
Laptop Friendly Workspace			0.0056^b	0.0062^b	
			(0.0029)	(0.0029)	
Safety Card			0.0119^a	0.0133^a	
			(0.0042)	(0.0042)	
Smoke Detector			-0.0178^a	-0.0099^a	
			(0.0042)	(0.0036)	
Carbon Monoxide Detector			0.0123^a		
			(0.0035)		
Hot Tub			0.0044		
			(0.0060)		

1.3.1.3 Cautions with Endogeneity

Hedonic price regressions often suffer from endogeneity due to omitted or unobserved variables. Even though OLS post lasso produced seemingly appropriately signed parameter estimates, it is susceptible to endogeneity. It is because both the lasso selector and post OLS use Gaussian or at best, heteroskedastic errors implicitly assuming that there is no endogeneity due to unobservables. Indeed, Ramsey test F-values for omitted variables in post selection OLS estimates are extremely high for all specifications in Table 1.7 and 1.8.

High-dimensional econometricians in fact, have provided post selection IV regressions with variable selection on both many controls and instruments with a small number of key endogenous variables such as treatment/policy indicators or prices. (Chernozhukov, Hansen, and Spindler (2015), Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018)) But they need a distributional separability assumption between the 'key' variables and lasso selection on controls and instruments, which still uses Gaussian errors. It is understandable in that unobservables are not in the dataset, and ML methods resort to the magnitudes of in-sample prediction errors such as RMSE to choose the 'right' subset of all covariates.

This paper hence proposes to use OLS post lasso only as a guide for efficient dimension reduction in attribute space and employ econometric methodologies to control for endogeneity due to unobservables: fixed effects and a GMM approach with a panel data. They do not pre-specify a small set of endogenous variables but accepts the possibility that unobservables could be correlated with any observed characteristics included in the model.

In fact, the same concerns on endogeneity arise with conventional alternatives including principal component analysis (PCA), Akaike information criterion (AIC), or Bayesian information criterion (BIC), plus being practically infeasible with hundreds of variables to consider. PCA coefficients are linear combinations of covariates which makes it impossible to isolate and identify coefficients for individual variables, and the stepwise nature of AIC and BIC dictates too high calculation costs for estimation and comparison to incur given that there are at most 150 binary indicators.

1.3.2 Time-Varying Correlated Unobservables

1.3.2.1 Possible Sources

Unobservables correlated with observable attributes and prices could be classified into two categories; one is time-invariant and another is time-varying. Time-invariant unobservables include seasonality and neighborhood specific attributes such as education levels, crime rates, and proximity to famous tourist attractions. If there is no time-varying unobservables, biases in parameter estimates could be controlled for with a fixed effects estimation.

$$log(p_{it}) - log(p_{it-1}) = \beta(X_{it} - X_{it-1}) + \epsilon_{it} \tag{1.6}$$

However, there are strong candidates for time-varying unobservables such as curb appeal, direct/indirect advertising, and geographical dynamics. For example, if a consumer finds the curb appeal of a rental unit perceived via web images quite attractive, then he/she could ignore some shortcomings in certain observed attributes. Rental unit images containing curb appeal change overtime in quantity, quality, and contents.

Also, if a potential guest were looking for information about travel accommodation choices using social network services (SNS), chances are the search log and cookies would pop up Airbnb advertisement on the screen. The schedule and coverage of such advertising campaigns vary over time. Concerts, plays, and other events being held in specific areas of NYC could also affect consumers' rental unit choices. Anticipating a traffic jam, a tourist can sacrifice some personal standards on other attributes for the sake of locational merits.

1.3.2.2 Markov Process and Consumer Rationality

Following the ideas of Bajari, Fruewirth, Kim and Timmins (2012), this paper imposes a Markov (AR(1)) process on the errors to describe time-varying unobservables.

$$log(p_{it}) = \alpha + \beta X_{it} + \tau_{it} \tag{1.7}$$

$$\tau_{it} = \rho\tau_{it-1} + \eta_{it}$$

With a few algebraic manipulations, equation (1.7) implies

$$\begin{aligned} log(p_{it}) &= \alpha + \beta X_{it} + (\rho\tau_{it-1} + \eta_{it}) \\ &= (1-\rho)\alpha + \beta(X_{it} - \rho X_{it-1}) + \rho log(p_{it-1}) + \eta_{it} \end{aligned} \tag{1.8}$$

The rationale behind such error structure is twofold; time-invariant and a certain portion of time-varying unobservables could be controlled for with a relatively high value of the persistency parameter ρ. Given that Airbnb rentals are originally individual real estate properties, a rather stable time-series modelling is proposed to describe consumers' expectation of implicit prices for unobservables. Also, accommodation capacities like the number of bedrooms show enough variations to identify our dynamic models, but are stable over time. Idiosyncratic shock η_{it} captures unexpected changes in time-varying unobservables.

The third and final identifying assumption is consumer rationality (equation (1.9)) that time-varying unobservables η_{it} do not affect Airbnb guests' price expectations based on observable characteristics. This orthogonality moment condition implies in other words, that consumers do not make systematic errors in implicit pricing of attributes with the current information set I_t due to unexpected changes in η_{it}.

$$E[log(p_{it}) - \rho log(p_{it-1}) - (1-\rho)\alpha - \beta(X_{it} - \rho X_{it-1})|I_t] = 0 \tag{1.9}$$

Since time-varying unobservables are correlated with observable attributes X_{it}, consumers' current information set I_t includes a set of instruments Z_{it-1} along with ($log(p_{it-1})$, X_{it}, X_{it-1}). This paper uses further lags of observables, X_{it-2} as instruments. Appendix B.4 provides evidence for strong relevance of each x_{it-2} to x_{it} after controlling for $log(p_{it-1})$ and x_{it-1}. The estimation is easily implemented with a standard GMM command in STATA.

Together with fixed-effects estimation, the GMM (Generalized Methods of Moments) approach based on consumer rationality assumption regresses dynamic price adjustments on changes in observed attributes. Given small annual variations in observed attributes and prices (Appendix B.3), there could be a risk of relatively less precise estimates for the implicit prices from the dynamic models. This paper presents estimation results for both fixed-effects and GMM to investigate the research question under various error structures.

1.3.2.3 Previous Empirical Research on Airbnb

Existing hedonic studies on price determinants of Airbnb rental units in tourism and hospitality research employed simple OLS. (Wang and Nicolau (2017), Chen and Xie (2017), Teubner, Hawlitschek, and Dann (2017), Gibbs, Guttentag, Gretzel, Morton, and Goodwill (2018)) Controlling for an obvious existence of unobservables seems necessary to identify implicit prices of attributes more accurately. Also, they do not handle the issue of model selection in the presence of many binary indicators for amenity and service features.

Others employed quasi-experimental identification strategies; Ert, Fleischer, and Magen (2016) hired 900 Amazon Mechanical Turks to study the impacts of the 'Beauty Scores' from host photographs. Edelman, Luca, and Svirsky (2017) created Airbnb guest accounts with names strongly suggestive of African-American ethnicity, and found strong evidence for racial discriminations in booking processes. However, quasi-experimental methods cannot identify multiple attributes like regression models because of the narrow windows focused only on one target variable. Also, artificial environments to generate data are susceptible to biases, not reflecting natural choices of real customers. Finally, it is hard to find a common geographical or regulational break that would affect entire NYC Airbnb market.

Table 1.9 and Table 1.10 report estimation results using fixed effects and GMM for time-varying unobservables on the panel dataset. Columns (1) and (3) report estimation results with amenity and service features selected by ML methods and (2) and (4) without. It is to check if this paper's hypothesis is robust to the variable selection procedures. Platform enforced quality certifications including 'Superhost' badge and 'Verification Accounts' show strongly positive signs for all methods and specifications. The coefficients and significance are higher for GMM than fixed effects, comparing (1) and (3). 'Superhost' units have 0.93 to 1.01% price premiums over normal host units.

1.4 Results

1.4.1 Fixed Effects vs. GMM Based on Consumer Rationality

Table 1.9: Fixed Effects vs. GMM (1)

$obj : log(p_{it})$	Fixed Effects		GMM	
obs: 37,618	(1)	(2)	(3)	(4)
Quality Certification				
Superhost	0.0099***	0.0094***	0.0101***	0.0093***
	(0.0020)	(0.0020)	(0.0020)	(0.0021)
Verification Accounts	0.0022***	0.0022***	0.0033***	0.0019**
	(0.0007)	(0.0007)	(0.0008)	(0.0008)
Review Scores				
Cleanliness	-0.0007	-0.0012	0.0007	0.0014
	(0.0018)	(0.0018)	(0.0022)	(0.0022)
Location	0.0050**	0.0052**	0.0087***	0.0064**
	(0.0021)	(0.0021)	(0.0026)	(0.0026)
Value	-0.0038*	-0.0035*	-0.0070***	-0.0046**
	(0.0020)	(0.0020)	(0.0023)	(0.0023)
Review Texts				
Negative Reviews	-0.0001	0	-0.0023***	-0.0027***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Positive Phrases	0.0036***	0.0035***	0.0034***	0.0036***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Seller Texts				
Positive Adjectives	-0.0006	-0.0006	-0.0004	-0.0007
	(0.0005)	(0.0005)	(0.0007)	(0.0007)
Location Phrases	0.0021***	0.0021***	0.0013***	0.0012***
	(0.0003)	(0.0003)	(0.0004)	(0.0004)
Constant			0.1720***	0.1793***
			(0.0079)	(0.0074)
ρ			0.9654***	0.9643***
			(0.0017)	(0.0015)

***: 1% significant, **: 5%, *: 10%, standard errors in parentheses

Table 1.10: Fixed Effects vs. GMM (2)

$obj : log(p_{it})$	Fixed Effects		GMM	
obs: 37,618	(1)	(2)	(3)	(4)
ACCOMMODATION CAPACITIES†				
Default Guests	0.0269***	0.0365***	0.0304***	0.0409***
	(0.0018)	(0.0018)	(0.0034)	(0.0038)
Bathrooms	0.0151**	0.0139*	0.0260**	0.0236*
	(0.0076)	(0.0077)	(0.0122)	(0.0130)
Bedrooms	0.0420***	0.0583***	0.0356***	0.0607***
	(0.0042)	(0.0041)	(0.0079)	(0.0088)
Beds	0.0125***	0.0133***	0.0136***	0.0188***
	(0.0026)	(0.0026)	(0.0040)	(0.0045)
Included Guests	0.0167***	0.0173***	0.0153***	0.0172***
	(0.0018)	(0.0018)	(0.0029)	(0.0030)
AMENITY AND SERVICE				
Air Conditioner	0.0005		-0.0054	
	(0.0048)		(0.0071)	
Buzzer Wireless Intercomm	0.0120***		0.0111*	
	(0.0046)		(0.0062)	
Cable TV	-0.0005		0.0064	
	(0.0042)		(0.0051)	
Free Parking on Street	-0.0059		-0.0033	
	(0.0057)		(0.0077)	
Indoor Fire Place	0.0141		0.0295*	
	(0.0128)		(0.0165)	
Lock on Bedroom Door	-0.0033		-0.0170***	
	(0.0047)		(0.0060)	
Cats Allowed	-0.0330***		-0.0174	
	(0.0076)		(0.0127)	
Internet	-0.0024		0.0007	
	(0.0043)		(0.0051)	
Shampoo	-0.0029		0.0004	
	(0.0035)		(0.0044)	
ROOM TYPE				
Entire Home/Apt	0.1499***		0.1700***	
	(0.0060)		(0.0114)	
Shared Room	-0.0478***		0.0079	
	(0.0157)		(0.0389)	

†: One concern with fixed effects and GMM could be that within a year (from 2016 to 2017), there may not be enough variations in accommodation capacities for identification. It turns out that unlike hotels, Airbnb hosts make non-negligible changes to accommodation capacities. (see Appendix B.3)

For review scores, the implicit price estimates and their statistical significance for 'Location' and 'Value' improve in GMM compared to fixed effects specifications: from 0.5% and -0.38% impacts with fixed effects to 0.87% and -0.7% with GMM, respectively. 'Negative Reviews' have expected negative sign and are highly significant with GMM, but not with fixed effects estimation. 'Positive Phrases' extracted from consumer review texts have positive impact on rental prices for both methods. Leaving phrases like 'would definitely come back' implies a strong satisfaction of customers from the past, and hence it is expected that consumers would find them trustworthy. However, the coefficients and statistical significance of sellers' advertisement texts both are less than consumer review text variables; non-verified 'Positive Adjectives' are insignificant whereas 'Location Phrases' are significant, which is likely because it is hard to lie about locational merits given that potential guests can check the location with Google Maps API on web listings instantly.

The coefficients on variables for accommodation capacities show expected signs, with 'Default Guests', 'Bathrooms', and 'Beds' showing noticeable increments in coefficients with GMM method. Also, the price premiums for room type 'Entire Home/Apt' become greater and more significant with GMM. Among amenity and service features, 'Indoor Fire Place' and 'Lock on Bedroom Door' become significant in GMM estimation, though 'Cats Allowed' and the room type 'Shared Room' become insignificant. 'Buzzer Wireless Intercomm', 'Indoor Fire Place', and 'Entire Home/Apt' usually come with private house or apartment which make them as indicators of price premiums; 'Lock on Bedroom Door' is often associated with rental units of shared spaces such as youth hostels, thus a good indicator of cheap prices.

Appendix B.2 reports fixed effects and GMM estimation results over Manhattan and other neighborhoods. More than 47% of NYC Airbnb rental units are concentrated in Manhattan area with a higher average price by more than $50. In a less than 3 mile distance circle, central Manhattan area contains most of the tourist attractions and famous places; if a rental unit belongs to Manhattan could be an indicator for many unobservables that could be correlated with prices. The results confirm the dominance of enforced and ex-post verified reputation/feedback contents over non-verified seller side disclosures in both regions.

1.4.2 Over Price Levels

Table 1.11: GMM: Results for Reputation/Feedback Contents over Price Levels

$obj : log(p_{it})$	Fixed Effects			GMM		
obs: 37,618	1Q	2/3Q	4Q	1Q	2/3Q	4Q
Quality Certification						
Superhost	0.0187***	0.0061**	0.0071**	0.0190***	0.0080***	0.0081**
	(0.0047)	(0.0028)	(0.0036)	(0.0045)	(0.0030)	(0.0033)
Host Verification	0.0061***	0.0047***	-0.0061***	0.0030**	0.0043***	-0.0004
	(0.0016)	(0.0010)	(0.0013)	(0.0014)	(0.0011)	(0.0015)
Review Scores						
Cleanliness	-0.0089**	0.0067**	-0.0080**	-0.0010	0.0089**	-0.0044
	(0.0039)	(0.0025)	(0.0034)	(0.0047)	(0.0031)	(0.0034)
Location	0.0127***	0.0038	-0.0029	0.0105**	0.0083**	0.0075*
	(0.0042)	(0.0029)	(0.0044)	(0.0050)	(0.0034)	(0.0045)
Value	0.0029	-0.0089***	-0.0030	0.0011	-0.0107***	-0.0031
	(0.0045)	(0.0028)	(0.0036)	(0.0050)	(0.0033)	(0.0037)
Review Texts						
Negative Reviews	0.0007	-0.0006	-0.0030***	-0.0022***	-0.0030***	-0.0014*
	(0.0009)	(0.0006)	(0.0009)	(0.0008)	(0.0006)	(0.0008)
Positive Phrases	0.0073***	0.0038***	0.0032***	0.0051***	0.0028***	0.0025***
	(0.0014)	(0.0008)	(0.0010)	(0.0012)	(0.0008)	(0.0009)
Seller Texts						
Positive Adjectives	0.0002	-0.0013*	-0.0023**	0.0009	-0.0014	-0.0011
	(0.0010)	(0.0007)	(0.0010)	(0.0011)	(0.0010)	(0.0014)
Location Phrases	0.0045***	0.0021***	-0.0002	0.0025***	0.0014***	-0.0001
	(0.0007)	(0.0004)	(0.0006)	(0.0008)	(0.0005)	(0.0006)
Constant				0.3237***	0.3844***	0.3472***
				(0.0280)	(0.0229)	(0.0356)
$\rho : rho$				0.9234***	0.9193***	0.9338***
				(0.0070)	(0.0049)	(0.0066)

1Q: rental units of lower 25% price range, 2/3Q: middle 50%, and 4Q: upper 25%

***: 1% significant, **: 5% ,*: 10%, standard errors in parentheses

Table 1.12: GMM: Results for Amenity and Service Features

$obj\ :\ log(p_{it})$	Fixed Effects			GMM		
obs: 37,618	1Q	2/3Q	4Q	1Q	2/3Q	4Q
Accommodation Capacities						
Default Guests	0.0236***	0.0271***	0.0243***	0.0319***	0.0360***	0.0236***
	(0.0052)	(0.0024)	(0.0029)	(0.0088)	(0.0046)	(0.0048)
Bathrooms	-0.0020	0.0337***	0.0188	-0.0070	0.0130	0.1240***
	(0.0144)	(0.0106)	(0.0181)	(0.0182)	(0.0162)	(0.0279)
Bedrooms	0.0239**	0.0516***	0.0364***	0.0136	0.0345***	0.0466***
	(0.0119)	(0.0053)	(0.0076)	(0.0193)	(0.0091)	(0.0123)
Beds	0.0019	0.0186***	0.0115***	-0.0006	0.0184***	0.0029
	(0.0079)	(0.0035)	(0.0039)	(0.0113)	(0.0062)	(0.0050)
Included Guests	0.0299***	0.0281***	0.0062***	0.0105	0.0314***	0.0055*
	(0.0059)	(0.0029)	(0.0024)	(0.0103)	(0.0047)	(0.0032)
Amenity and Service						
Air Conditioner	-0.0074	0.0033	-0.0188	0.0136	-0.0009	-0.0183
	(0.0089)	(0.0064)	(0.0134)	(0.0124)	(0.0093)	(0.0156)
Buzzer Wireless Intercomm	0.0262***	0.0212***	-0.0299***	0.0248**	0.0169**	-0.0126
	(0.0098)	(0.0063)	(0.0086)	(0.0127)	(0.0079)	(0.0096)
Cable TV	-0.0152	-0.0020	0.0146**	-0.0111	0.0062	0.0209***
	(0.0097)	(0.0058)	(0.0074)	(0.0113)	(0.0073)	(0.0075)
Free Parking on Street	0.0288**	-0.0202***	-0.0159	0.0269*	-0.0184*	-0.0135
	(0.0115)	(0.0078)	(0.0114)	(0.0153)	(0.0105)	(0.0147)
Indoor Fire Place	0.1049***	-0.0217	0.0197	0.1123***	0.0296	0.0196
	(0.0324)	(0.0201)	(0.0181)	(0.0434)	(0.0289)	(0.0178)
Lock on Bedroom Door	0.0139*	-0.0201***	-0.0852***	0.0018	-0.0301***	-0.0710***
	(0.0077)	(0.0067)	(0.0168)	(0.0090)	(0.0078)	(0.0260)
Cats Allowed	-0.0273**	-0.0410***	-0.0267	-0.0306**	-0.0581***	0.0097
	(0.0129)	(0.0112)	(0.0186)	(0.0149)	(0.0191)	(0.0298)
Internet	0.0134	-0.0078	-0.0116	0.0045	-0.0029	-0.0141
	(0.0093)	(0.0059)	(0.0084)	(0.0103)	(0.0070)	(0.0090)
Shampoo	-0.0024	-0.0034	0.0025	0.0047	-0.0001	0.0015
	(0.0081)	(0.0048)	(0.0061)	(0.0100)	(0.0061)	(0.0066)
Room Type						
Entire Home/Apt	0.2514***	0.1225***	0.1300***	0.3284***	0.2228***	0.1182***
	(0.0169)	(0.0077)	(0.0110)	(0.0327)	(0.0133)	(0.0202)
Shared Room	-0.1023***	0.0601**	-0.1076	-0.1137**	0.1239**	0.0367
	(0.0251)	(0.0258)	(0.0349)	(0.0552)	(0.0490)	(0.0596)

Price level itself acts as a quality indicator for consumers and there is a risk of reverse causality in hedonic price regressions with information variables on product quality as covariates. Also it is worth investigating the differential effects of information disclosures over price ranges. Table 1.11 and Table 1.12 report estimation results on the three panel datasets over three price ranges: lower 25%, middle 50%, and upper 25% with average nightly prices of $75, $159, and $337 respectively.

Consumers turn out to value 'Superhost' badge much more (2-3 folds) in rental units of lower prices. It has price effects of 1.9% compared to 0.6-0.8% of rental units with higher price ranges. The coefficients become greater and more significant with GMM estimation, namely controlling for endogeneity due to time-varying correlated unobservables. Host 'Verification Accounts' also show strongly positive and significant coefficients for rental units at both lower 25% and middle 50% price ranges. It might reflect the fact that consumers might demand more rigorous credibility and professionalism standard for relatively cheap rental units. Since cheap rentals could lure low quality sellers in the market segment inducing high risk of information asymmetry, credible measures of product quality would be appreciated by potential guests much more.

For review scores, Airbnb guests for rental units of the middle 50% price range turn out to care about all three categories of review scores: 'Cleanliness', 'Location', and 'Value'. Customers of relatively cheap and expensive rental units only consider 'Location' scores, with implicit price estimates of 1.05% and 0.075%, respectively. It could be stated that locational merits are indeed an important source of price premiums over all price levels. Comparing estimation methods, 'Cleanliness' score had a strongly negative coefficient with fixed effects but become insignificant with GMM for rental units of lower and upper 25% price ranges . Another sign change and gain in significance occur for 'Location' score in the top 25% rentals with GMM. There are overall improvements in statistical significance and increments in coefficients over all review score categories for the units of middle 50% price range.

Coefficients for 'Negative Reviews' become highly significant and show expected negative signs over all price levels with GMM. Coefficients for 'Positive Phrases' from review texts are also highly significant and the magnitude is particularly higher for cheap rentals, confirming the finding that consumers require ex-post verified trustworthiness more for cheap Airbnb units, similar to the differential impacts of 'Superhost'. It also conforms to the finding in online commerce research that prices of cheaper products are more sensitive to positive e-WOM. (electronic word of mouth, Shin, Hanssesns, and Kim (2016))

'Positive Adjectives' extracted from sellers' advertisement texts are insignificant with GMM. 'Location Phrases' are highly significant for rental units of lower 25% and middle 50% price ranges, but the magnitudes are much smaller than those of quality certifications, review scores, and review text variables.

Regarding accommodation capacities, rental guests who book low priced rentals seem to only care about the number of 'Default Guests'; higher priced rental customers seem to consider the number of bathrooms, bedrooms, beds, and included guests altogether. A likely explanation for this could be that a tourist or tourists who visit a rental unit of an average price of $75 in NYC mostly are finding a place to spend the night.

A few interesting sign changes occur for amenity and service features. 'Free Parking on Street' adds a price premium for lower priced rentals, but is a minus factor for rentals of higher prices; customers who are visiting rentals with prices coming close to three star hotels (more than $158) usually anticipate a designated parking space or a private garage. 'Lock on Bedroom Door' could be welcomed by tourists who visit a cheap rental unit expecting the space being shared with other guests but it is definitely a minus factor indicating low level of privacy for higher priced rentals.

Also, the room type 'Entire Home/Apt' show differential impacts over price levels. The level of privacy it presents is much more appreciated in cheaper rentals. The fact that 'Shared Room' has negative price impacts for cheaper rentals and positive for middle 50% price range could reflect the difference between the dominant property types each price level implies; rentals of $75 nightly price in NYC usually mean a youth hostel or a smoking guy's couch. However, rentals with prices of more than $158 typically means private housing units where a customer shares the house with the host or family. The so advertised 'home out of home' experiences and interactions with locals seem to come only above a certain price level.

1.5 Conclusion

Sharing economy such as Airbnb could suffer more severe information asymmetry in that products and services offered have never been tested as marketable, and involve further risks than monetary losses. This paper shows that quality certifications and consumer reviews resolve adverse selection as verified in other online marketplaces over various error structures and specifications. Non-verified voluntary disclosures from sellers turn out to be less influential than enforced and ex-post verifiable information on product quality. Machine learning methods, text processing techniques, and flexible identifying assumptions were proposed to deal with identification challenges modern that online platform data present.

CHAPTER 2

DEMAND ESTIMATION FOR NYC AIRBNB MARKET: VALUE OF REPUTATION/FEEDBACK CONTENTS AND VOLUNTARY DISCLOSURES

2.1 Introduction

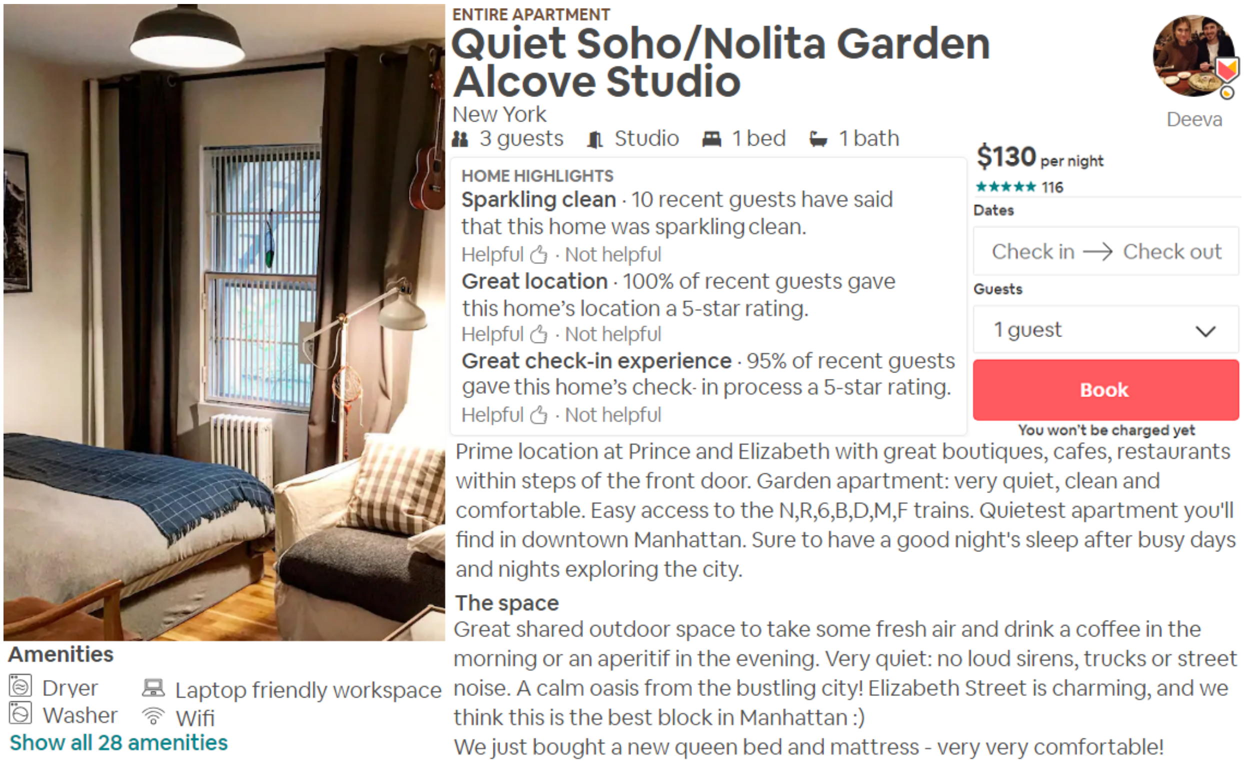
2.1.1 P2P Online Marketplaces and Asymmetric Information

P2P (Peer-to-Peer) online marketplaces can be thought of as a matching platform between sellers of underutilized idle assets and buyers who are willing to pay for temporary occupation of the assets. For example, Uber or Lyft makes an ordinary car owner a taxi driver and Turo lets people lend and borrow cars from each other. The same business model applies to leftover parking spaces (Parking Panda), bikes, surfboards, ski equipments (Spinlister), and many others. Airbnb also falls into this category, where home owners can be travel accommodation hosts and travelers can enjoy 'home away from home' experiences. Airbnb's 'sharing economy' platform recorded a market value of $31 billion in 2017, and is being considered as a serious threat to the existing accommodation businesses.

This paper empirically tests if the insight of classical information economics (i.e., disclosure models) contributes to the remarkable success of Airbnb in NYC vacation rental market (Akerlof (1970), Grossman and Hart (1980), Grossman (1981), and Milgrom (1981)). The insight from disclosure models, put simply, means that if the platform provides trustworthy (verifiable) information on product quality, it can prevent market failure caused by adverse selection due to information asymmetry. Conversely, providing non-trustworthy (non-verifiable) information on product quality would not prevent the market failure due to the asymmetric information.

This paper tests how various information contents on Airbnb websites affect consumer choices and quantifies how much value (in $) each content created for consumers. Specifically, it quantifies the value of non-verified information provided by the sellers and verifiable information provided by prior consumers. Figure 2.1 shows an example Airbnb listing, with product information stored

Figure 2.1: Example Airbnb Listing, a 'Superhost' Rental



ENTIRE APARTMENT

Quiet Soho/Nolita Garden Alcove Studio

New York

3 guests Studio 1 bed 1 bath

Deeva

HOME HIGHLIGHTS

Sparkling clean · 10 recent guests have said that this home was sparkling clean.

Helpful · Not helpful

Great location · 100% of recent guests gave this home's location a 5-star rating.

Helpful · Not helpful

Great check-in experience · 95% of recent guests gave this home's check- in process a 5-star rating.

Helpful · Not helpful

$130 per night

★★★★★ 116

Dates

Check in → Check out

Guests

1 guest

Book

You won't be charged yet

Prime location at Prince and Elizabeth with great boutiques, cafes, restaurants within steps of the front door. Garden apartment: very quiet, clean and comfortable. Easy access to the N,R,6,B,D,M,F trains. Quietest apartment you'll find in downtown Manhattan. Sure to have a good night's sleep after busy days and nights exploring the city.

The space

Great shared outdoor space to take some fresh air and drink a coffee in the morning or an aperitif in the evening. Very quiet: no loud sirens, trucks or street noise. A calm oasis from the bustling city! Elizabeth Street is charming, and we think this is the best block in Manhattan :)

We just bought a new queen bed and mattress - very very comfortable!

Amenities

Dryer Laptop friendly workspace

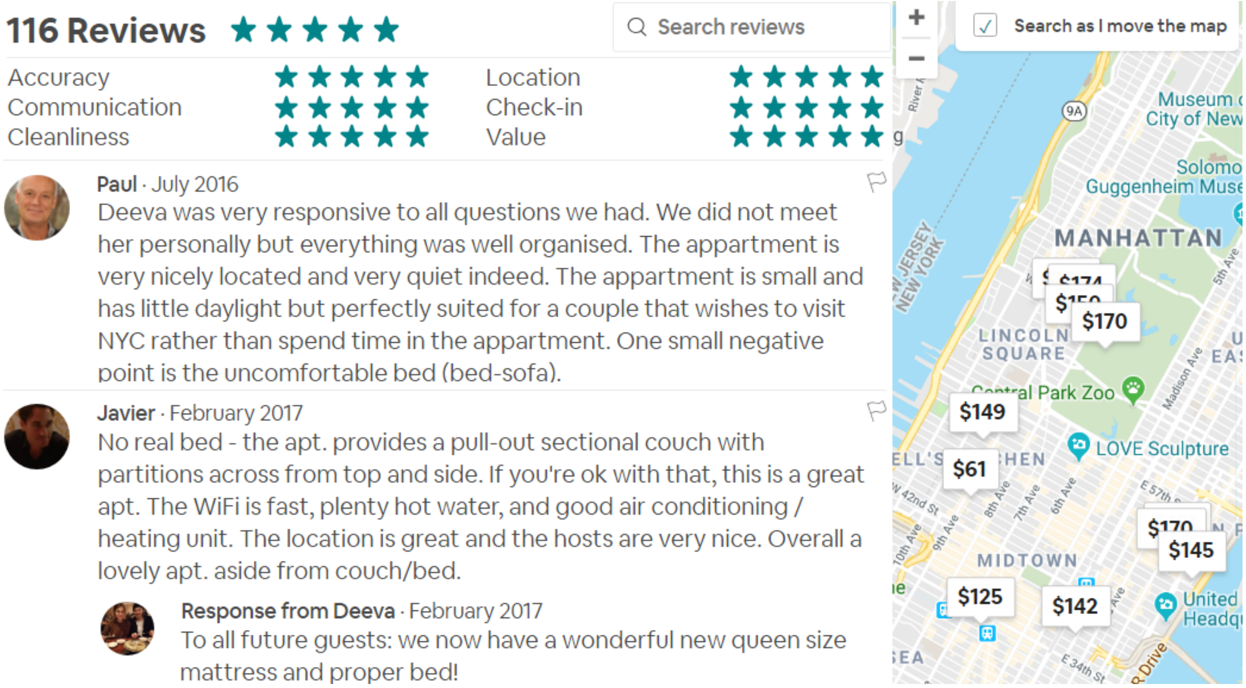
Washer Wifi

Show all 28 amenities

in photographs and texts. Some of them are standard and formatting is provided by the platform, such as the room type ('Entire Apt'), accommodation capacities (the number of guests, beds, and bathrooms), 'Home Highlights' and 'Superhost' badge right next to the host's photo. While Airbnb provides standardized formatting, sellers have discretion in terms of the pictures, texts, and specific information to be provided.

Figure 2.2 shows the review ratings and texts from the past travelers. Airbnb lets only the actual guests who visited the rental unit leave reviews, and hence they are more trustworthy or 'ex-post verified'. Potential future guests can check the validity of sellers' contents indirectly by the reputation they accumulated over time, and it functions as a feedback mechanism for sellers to enhance their product quality. In addition, the platform provides quality certifications ('Superhost' badge and 'Home Highlights') based on past buyer review ratings. Textual descriptions such as 'Prime location' or 'Quietest apartment' provided by sellers can be deemed as non-verified information on product quality or, 'Cheap Talk'.

Figure 2.2: Example Review Ratings, Texts, and Google Maps API



116 Reviews ★★★★★

Search reviews

Accuracy	★★★★★	Location	★★★★★
Communication	★★★★★	Check-in	★★★★★
Cleanliness	★★★★★	Value	★★★★★

Paul · July 2016
Deeva was very responsive to all questions we had. We did not meet her personally but everything was well organised. The appartment is very nicely located and very quiet indeed. The appartment is small and has little daylight but perfectly suited for a couple that wishes to visit NYC rather than spend time in the appartment. One small negative point is the uncomfortable bed (bed-sofa).

Javier · February 2017
No real bed - the apt. provides a pull-out sectional couch with partitions across from top and side. If you're ok with that, this is a great apt. The WiFi is fast, plenty hot water, and good air conditioning / heating unit. The location is great and the hosts are very nice. Overall a lovely apt. aside from couch/bed.

Response from Deeva · February 2017
To all future guests: we now have a wonderful new queen size mattress and proper bed!

☑ Search as I move the map

This paper's primary focus is to estimate the values of verifiable user information and the non-verifiable seller information on product quality. It is from the key insight from Kreps (1982, 1990) and Tadelis (2016) that a well functioning public repositories of reputation/feedback mechanisms discipline sellers to act honestly even with a total stranger in every future transaction, which sustains a market of trades among anonymous individuals like Airbnb. Using utility parameter estimates from logit demand models, I find that the compensating variations for review ratings and seller texts are about $38.54 and $3.65 million respectively, with a total of 708,308 reservations during 2016 and 2017 in NYC for a counterfactual scenario of complete absence. It shows that ex-post verified information on product quality affects consumer choices more than non-verified seller side voluntary disclosures.

There are three estimation challenges. They originate from extreme product heterogeneity that is inevitable due to the business model of P2P platforms, gathering as many unique individual assets as possible. First is 'too many products' for consumers to choose from. If an econometrician fails to consider consumers' realistic choice set formation, demand parameters will suffer biases. Second is high dimensional attribute space. A new set of 150 diverse amenity and service features such as ’video game consoles’ and ’EV chargers’ is now on search filters and product descriptions. In the dataset, they are all binary indicators causing the risk of multicollinearity and irrelevant variables. Third is product information stored in unstructured or non-numerical text format i.e., consumer reviews and sellers’ advertisement texts as can be seen in Figure 2.1 and 2.2. Given that they affect consumer choices, appropriate data science techniques to transform such texts into numerical variables are called for.

This paper contributes to the literature by addressing the estimation challenges listed above using a set of tools recently developed in econometrics, i.e., machine learning (high dimensional metrics) and text processing. They can be flexibly adjusted to most of general linear models used in applied microeconomics research and possess a great potential for extracting policy insights or business intelligence from massive size datasets (’Big Data’).

2.1.2 Estimation Challenges with Airbnb Platform Data

The first and primary concern of demand estimation with many products is to control for consumer choice sets. There are on average 40,000 unique individual properties operating as Airbnb rentals in NYC alone. If a researcher falsely assumes and models a consumer to choose a rental unit from all of the tens of thousands products, utility parameter estimates will definitely suffer biases. The dataset used in this paper is aggregate (market) level, which means it includes individual rental units’ market shares but does not contain individual demographics that could directly identify a consumer’s choice set formation processes.

Too many choice alternatives indeed have been one of the most challenging identification problems in empirical industrial organization. (Berry, Linton, and Pakes (2004) and Berry and Pakes (2007)) Marketing researchers also showed that consumers pay attention to only a small subset of all products. (Draganska and Klapper (2011), and Kim, Albuquerque, and Bronnenberg (2010)) If additional information on consumers’ choice set formation is available such as scanner data in retail demand, discrete choice models and GMM estimation can be flexibly adjusted to utilize such information (Kim and Kim (2017)).

To describe a realistic choice set formation with utility models, this paper borrows the findings from careful observations on Airbnb guests' web search behaviors. As Fradkin (2017) notes, Airbnb rental searchers heavily rely on Google Maps API on the search screen. Among searchers who sent a reservation request, more than 64% of them changed the default map locations and more than 50% used the zoom-in function to further reduce choice sets. Another key search filter was 'Room Type': Entire Home/Apt, Private, and Shared Room. Nearly 70% of potential guests applied this filter to find a rental unit.

Hence this paper employs a three level nested multinomial logit (NMNL) model to reflect consumers' preferences for geographic locations in choice set formation during web search process. The nesting structure is based on the mutually exclusive service neighborhood designations by Airbnb. It first divides NYC into five precincts: Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Each precinct contains 32 to 53 neighborhoods, such as 'SoHo' in Manhattan. The number of rentals in a neighborhood varies from 2 to more than 5,500.

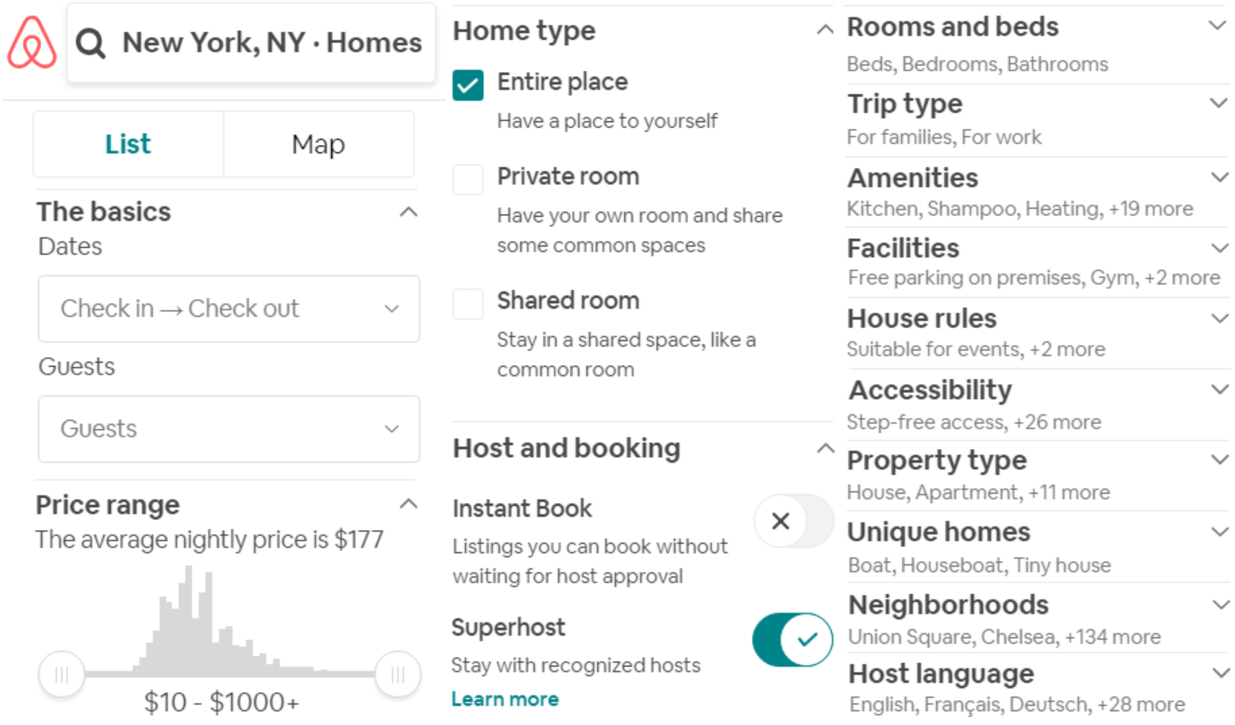
The first level nesting parameter for the big five regions (precincts) is intended to capture the 'changing default map location' behaviors, which occurs at a wider scale on the map. The second level nesting parameter for neighborhoods is to capture the zoom-in/out behaviors; tourists often have intentions to visit or preferences for famous neighborhoods such as 'Hell's Kitchen', 'Midtown', or 'Financial Districts'. I then include the binary indicators for 'Entire Home/Apt' and 'Shared Room'. Each counts for 51% and 2.79% of entire rental units in the sample. The estimation results show highly significant and economically meaningful estimates for the nesting parameters and 'Room Type' filters, suggesting that the modeling choice was able to capture the targeted aspects of consumer preferences.

The second identification challenge due to extreme product heterogeneity is high dimensional attribute space. To differentiate each rental from another and match it to heterogeneous consumer preferences, a new set of amenity and service features that traditional hotel chains cannot provide is added to search filters, e.g., baby beds, children's dinnerware, EV charger, and video game consoles. Property types also show surprising variety, including cabins, castles, farm barns, camping cars, and tree houses. The dataset at our hands as a result attaches 150 binary indicators, giving a high dimensional attribute space (Figure 2.3).

Such high dimensionality with many binary indicators poses two serious threats; some characteristics are common and some are scarce, causing multicollinearity. There are irrelevant variables that do not affect a consumer's purchase decision significantly. The demand parameter estimates could suffer biases or misspecifications.

An efficient way of dimension reduction is called for. This paper hence proposes to adopt sparsity assumption and use variable (model) selection by a lasso variant (Belloni and Chernozhukov (2013)), choosing a subset of attributes that explains variations in sales (market shares) the best. The demand model hence includes only the selected attributes, and exact inferences for the parameter estimates adjusted to reflect additional uncertainty due to the model selection procedures were proposed directly following Lee, L. Sun, Sun, and Taylor (2016).

Figure 2.3: Search Filters for NYC Airbnb Rentals



New York, NY · Homes

List | Map

The basics

Dates

Check in → Check out

Guests

Guests

Price range

The average nightly price is $177

$10 - $1000+

Home type

☑ Entire place
Have a place to yourself

□ Private room
Have your own room and share some common spaces

□ Shared room
Stay in a shared space, like a common room

Host and booking

Instant Book ✕
Listings you can book without waiting for host approval

Superhost ✓
Stay with recognized hosts
Learn more

Rooms and beds
Beds, Bedrooms, Bathrooms

Trip type
For families, For work

Amenities
Kitchen, Shampoo, Heating, +19 more

Facilities
Free parking on premises, Gym, +2 more

House rules
Suitable for events, +2 more

Accessibility
Step-free access, +26 more

Property type
House, Apartment, +11 more

Unique homes
Boat, Houseboat, Tiny house

Neighborhoods
Union Square, Chelsea, +134 more

Host language
English, Français, Deutsch, +28 more

The third identification challenge is to incorporate textual data as shown in Figure 2.1 and 2.2 into the econometric models. The platform allows each Airbnb hosts to post text descriptions and photographs. Through advertisement texts, a host extols the merits of his/her rental unit and neighborhood: nearby tourist attractions, transportation logistics, restaurant/shopping mall recommendations, and house rules guests should abide by. They characterize the identity of each unique product that cannot be transmitted via numerical variables or search filters (binary indicators), letting sellers better differentiate from another.

This paper proposes to generate numerical variables from text data using processing techniques that are widely accepted in online tourism/hospitality research: extracting keywords/phrases and sentiment analysis on consumer reviews.[1] In other words, the frequency of appearance of certain words/phrases and the number of reviews that were classified as negative by a supervised machine learning will be used as additional product attributes. Image processing is beyond the scope of this paper, but represents another source of product information for consumers. Economists also used text analysis in studying online marketplaces such as eBay Motors and real estate (Lewis (2011) and Nowak and Smith (2017)).

2.1.3 Literature Review and My Contribution

This paper will be the first to employ standard logit demand models to quantify the value of information contents in NYC Airbnb market, with a newly constructed dataset of the actual NYC Airbnb tourists' rental unit choices between 2016 and 2017. It proposes a new set of empirical toolsets for the pervasive estimation challenges for P2P online platform data: too many choice alternatives, high dimensional attribute space, and unstructured texts. Nested logit models for consumer choice sets, dimension reduction using machine learning, and text processing for 130,000 seller texts and 850,000 review texts differentiate this paper from the previous empirical research for Airbnb market using hedonic price regressions (Wang and Nicolau (2017), Chen and Xie (2017), Teubner, Hawlitscheck, and Dann (2017), and Gibbs, Guttentag, Gretzel, Morton, and Goodwill (2018)) and quasi-experimental approaches. (Ert, Fleischer, and Margen (2016) and Edelman, Luca, and Svirsky (2017))

[1] See Alaei, Becken, and Stantic (2017) for a comprehensive review of methods and literature up to date.

My methodological contributions also extend to previous literature in economics and marketing/management on the role of reputation/feedback systems and voluntary disclosures in reducing information asymmetry. Economists found that eBay's 'Superseller', seller rating scores, and text/photo descriptions have causal relationships with transaction prices (Resnick, Zeckhauser, Swanson, and Lockwood (2006), Houser and Wooders (2006), Jin and Kato (2006), and Lewis (2011)). Marketing/management researchers investigated the role of online reviews in business outcomes in various markets for books, movies, music, retail, and etc (Chavalier and Mayzlin (2006), Liu (2006), Dellarocas, Zhang, and Awad (2007), Duan, Gu, Whinston (2008), Chintagunta, Gopinath, and Venkataraman (2010), Dhar and Chang (2009), Floyd et al. (2014), Cui, Lui, and Guo (2012), Ghose and Ipeirotis (2011)).

This paper follows closely Lewis and Zervas (2016), wherein the authors study the welfare impacts of consumer review ratings for U.S. hotel industry. The authors estimated a series of logit demand models using a proprietary 10 year monthly panel data from Smith Travel Research containing 5,944 hotels in Arizona, California, Nevada, Oregon, and Washington (45% of all hotels). They augmented the dataset with a panel of consumer reviews from three online travel review platforms: TripAdvisor, Expedia, and Hotels.com each of which containing 807,140, 1,410,488, and 1,544,883 review ratings. The welfare implications (compensating variations) of the ratings in a counterfactual scenario of complete absence of review ratings vary over how counterfactual prices were calculated. The aggregate consumer surplus falls about $123 million without price adjustments, $107 million with the conventional nash equilibrium prices, and $546 million in the case of reduced form price changes.

The methodological contributions of this paper compared to Lewis and Zervas (2016) are clear. First, I focus on a regional market (NYC) and reflect the actual consumers' choice set formation principles into the modeling approach. Though they included market-year-monthly fixed effects in the utility specifications, the fact that their dataset covers a wide range of locational segments make the analysis focused more on the hotel market as a whole, and less on correctly describing consumers' decision making processes leaving concerns of bias.

Secondly, I employ machine learning and high dimensional econometrics to handle high dimensional attribute space. Though machine learning could provide an efficient way of dimension reduction, econometric modeling and identification techniques for endogeneity control turn out to be essential for successful identification. Plus, I show that product information stored in text format is worth being incorporated into empirical analyses.

A couple of clear shortcomings of this paper compared to Lewis and Zervas (2016) and key demand literature in industrial organization (BLP (1995), Nevo (2001), and Petrin (2002)) include first, that I do not estimate supply side moments. One excuse for this would be that each of Airbnb vacation rentals is a unique individual housing unit, which makes it hard to justify a simple Nash equilibrium marginal costs modeling. Another shortcoming is not to incorporate random coefficients into the nested logit models, mainly due to the extremely small market shares of a single rental property and the resulting numerical instability sensitive to pre-set initial values for the iterative BLP estimation routine.

Instead, I used hedonic price adjustments in producing counterfactual prices following Hausmann and Leonard (2002). The resulting compensating variations for information contents from the demand estimates support this paper's hypothesis that the platform's quality certifications and consumer reviews (ex-post verified) show greater welfare impacts than non-verified seller side advertisement texts.

The rest of this paper is organized as follows. Section 2.2 introduces NYC accommodation market, the dataset, and processing details. Section 2.3 discusses actual NYC Airbnb guests' web searching behaviors in more detail, presents the demand model, and estimation methods. Section 2.4 reports the demand parameter estimates, price elasticities, and welfare measures for key reputation/feedback and disclosure devices. Section 2.5 concludes.

2.2 Data

2.2.1 NYC Accommodation Market

2.2.1.1 Market Definition and Size

This paper defines the market size of NYC accommodation market as the total potential number of accommodation reservations. Table 2.1 lists the annual number of visitors to NYC and since 2016, more than 60 million tourists came to NYC with about 20% of international arrivals.[2] Assuming each of them reserves an accommodation facility, the total potential annual number of reservations can be obtained by dividing the number of annual NYC visitors by the average length of stays. Based on the actual booking records obtained from both Expedia.com and Airbnb rentals (Table 2.2), I roughly assume that the average length of stays is four days.[3]

The rationale behind such a comprehensive market definition comes from the fact that Airbnb is indeed taking market shares in various accommodation segments, not just competing with hotels. A survey on 4,000 potential Airbnb tourists found that[4]: (1) As of November, 2015 the market share of Airbnb.com is occupying 12% of leisure and business travelers and projected to reach 16-18% in 2016; (2) 42% of Airbnb demand is coming at the expense of hotels. Moreover, it is also replacing other non-traditional accommodations such as bed and breakfast inns, vacation rentals and stays with friends and family. The last segment makes up around 60% of overnight accommodations, and is thus larger than hotels. Another report by an accommodation market research company says that the room nights share of Airbnb amounts to 8% compared only to hotels as of August, 2015.[5]

Defining the market size with the maximum purchasing capacity is a fairly conventional approach taken by key demand literature in empirical industrial organization; Berry, Levinsohn, and

[2] The visitor poll is from NYC & Company.

[3] Expedia.com launched a prediction contest in 2016. The task was to predict top five hotel recommendations based on the distributed data. The locations of hotels were anonymized but a participant decoded the regional codes by the distances between users and destination hotels. The contest operator confirmed the leak. The sales records for NYC Airbnb rentals were purchased from Airdna.

[4] Who Will Airbnb Hurt More - Hotels or OTAs (Online Travel Agency)? - JP Morgan Global Insight

[5] Airbnb and Impacts on the New York City Lodging Market and Economy - Hospitality Valuation Services

Pakes (1995) defined the total annual market size for automobile as the number of households in each year. Nevo (2001) used the total potential number of servings in a city per quarter as the quarterly market size for ready to eat cereals.

Table 2.1: NYC Visitors and Potential Reservations for Travel Accommodations

Year	Total	Domestic	International	Potential Reservations	Record
2015	58,500,000	46,200,000	12,300,000	14,625,000	
(Jan - Jun)	23,400,000	18,480,000	4,920,000	5,850,000	
(Jul - Dec)	35,100,000	27,720,000	7,380,000	8,775,000	Actual
2016	60,300,000	47,600,000	12,650,000	15,075,000	
(Jan - Jun)	24,120,000	19,040,000	5,060,000	6,030,000	
(Jul - Dec)	36,180,000	28,560,000	7,590,000	9,045,000	
2017	61,800,000	48,700,000	13,100,000	15,450,000	Projected

Table 2.2: Actual Booking Data Summary for NYC Market for Hotels and Airbnb

Source	Expedia.com			Airbnb.com		
Sample	Random Sample			All Sales		
Coverage	All Segments			All Segments		
Average Length	3			4.984		
S.D. of Length	2.106			5.017		
Reservations	96,262			1,987,362		
Sale Periods	201301 - 201412			201408 - 201704		
Length of Stay	Frequency	%	Cum. %	Frequency	%	Cum. %
1	26,199	27.22	27.22	270,076	13.59	13.59
2	21,255	22.08	49.30	363,177	18.27	31.86
3	18,386	19.10	68.40	331,128	16.66	48.53
4	12,751	13.25	81.64	260,940	13.13	61.66
5	7,212	7.49	89.13	179,106	9.01	70.67
6	3,988	4.14	93.28	126,477	6.36	77.03
...	...	...	...	...	...	...
28	11	0.01	100	6,025	0.30	98.56

2.2.1.2 Purchase Units and Market Share of a Product

Following recreational demand literature, this paper takes the number of short-term vacation rental reservations (in other words, the number of trips a recreation site received from visitors) as the number of sales an Airbnb rental unit recorded. By 'short-term', I mean reservations with lengths of stays up to four days, keeping in line with the market definition. Using the number of reservations instead of nights sold will reduce the risk of inflating the market shares of cheap rentals operating many rooms or hostel type Airbnb listings and long-term extended stays for specific purposes. The unit price is defined to be the rental price per night plus cleaning fees, which is the actual transaction price consumers pay.

Frequency of leisure and business travels a person takes in a year is also an important factor for both decision modeling and how many cross sections (time periods) the demand model should include. Without individual level choice data, I rely on a previous market research. According to AARP (American Association of Retired Persons) 2016 and 2017 travel research, Americans across all generations take on average 3.5 domestic leisure trips. I assume a person on average consider visiting NYC Airbnb rentals three times a year.

2.2.2 Summary Statistics

2.2.2.1 Rating Score Inflation

Table 2.4 presents the summary statistics from the dataset used for demand models. Prices and rental unit attributes are from InsideAirbnb.com, a public repository of Airbnb data. Sales records were purchased from Airdna, a data consulting branch firm of Airbnb. The dataset consists of four cross sections: April, 2016, August, 2016, December, 2016, and April, 2017. Rental units without any reviews or sales records were dropped from the analysis. Also, rental units with nightly prices greater than \$5,000 were excluded.

One important issue with the data is review rating score inflation. There are seven categories of review ratings for Airbnb rentals, and Table 2.3 presents the questions Airbnb asks for consumers during the rating process. As can be seen in Table 2.4, rating scores over all categories are near perfection. Rating inflation seems exacerbated in Airbnb particularly compared to other vacation rental portals. Zervas, Proserpio, and Byers (2015) compares review scores of two rental platforms: Airbnb and TripAdvisor. Not only are the average ratings higher for Airbnb listings than TripAdvisor (4.7/5 stars versus 3.8/5 stars), but it is also the case for cross-listed rentals. (0.1/5 stars differences in mean) The authors conjecture that this phenomenon is from reciprocity or fears of retaliation due to the bilateral review policy and strategic manipulation by sellers.

Also, there seems to exist strong collinear relationships among review score categories (see Appendix C.4). To avoid multicollinearity, I use an average of the six rating score categories (2 - 10 scale, 'Ratings Average') rather than overall rating (20 - 100 scale), which was more severely inflated. Due to the time-cumulative nature of review scores, there is a risk of sellers' strategic manipulations; rating scores of a rental unit with a small number of extremely positive reviews cannot be trusted. Hence I drop rental unit observations that both have a 'Ratings Average' greater than 9.99999 and a total number of reservations received during the time periods of our empirical analysis (201604 - 201704) less than 5.

'Superhost' designation implies a rigorous quality certification that cumulates past reputation/feedback performances. A host must have accommodated more than ten parties of guests, maintained a 90% response rate to booking requests or higher, received a five star review, i.e., review scores higher than 80 out of 100 - at least 80% of the time, and completed each confirmed reservations without canceling. 'Verification Accounts' means the number of contact methods a host maintains including emails, phones, and social media network accounts. It implies that hosts with multiple 'Verification Accounts' have gone through the government issued ID check and uploaded self-introduction with photographs. They represent that the basic contract enforceability by Airbnb is active, which is the key prerequisite for reputation/feedback systems to work. (Milgrom, North, and Weingast (1990))

Table 2.3: Definitions for Review Score Categories

Category	Questions Asked in Reviewing Process
Overall Rating	Overall experience
Accuracy	How accurately did the photos and description represent the actual space?
Cleanliness	Did the cleanliness match your expectations of the space?
Check-in/out	How smooth was the check-in process, within control of the host?
Communication	How responsive and accessible was the host before and during your stay?
Location	How appealing is the neighborhood (safety, convenience, desirability)?
Value	How would you rate the value of the listing?

Table 2.4: Summary Statistics

	All Sample		Superhost	Normal Host		
obs: 62,673	Mean	S.D.	Mean	Mean	Min	Max
Price per Night ($)	171.8356	118.7063	187.3653	169.8143	30	4770
Nights Sold	27.4482	24.0788	36.6391	26.2519	1	168
Reservations	11.2973	10.2259	14.7591	10.8468	1	88
Quality Certifications						
Superhost Indicator	0.1152		1	0	0	1
Verification Accounts	4.3243	0.9975	4.4341	4.3100	1	12
Review Rating						
Overall Rating	92.0813	7.4158	96.8795	91.4568	20	100
Accuracy	9.4650	0.7756	9.8828	9.4107	2	10
Cleanliness	9.1121	1.0109	9.7431	9.0300	2	10
Check-in/out	9.6555	0.6690	9.9498	9.6172	2	10
Communication	9.7074	0.6290	9.9701	9.6732	2	10
Location	9.3123	0.8182	9.5590	9.2802	2	10
Value	9.2172	0.7966	9.6784	9.1571	2	10
Review Texts						
Number of Reviews	26.8576	33.2056	41.4124	24.9632	1	380
Negative Reviews	3.5813	4.9652	5.9722	3.2701	0	67
Seller Texts						
Positive Adjectives	2.8890	1.6858	3.2350	2.8440	0	10
Location Words	3.5993	2.3112	4.0855	3.5360	0	15
Accommodation Capacities						
Default Guests	2.7090	1.2622	2.7544	2.7031	1	6
Bathrooms	1.0826	0.3159	1.0786	1.0831	0	5
Additional Guests	1.4769	0.8881	1.5799	1.4635	0	14
Instant Bookable	0.1793	0.3836	0.1607	0.1817	0	1
Amenity and Service						
24 Hour Check-in	0.2921	0.4547	0.3935	0.2789	0	1
Hangers	0.5286	0.4992	0.6730	0.5098	0	1
Heating	0.9570	0.2028	0.9805	0.9540	0	1
Shampoo	0.6499	0.4770	0.7929	0.6313	0	1
(Room Type)						
Entire Home/Apt	0.5099	0.4999	0.5169	0.5089	0	1
Shared Room	0.0279	0.1648	0.0197	0.0290	0	1

* Variables were selected by a lasso variant (Belloni and Chernozhukov (2013)) designed for a successful asymptotic approximation to the objective $ln(s_{jt}/s_{ot})$ with only a subset of all 180 variables. Subsection 2.3.3 for high dimensional metrics introduces the selection principles and inference for the post selection parameter estimates.

Seller texts include rental unit titles, sub-titles, descriptions of various aspects of the rentals, such as neighborhoods, transportation, pros and cons, and etc. The selected accommodation capacities conform to the empirical studies on price determinants of hotels and Airbnb.[6] Amenity and service features in Table 2.4 were in fact, cross-selected by other popular machine learning methods other than Belloni and Chernozhukov (2013)'s lasso.[7]

2.2.2.2 Text Processing

This paper employs n-gram word/phrase extraction (bag of words) and sentiment analysis (classification) using a supervised machine learning to process seller and buyer texts. N-gram bag of words means extracting words/phrases purely according to the frequency of occurrences and use them as regressors. Selected features are often reduced and categorized at a researcher's discretion. For seller texts, 36 words/phrases including up to four words ('Quadrigram') out of the 3,000 most frequently appearing ones were selected among 120,000 advertisement texts. The counts for each rental were summed over the regarding two categories: 'Positive Adjectives' and 'Location Words'.

Table 2.5: Selected Words/Phrases from Airbnb Hosts' Advertisement Texts

Category	Positive Adjectives	Location Words
Unigram	Amazing, Beautiful, Cozy, Friendly, Spacious, ...	Broadway, Manhattan, SoHo Brooklyn, Chelsea, ...
Bigram		Central Park, Columbia University, Hell's Kitchen, Brooklyn Bridge, ...
Trigram		Empire State Building, The G train, Major subway lines, ...
Quadrigram		Metropolitan Museum of Art Museum of Natural History, ...

Supervised machine learning means fitting a function that maps an input to an output based on an example input-output pair dataset. Sentiment classification for review texts includes the following steps. A researcher conducts a pre-processing such as removing non-alphabetical components, e.g., arabic numbers, commas, and etc, trimming white spaces, and converting to lower case letters.

[6] See Wang and Nicolau (2017) for a comprehensive review up to date.

[7] For Belloni and Chernozhukov (2013), see Subsection 2.3.3 and for machine learning, Appendix A.1 and A.2.

A classification machine is then trained on sample reviews, with emotional polarity as outputs and words/phrases as inputs. The choice on words/phrases could either rely on pre-established dictionaries (lexicons) or n-gram words/phrases from sample reviews whichever yields the best in-sample prediction rates. The trained machine is then scaled up on the whole review corpus.

This paper trains a classification machine on a set of 1,000 sample reviews collected from four major U.S. cities other than NYC: Ashevill (NC), Austin (TX), Denver (CO), and Washington D.C. Using 3,500 n-gram bag of words/phrases, multiple supervised machine learning models were constructed and the highest in-sample prediction rate (87%) was achieved with classification tree in 'Caret' R package over Naive Bayes and Support Vector Machine. The classification was then applied to the whole 850,000 NYC Airbnb review texts. Table 2.6 and Table 2.7 provide conceptual examples. One caution with n-gram dictionaries is that they could contain indicators for expressions of no particular meaning (e.g., were not, was indeed) or opposite meaning ('dirty' for positive reviews) for pure prediction performances.

Table 2.6: Example Guest Reviews and Sentiment Labels

Reviews	Raw Texts
Ex.1	This is a ***dirty*** frat house.
(Negative)	No locks other than main building door.
	Dirty toilets. No host present. ***Rotting*** food in the fridge.
Ex. 2	My room at the BPS Hostel was ***clean*** and ***cool***.
(Nonnegative:	The staff and fellow guests were friendly and ***helpful***.
Neutral)	The location is very convenient for local eateries, coffee shops, pubs and deli's.
	However, I do not feel it was good value for money at $72 per day.
	There was no room service,
	I shared a bathroom with upto 8 others and the ***breakfast*** was weak.
Ex. 3	***Great*** location just outside of downtown Asheville.
(Nonnegative:	I stayed here with three other people. ***Plenty*** of space.
Positive)	Mike was very easy to work with, and made sure we had everything we needed.

Table 2.7: Bag of Words Matrix for Example Guest Reviews

Label		breakfast	clean	cool	dirty	great	helpful	however	plenty	rot	
Ex.1		0	0	0	2	0	0	0	0	1	
Ex.2	...	1	1	1	0	0	1	1	0	0	...
Ex.3		0	0	0	0	1	0	0	1	0	

2.3 Model

2.3.1 Potential Guests' Rental Searching Behavior

The dataset used in this paper does not include any individual demographics. The market share of an individual product is extremely small, because for each time periods, there are about 12,000 - 15,000 rental units in operation even after multiple truncation processes. Given such product heterogeneity, it is elusive to find a homogeneous product groups among which the demand analysis could find a targeted insight on substitution patterns, like automobile markets (BLP (1995) and Petrin (2002)) and retail applications (Nevo (2001)).

However, setting up a utility function with a proper description of consumers' choice set formation could suffice to answer the research question with aggregate level data: evaluating welfare implications of information contents from realized purchase decisions. I propose to employ a three level nesting structure based on Airbnb's hierarchial service neighborhood designations and, use 'Room Type' filters as another set of observed attributes, which were found to be the most popular tools for reducing choice sets during web search processes of actual NYC Airbnb guests.

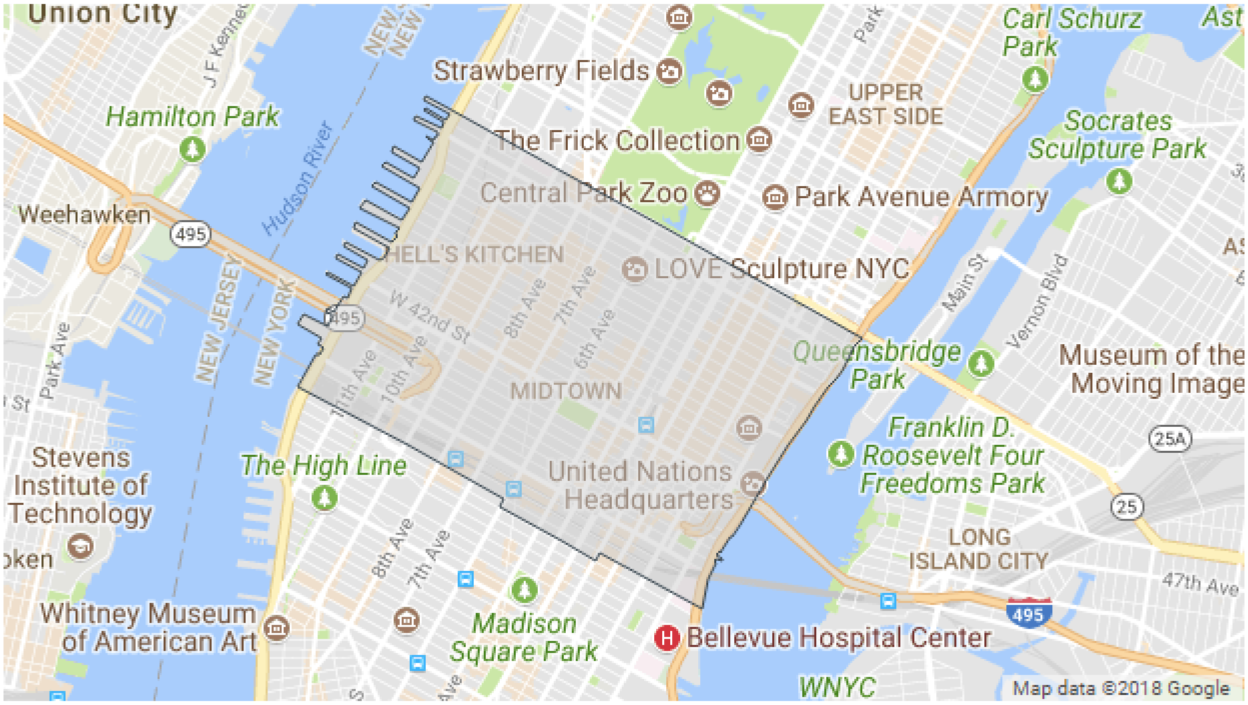
This idea is from Fradkin (2017), who investigates the impacts of search and matching performance of Airbnb platform designs with a detailed consumer web search log data for a major U.S. city between September 2013 and September 2014.[8] A consumer's search is fairly limited in that he/she only sees about 4 to 5% out of over a thousand rental units popping up after the initial search command. During initial search steps, a consumer typically sets the number of guests, which is included in the utility model.

Though limited, a consumer puts a significant amount of effort and time, checking out 88 rental units during 58 minutes on average. Among web searchers who sent a reservation request, more than 64% of them changed the default map location and 50% used the zoom-in/out function to further reduce the choice sets. Figure 2.4 contains an actual Google Maps API example from NYC

[8] The name of the city was anonymized but he says it is the first city Airbnb made a success, which is highly likely to be NYC. The platform is known to be taking off in NYC after the first significant capital investment from Sequoia capital and changing the company's name from Air Bed & Breakfast to Airbnb.

rental unit search on Airbnb.com. It seems that consumers first choose a relatively greater region on the default scale map, and then zoom-in to further narrow down on a neighborhood, wherein he/she picks a rental unit (a three step choice).

Figure 2.4: Neighborhood Designation Example: ’Midtown’ in Manhattan



Airbnb divides NYC area into five big regions: Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Each region is again divided into neighborhoods. For example, ’Midtown’ or ’Harlem’ in Manhattan. Each region contains from 32 to 53 neighborhoods, and the number of listings contained in a neighborhood varies from 2 to more than 5,500. Neighborhood designations are mutually exclusive for all samples and observations, which suggests a hierarchial three level nesting structure (for a comprehensive list, see Appendix C.3).

The first level nesting is based on the five big region indicators, and it is for capturing ’changing default map location’ behaviors. The second level nesting is based on the neighborhood dummies, and it is for capturing zoom-in/out behaviors of consumer search. For the third or rental unit level, I included indicators for ’Entire Home/Apt’ and ’Shared Room’ as attributes based on the finding that 70% of web searchers applied the ’Room Type’ filter.

2.3.2 Nested Multinomial Logit (NMNL) Model

Berry (1994) provides a transformation to estimate a (two level) nested logit model with aggregate level data. Three level nesting structure is an extension by Verboven (1996) and has been adopted in various applications on markets for drugs, automobiles, and agricultural products (Bjornerstedt and Verboven (2016), Brenkers and Verboven (2010), and Ciliberto, Moschini, and Perry (2017)).

I estimate four models. A simple OLS, IV, and two and three level nested logit models. This is done to show stepwise improvements in utility parameter estimates over the model changes to resolve identification issues of price endogeneity and choice set formation. The nesting structures are expected to provide a more accurate modeling for consumers' choice set formation, reducing possible sources of biases due to unobserved variables or decision making principles. The utility function for a Berry (1994) style IV logit model consists of a mean utility term δ_{jt} and an idiosyncratic Type I extreme value error ϵ_{ijt};

$$\begin{aligned} u_{ijt} &= \delta_{jt} + \epsilon_{ijt} \\ &= x_{jt}\beta - \alpha p_{jt} + \xi_{jt} + \epsilon_{ijt} \end{aligned} \tag{2.1}$$

where x_{jt} is the attributes vector, p_{jt} is the per night rental price, and ξ_{jt} captures the unobservables of rental unit j at time period t. Nested logit models impose additional structures on ϵ_{ijt} for each consumer i;

$$u_{ijt} = x_{jt}\beta - \alpha p_{jt} + \xi_{jt} + (\zeta_{igt} + (1 - \sigma)\epsilon_{ijt}) \tag{2.2}$$

$$u_{ijt} = x_{jt}\beta - \alpha p_{jt} + \xi_{jt} + (\zeta_{igt} + (1 - \sigma_2)\epsilon_{ihgt} + (1 - \sigma_1)\epsilon_{ijt}) \tag{2.3}$$

where equation (2.2) and (2.3) represent the utility functions for two and three level nested logit models, respectively. ζ_{igt} captures the impact of nesting 'groups' or in our case the big five regions: Bronx, Brookyln, Manhattan, Queens, and Staten Island ($g = 1, \ldots, G$).

The nesting parameter $0 \leq \sigma < 1$ (for three level nested logit, σ_2) captures how strong the substitution within each group is. For example, if an estimate of $\sigma(\sigma_2)$ is positive and significant, then a tourist is likely to choose rental units in the same region like Bronx, but not in a different region, like Brooklyn. On top of the big regions, the three level nested logit model captures a stronger correlated preferences for units in a neighborhood ($h = 1, ..., H_g$) of a group g with parameter σ_1. The total number of products J is then $\sum_{g=1}^{G} \sum_{h=1}^{Hg} 1_{(j \in h)}$.

ζ_{igt} is common to all products in group g for consumer i and follows a distribution that depends on σ for two level, or σ_1 and σ_2 for three level nested logit model. Cardell (1997) shows that then ζ_{igt} follows a distribution with $(\zeta_{igt} + (1 - \sigma)\epsilon_{ijt})$ or $(\zeta_{igt} + (1 - \sigma_2)\epsilon_{ihgt} + (1 - \sigma_1)\epsilon_{ijt})$ also following extreme value distribution. As the values of nesting parameters approach to zero, i.e., $\sigma(\sigma_1, \sigma_2) \rightarrow 0$, the within group correlation goes to zero and hence the model becomes a simple logit model with a Type I extreme value error. As $\sigma(\sigma_1, \sigma_2) \rightarrow 1$, the within group correlation goes to one.

$$s_{jt}^{OLS(IV)} = \frac{exp(\delta_{jt})}{\sum_{k=0}^{J} exp(\delta_{kt})} \tag{2.4}$$

$$s_{jt}^{NL2} = \frac{exp[\delta_{jt}/(1 - \sigma)]}{\sum_{k \in g} exp[\delta_{kt}/(1 - \sigma)]} * \frac{\left(\sum_{k \in g} exp[\delta_k/(1 - \sigma)]\right)^{1-\sigma}}{1 + \sum_{g=1}^{G} \left(\sum_{k \in g} exp[\delta_k/(1 - \sigma)]\right)^{1-\sigma}} \tag{2.5}$$

$$s_{jt}^{NL3} = \frac{exp[\delta_{jt}/(1 - \sigma_1)]}{exp[I_{hg}/(1 - \sigma_1)]} * \frac{exp[I_{hg}/(1 - \sigma_2)]}{exp[I_g/(1 - \sigma_2)]} * \frac{exp(I_g)}{exp(I)} \tag{2.6}$$

$s_{jt}^{OLS(IV)}$, s_{jt}^{NL2}, and s_{jt}^{NL3} are the resulting stepwise choice probabilities or market shares of a rental unit j for a simple logit, two level nested logit, and three level nested logit models. The inclusive values I_{hg}, I_g, and I for three level nested logit models are defined by:

$$\begin{aligned} I_{hg} &= (1 - \sigma_1) * ln \sum_{k=1}^{J_{hg}} exp[\delta_{kt}/(1 - \sigma_1)] \\ I_g &= (1 - \sigma_2) * ln \sum_{h=1}^{H_g} exp[I_{hg}/(1 - \sigma_2)] \\ I &= ln\left(1 + \sum_{g=1}^{G} exp(I_g)\right) \end{aligned} \tag{2.7}$$

McFadden (1978) gives the condition for nesting parameters to be consistent with the utility theory: $0 \leq \sigma_2 \leq \sigma_1 < 1$ which comes natural in that the correlation of preferences is stronger for rental property choices on a neighborhood level (σ_1) than neighborhood choices out of a big locational segment (σ_2). The inverted aggregate level estimating equations based on (2.4), (2.5), and (2.6) were provided by Berry (1994) and Verboven (1996):

$$ln(s_{jt}/s_{ot}) = x_{jt}\beta - \alpha p_{jt} + \xi_{jt} \tag{2.8}$$

$$ln(s_{jt}/s_{0t}) = x_{jt}\beta - \alpha p_{jt} + \sigma ln(s_{j|gt}) + \xi_{jt} \tag{2.9}$$

$$ln(s_{jt}/s_{0t}) = x_{jt}\beta - \alpha p_{jt} + \sigma_1 ln(s_{j|hgt}) + \sigma_2 ln(s_{h|gt}) + \xi_{jt} \tag{2.10}$$

where s_{0t} is the outside market share at time period t, $s_{j|gt}$ is the market share of rental unit j in region $g = 1, ..., 5$, $s_{j|hgt}$ is j's share in neighborhood h in region g, and finally, $s_{h|gt}$ is the share of all units in neighborhood h in region g.

The idea of aggregate level estimating equations (2.8), (2.9), and (2.10) for identifying utility parameters is similar to regressing ASC (Alternative Specific Constants) on observable attributes in recreational demand literature (Murdock (2006)). Also, though nested logit models partially alleviate the pervasive IIA problem, with individual level data a practitioner can estimate more comprehensive substitution patterns across recreation sites with mixed logit models and consider nested logit as a special case (Herriges and Phaneuf (2002)).

2.3.3 High-Dimensional Attributes and Machine Learning

2.3.3.1 Lasso Selector and Oracle Property

Candidates for attributes in x_{jt} include information contents, accommodation capacities, and 150 binary indicators for amenity and service features. Such a high dimensional characteristic space with many binary indicators originating from extreme product heterogeneity poses a threat of multicollinearity and irrelevant variables. This paper hence assumes sparsity, which is frequently introduced in high dimensional metrics. Sparsity assumption is that given a p-dimensional vector $[x_{jt}, p_{jt}] \in R^p$, there exist $s = o(n) \ll p$ variables that asymptotically capture most of the impacts of all p regressors onto the objective $ln(s_{jt}/s_{ot})$.

A practical implication of sparsity for general linear regression models with Gaussian or heteroskedastic errors ($\epsilon \sim N(0, \sigma)$) is that an econometrician first chooses a set of s variables (the observed model $\hat{M}$) that affects $ln(s_{jt}/s_{ot})$ the most by lasso and then do OLS only with the selected variables. Such OLS post lasso, with theoretically suggested conditioning parameters for the first step lasso selector achieves a 'successful' asymptotic approximation to the 'true' $ln(s_{jt}/s_{0t})$ objective function (Belloni and Chernozhukov (2013), Chernozhukov, Hansen, and Spindler (2015), Belloni, Chernozhukov, and Wang (2014)).

Specifically, the risk minimization problem of balancing bias and variance for demand estimation with sparsity can be stated as the following (Belloni and Chernozhukov (2013)).

$$\min c_s^2 + \sigma^2 \frac{s}{n} \tag{2.11}$$

$$c_s^2 = \min_{dim(\beta,\alpha)\leq s} E[(ln(s_{jt}/s_{0t}) - x_{jt}\beta + \alpha p_{jt})^2]$$

$c_s^2 + \sigma^2 \frac{s}{n}$ is the upper bound of the risk for the best market share estimator using only $s \ll p$ covariates. This 'oracle risk' is achieved if the first stage lasso chose the correct s variables which by sparsity assumption captures the most of the impacts of all p regressors. Then the resulting 'oracle rate' of error convergence rate is given by $\sqrt{s/n}$.

One important appeal of OLS post lasso is that even if lasso selector gives only a subset of s covariates, post selection OLS estimator still achieves the 'near oracle rate' of $\sqrt{s * log(p)/n}$. In other words, $\sqrt{E[(log(s_{jt}/s_{0t}) - x_{jt}^{\hat{M}}\beta^{\hat{M}} + \alpha p_{jt})^2]} = O_p(c_s + \sigma\sqrt{s * log(p)/n})$, where $x^{\hat{M}}$ and $\beta^{\hat{M}}$ represent the vector of attributes chosen by lasso (the observed selected model $\hat{M}$) and the corresponding post selection OLS coefficients.

A lasso selection to get $\hat{M}$ (including price p_{jt}) means choosing variables of non-zero coefficients in solving the following penalized regression problem. Letting $\beta' = [\beta, \alpha]$,

$$\hat{\beta'} = argmin_{\beta' \in R^p}\, \hat{Q}(\beta') + \frac{\lambda}{n}||\hat{\Psi}\beta'||_1 \tag{2.12}$$

$$\hat{Q}(\beta') = \frac{1}{n}\sum_n (ln(s_{jt}/s_{ot}) - x_{jt}\beta + \alpha p_{jt})^2$$

where $||\beta'||_1 = \sum_{l=1}^{p-1} |\beta_l| + |\alpha|$ and $\hat{\Psi} = diag(\hat{\psi}_1, ..., \hat{\psi}_p)$. The penalty loadings $\hat{\Psi}$ and penalty level λ for post OLS oracle rate in the heteroskedastic case are;

$$\hat{\psi}_k = \sqrt{\frac{1}{n}\sum_n (x_{ik}^2 \hat{\epsilon}_i^2)} \tag{2.13}$$

$$\lambda = 2c\sqrt{n}\Phi^{-1}(1-\gamma/(2p))$$

where Φ denotes the cumulative standard normal distribution and $\hat{\epsilon}$ is an empirical estimate of errors (residuals). The suggested preset values for c and γ is 1.1 and 0.1. $\hat{\Psi}$ and λ for homoskedastic errors result in similar variable selection results.

The attributes in the summary statistics (Table 2.4) were in fact chosen by this process using R package **hdm**. The observed model $\hat{M}$ is stable over a range of c from 0.9 and 1.3 with 0.05 increments. To check the validity of selection results, four other data-driven machine learning models were estimated: lasso, ridge, elastic net, and gradient boosting. They focus on the prediction accuracy (reducing RMSE) rather than 'oracle rate'. Again, all attributes in Table 2.4 were unanimously chosen by all and hence included in x_{jt}.[9]

2.3.3.2 Post Selection Inference

However, if one proceeds to OLS with the selected (observed) model $\hat{M}$ by lasso, there are two possible pitfalls. First is that classical inferences (confidence intervals and p-values) on $\hat{\beta}^{\hat{M}}$ are no longer valid. It is because of the non-selected (omitted) variables, making the post selection OLS only with the attributes in $X^{\hat{M}}$ biased. Though the asymptotic distribution of lasso coefficients for our case of $n \gg p$ is available (Fu and Knight (2000)), an exact post selection inference for OLS post lasso is the primary target of interest.

Such 'post selection inference' after variable selection with machine learning is a relatively new and still developing area. This paper follows Lee, L. Sun, Sun, and Taylor (2016) which provides an exact distribution of post selection OLS estimates and hence, exact confidence intervals (C.Is), p-values, and tail areas. The idea is that given a response $y \sim N(\mu, \sigma^2 I_n)$, the model selection

[9] See Appendix A.1 and A.2 for details on the methodologies for data-driven machine learning.

event $\{\hat{M} = M\}$ by lasso can be expressed as a form of polyhedron $\{Ay \leq b\}$ which once again can be transformed into an interval with low and upper endpoints being functions of residuals z_j of y in the direction of x_j, $\{\nu^{-}(z) \leq y \leq \nu^{+}(z)\}$. Due to the independence between y and z_j, the distribution of an individual coefficient $\hat{\beta}_j^{\hat{M}}$ (a linear transformation of y) from OLS conditional on the lasso selection is a truncated normal.

One advantageous fact about Lee et al. (2016) is that a practitioner can produce exact C.Is and p-values with a fixed penalty parameter λ'. Specifically, the lasso formulation for the exact post selection inference is the original data-driven lasso by Tibshirani (1996).

$$\hat{\beta} = argmin_{\beta \in R^p} \hat{Q}(\beta) + \lambda' ||\beta||_1 \tag{2.14}$$

Therefore, with a range of values of λ' that produces the same model $\hat{M}$ including variables of non-zero coefficients from the penalized regression problem in equation (2.12), a practitioner can produce exact inference for OLS post lasso, still achieving c_s asymptotically.

2.3.3.3 Cautions on Endogeneity and Post Selection Estimator

The second concern with OLS post lasso is endogeneity. In fact, endogeneity lurks under both lasso selection and the subsequent demand estimation with the chosen model. Lasso selector (equation (2.12) and (2.14)) uses Gaussian or at best, heteroskedastic errors implicitly assuming there is no endogeneity due to omitted/unobserved variables. The post selection estimating equations (2.8), (2.9), and (2.10) under OLS structure, could suffer endogeneity in prices and group shares for nesting structures ($s_{j|gt}$, $s_{j|hgt}$, $and\ s_{h|gt}$).

High dimensional econometricians have provided post selection IV regressions after a variable selection on both many controls and instruments with a small number of key endogenous variables such as treatment/policy indicators or prices (Chernozhukov, Hansen, and Spindler (2015) and Chernozhukov et al. (2018)). But still, to the best of my knowledge, a variable selection approach under the presence of unobserved variables coupled with post selection estimation and inference has not been established well. It is understandable in that unobservables are not in the dataset, and we resort to the magnitudes of in-sample prediction errors such as RMSE to choose the 'right' subset of all covariates.

The same concern arises with the conventional alternatives such as principal component analysis (PCA), Akaike information criterion (AIC), or Bayesian information criterion (BIC), in addition being practically infeasible with hundreds of variables to consider. PCA coefficients are linear combinations of covariates which makes it impossible to isolate and identify coefficients for individual variables, and the stepwise nature of AIC and BIC dictates too high calculation costs for estimation and comparison to incur given 150 binary indicators.

This paper does not attempt to provide an analytic methodology for the first step model selection by lasso under the presence of unobservables, but shows that the variable selection results vary when the dataset include variables of possible sources of endogeneity. Specifically, lasso selection was conducted on four different datasets for each estimation methods: simple OLS logit, IV logit (with instrumented price $\hat{p}_{jt}$, equation (2.15)), two level, and three level nested logit models (also with $\hat{p}_{jt}$, equation (2.16) and (2.17)).

$$ln(s_{jt}/s_{ot}) = x_{jt}\beta - \alpha\hat{p}_{jt} + \xi_{jt} \tag{2.15}$$

$$ln(s_{jt}/s_{0t}) = x_{jt}\beta - \alpha\hat{p}_{jt} + \sigma ln(\hat{s}_{j|gt}) + \xi_{jt} \tag{2.16}$$

$$ln(s_{jt}/s_{0t}) = x_{jt}\beta - \alpha\hat{p}_{jt} + \sigma_1 ln(\hat{s}_{j|hgt}) + \sigma_2 ln(\hat{s}_{h|gt}) + \xi_{jt} \tag{2.17}$$

For OLS logit, the dataset for lasso selection contains all attributes except variables for nesting structures. For IV logit, p_{jt} was replaced with the instrumented price $\hat{p}_{jt}$. $ln(\hat{s}_{j|gt})$ was added for two level nested logit, and $ln(\hat{s}_{j|hgt})$ and $ln(\hat{s}_{h|gt})$ for three level nested logit. Group shares were instrumented due to endogeneity concerns proposed by Berry (1994). The selected model $\hat{M}$ differs over datasets, gauging a suspicion on the instability of variable selection results due to unobserved variables. 'Location Words' was not selected in OLS logit case, and for three level nested logit case, a few key parameters including σ_2 for the precinct level correlated preferences were not chosen.[10]

[10] Hence for three level nested logit, I proceed with x_{jt}'s selected in IV and two level nested logit case.

Actual estimation with equation (2.15), (2.16), and (2.17) relies on the moment condition $E[\xi_{jt}|z_{jt}]$, using the separable unobservables term ξ_{jt} and z_{jt} including the selected observables x_{jt} and instruments, following Berry (1994) and BLP (1995). It is a simple two step least squares with the first stage regressions to produce $\hat{p}_{jt}$ and group shares. For instruments, I used variables related to supply decisions: lagged base per night rental price (without 'Cleaning Fee', recorded one year before), starting date of an Airbnb host's rental business, long term availabilities (30, 60, 90, and 365 days), and cancellation policy.[11]

Table 2.8 and 2.9 report the post selection estimation results. Exact C.Is and tail areas reflecting additional uncertainty due to lasso were produced using Lee et al. (2016).[12]

2.4 Results

2.4.1 Parameter Estimates

The main interest of this paper lies on evaluating the value of information contents produced by the platform, consumers, and sellers. Also, the own and cross price elasticities of the new accommodation products could provide an insight on consumers' substitution patterns. For such purposes, it is important to check if the econometric (structural) modeling approach controls endogeneity properly. Together with instrumenting prices and group shares for the key parameters in calculating compensating variations and elasticities, the nesting structures were introduced to target consumers' realistic choice set formation which was expected to reduce biases in utility parameter estimates.

Strong evidence of endogeneity with the simple OLS logit model can be found in the coefficients of 'Room Type' indicators for 'Entire Home/Apt' and 'Shared Room' (Table 2.9). More than 50% of the total NYC Airbnb rentals are 'Entire Home/Apt' and they occupy more than 48% share of total reservations and enjoy a significant amount of price premium. Hence a highly negative

[11] See Appendix C.2 for a detailed discussion on instruments and first stage regressions.

[12] See Appendix A.3 and C.1 for the methodology and comparison between OLS inference and Lee et al. (2016)

coefficient for 'Entire Home/Apt' indicator gauges a suspicion of endogeneity due to unobservables or misspecification in consumers' decision making principles. Also, a positive and significant coefficient for 'Shared Room' looks strange given that it represents the lowest grade 'Room Type' occupying only 2.8% of total listings in the dataset.

For IV logit model, even after instrumenting prices parameter estimates seem to be inflated in overall scale compared to both OLS and nested logit models. This is not specific to the set of instruments reported in the Appendix C.2 (first stage regressions), but fairly stable over various sets of instruments tried. It seems to indicate that there are unobserved variables that significantly affect consumer choices.

The nesting parameters show relatively high coefficients and statistical significance; the Z-scores for σ, σ_1, and σ_2 are 99.67, 83.15, and 4.33, respectively. It would be safe to say that the imposed nesting structures were able to capture consumers' preference for location, as observed in web search log data. The condition $0 \leq \sigma_2 \leq \sigma_1 < 1$ is satisfied, showing that the results are consistent with the random utility theory (McFadden (1978)).

The nesting parameters capturing either precinct or neighborhood preferences alleviate the IIA problem of the simple logit as will be shown in Subsection 2.4.2. The cross price elasticities using σ, σ_1, and σ_2 for nested logit models show that a consumer's substitution across rental units is confined within his/her geographical choice set in mind. To get a more comprehensive picture on substitution patterns, mixed logit models with individual level choice data or BLP type random coefficients could be useful for future research.

Interpretation of individual coefficients are quite straightforward with the standard formulas for either (maximum) willingness to pay (WTP) or attribute elasticities. WTP for a unit increase in attribute k is β_k/α or, the coefficient of a factor divided by the price coefficient. For example, the willingness to pay for one point increase in 'Ratings Average' in the three level nested logit case is $\frac{0.0810}{0.0039} = \$20.7692$. Attribute elasticities can be obtained by $\beta_k x_{jk}(1 - s_{jt})$. Given that the market share of a single rental unit j is extremely small, one could approximately use $\beta_k x_{jk}$. If a rental unit j has a 'Ratings Average' of 9, then the demand elasticity with respect to 'Ratings Average' is about $0.0810 * 9 = 0.7290$. Table 2.10 presents WTP and demand elasticities (average) with respect to each attributes listed in Table 2.8, namely the information variables of main interest.

Table 2.8: Demand Parameter Estimates (1): Price, Nesting, and Information

Obs: 62,673	OLS Logit	IV Logit	2 Level NL	3 Level NL††
Objective	$ln(s_{jt}/s_{0t})$			
Price	-0.0008***	-0.0156***	-0.0059***	-0.0039***
	(0.0000)	(0.0002)	(0.0002)	(0.0003)
Nesting Parameters				
$\sigma(\sigma_1)$			0.7505***	0.8125***
			(0.0075)	(0.0098)
σ_2				0.2293***
				(0.0529)
Quality Certifications				
Superhost Indicator	0.2133***	0.3378***	0.0649***	0.0873***
	(0.0119)	(0.0115)	(0.0110)	(0.0112)
Verification Accounts	0.0479***	0.0842***	0.0663***	0.0535***
	(0.0036)	(0.0035)	(0.0033)	(0.0035)
Consumer Review				
Ratings Average	0.1938***	0.3992***	0.1220***	0.0810***
	(0.0067)	(0.0069)	(0.0070)	(0.0081)
Number of Reviews	0.0171***	0.0161***	0.0045***	0.0036***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Negative Reviews	-0.0381***	-0.0386***	-0.0118***	-0.0091***
	(0.0018)	(0.0017)	(0.0016)	(0.0016)
Seller Texts				
Positive Adjectives	-0.0125***	-0.0591***	-0.0353***	-0.0206***
	(0.0022)	(0.0023)	(0.0022)	(0.0026)
Location Words†		0.0100***	0.0168***	0.0064***
		(0.0016)	(0.0015)	(0.0018)

***: 1% significant, **: 5%, *: 10%, standard errors in parentheses

†: 'Location Words' was not selected by lasso procedure on the dataset for OLS, of all attributes and p_{jt} except for group shares for nesting structures

†† : 'Location Words' and $ln(s_{h|gt})$ for σ_2 were not selected by lasso on the dataset for three level nested logit, of all observable attributes, $\hat{p}_{jt}$, $ln(\widehat{s_{j|hgt}})$, and $ln(\widehat{s_{h|gt}})$. Hence I estimated three level nested logit model with variables selected by lasso in IV and two level nested logit case.

Table 2.9: Demand Parameter Estimates (2): Amenity and Service Features

Obs: 62,673 Objective	OLS Logit	IV Logit $ln(s_{jt}/s_{0t})$	2 Level NL	3 Level NL
Accommodation Capacities				
Default Guests	0.0971***	0.3768***	0.1026***	0.0781***
	(0.0039)	(0.0052)	(0.0055)	(0.0060)
Bathrooms	0.0604***	0.7227***	0.3114***	0.1886***
	(0.0118)	(0.0141)	(0.0137)	(0.0184)
Additional Guests	0.0041	0.1725***	0.0706***	0.0454***
	(0.0048)	(0.0051)	(0.0048)	(0.0055)
Instant Bookable	0.4807***	0.3327***	0.0123	0.0398***
	(0.0096)	(0.0094)	(0.0093)	(0.0097)
Amenity and Service				
24 Hour Check-In	0.1318***	0.1768***	0.0566***	0.0291***
	(0.0090)	(0.0086)	(0.0081)	(0.0086)
Hangers	0.1437***	0.1415***	0.0346***	0.0278***
	(0.0084)	(0.0080)	(0.0075)	(0.0075)
Heating	0.0689***	0.2053***	0.1046***	0.0725***
	(0.0181)	(0.0174)	(0.0162)	(0.0165)
Shampoo	0.1424***	0.2292***	0.0752***	0.0413***
	(0.0081)	(0.0078)	(0.0074)	(0.0082)
(Room Type)				
Entire Home/Apt	-0.0837***	1.4138***	0.7256***	0.4182***
	(0.0235)	(0.0212)	(0.0209)	(0.0373)
Shared Room	0.2015***	-0.1068***	-0.0761***	-0.0413**
	(0.0223)	(0.0217)	(0.0201)	(0.0204)
Constant	-13.4340***	-15.5229***	-5.4277***	-5.9737***
	(0.0694)	(0.0688)	(0.1198)	(0.1317)

But WTP and attribute elasticities do not take into consideration the supply side responses due to the unit changes in attributes. For instance, if 'Ratings Average' decreases by one unit, not only does a consumer's WTP decreases, but also a seller's price premium does due to the reduction in reputation scores. Also, a practitioner should use nesting parameters σ, σ_1, and σ_2 in calculating consumer surpluses (utility before and after a unit change in attributes) and the resulting compensating variations to get more realistic welfare measures for variables of interest.

Table 2.10: WTP and Factor Elasticities for Information Variables

	2 Level NL		3 Level NL	
	WTP ($)	Elasticity	WTP	Elasticity
QUALITY CERTIFICATIONS				
Superhost Indicator	10.9470	0.0075	22.5587	0.0101
Verification Accounts	11.1819	0.2866	13.8191	0.2312
CONSUMER REVIEW				
Ratings Average	20.5807	1.1479	20.9310	0.7621
Number of Reviews	0.7633	0.1215	0.9209	0.0957
Negative Reviews	-1.9845	-0.0421	-2.3552	-0.0327
SELLER TEXTS				
Positive Adjectives	-5.9596	-0.1020	-5.3379	-0.0600
Location Words†	2.8321	0.0604	1.6474	0.0229

Willingness to pay (WTP) was calculated using the formula β_k/α. Factor elasticities ($\beta_k x_{jk}(1 - s_{jt})$) are the mean values of all observations. The small magnitude of demand elasticity with respect to 'Superhost' indicator is due to the fact that only about 11% of total rental units were designated as 'Superhost'.

Overall, the nesting structures reduce the magnitudes of coefficients and imply a realistic impacts of each for consumers' purchase decision making processes. Amenity and service features chosen by multiple ML methods seem to show significant impacts on purchase decisions both in statistical and economic senses. However, demand parameters cannot, by themselves tell much about substitution patterns and welfare implications ($ metric). To investigate this paper's research question of evaluating and comparing information contents on product quality, appropriate formulas should be applied.

2.4.2 Elasticities and Welfare Measures

The own price elasticities for the simple IV logit is $\alpha(1 - s_{jt})p_{jt}$, and the formulas for nested logit models are presented in equations (2.18).

$$\begin{aligned} \frac{\partial q_{jt}}{\partial p_{jt}} * \frac{p_{jt}}{q_{jt}}^{NL2} &= \alpha(s_{jt} - \frac{1}{1-\sigma} + \frac{\sigma}{1-\sigma} s_{j|gt})p_{jt} \\ \frac{\partial q_{jt}}{\partial p_{jt}} * \frac{p_{jt}}{q_{jt}}^{NL3} &= \alpha(s_{jt} - \frac{1}{1-\sigma_1} + (\frac{1}{1-\sigma_1} - \frac{1}{1-\sigma_2}) s_{j|hgt} + \frac{\sigma_2}{1-\sigma_2} s_{j|gt})p_{jt} \end{aligned} \tag{2.18}$$

It turns out that the demand for Airbnb rentals in NYC is quite elastic (3.5365 to 4.0817), which is not a surprise given the product heterogeneity and severe competition in a densely populated urban area. Cross price elasticities involve multiple cases due to the nesting structures. First possibility is that product j and k are in the same big region (for two level nested logit) or in the same neighborhood (for three level nested logit). The formulas for this case (Case 1 in Table 2.11) are as follows.

$$\begin{aligned}
\frac{\partial q_{jt}}{\partial p_{kt}} * \frac{p_{kt}}{q_{jt}}^{NL2} &= \alpha(s_{jt} + \frac{\sigma}{1-\sigma} s_{j|gt}) p_{jt} \qquad (2.19) \\
\frac{\partial q_{jt}}{\partial p_{kt}} * \frac{p_{kt}}{q_{jt}}^{NL3} &= \alpha(s_{jt} + (\frac{1}{1-\sigma_1} - \frac{1}{1-\sigma_2}) s_{j|hgt} + \frac{\sigma_2}{1-\sigma_2} s_{j|gt}) p_{jt}
\end{aligned}$$

The cross price elasticities between substitutes j and k are negligible for products in the same big region for two level nested logit model. It is because of the fact that the two big regions 'Brooklyn' and 'Manhattan' occupies nearly 40% and 50% of total rental units in the data respectively, making the precinct effect $\frac{\sigma}{1-\sigma} s_{j|gt}$ or $\frac{\sigma_2}{1-\sigma_2} s_{j|gt}$ minuscule. It is hard to expect that cross price elasticities would be as large as that of Coke and Pepsi given there are about 25,000 to 30,000 alternatives. Other neighborhoods also contain many alternatives. 'Queens', 'Bronx', and 'Staten Island' contain 5,480, 980, and 378 rental units inside, respectively.

Table 2.11: Price Elasticities

	2 Level NL	3 Level NL	
Own Price Elasticities			
Mean	4.0817	3.5365	
S.D.	2.8200	2.4439	
Cross Price Elasticities	Case 1	Case 1	Case 2
Mean	0.0007	0.0286	0.0001
S.D.	0.0020	0.1618	0.0001
Min	0.0000	0.0000	0.0000
Max	0.0763	6.3240	0.0050

On the other hand, cross price elasticities for three level nested logit models show more reasonable values and greater variations though the mean is still small (0.0268). The cross price elasticities range from almost zero to 6.3240, reflecting the fact that there are neighborhoods such as 'Midtown' in Manhattan with more than 4,000 substitutes and pretty small ones with only a few competitors. The 'Room Type' filters, travel dates and host availabilities, number of rooms and guests, and maximum price filters still leave at least couple of hundred alternatives to consider in a popular neighborhood. To identify more refined choice set formation of consumers, an econometrician may need an individual level data.

The second possibility is that product j and k are in different neighborhoods but in the same big region (for three level nested logit only). The formula for this case (Case 2 in Table 2.11) is the same as equation (2.19) (Case 1 for two level nesting) with σ replaced with σ_2.

The estimates seem to suggest that a substitution between products in different neighborhoods is not a realistic option for NYC Airbnb tourists. Such small cross price elasticities is because of still a large number of alternatives in a precinct that contains the neighborhoods rental unit j and k reside, similar to the two level nested logit case.

The last possibility is when product j and k are in different big regions. Then the cross elasticities reduce to the simple logit case $\alpha s_{jt} p_{jt}$, which are close to zero meaning a negligible substitution among rental units far away from the locational preference of a consumer.

Table 2.12: Compensating Variations over Counterfactual Scenarios

	2 Level NL		3 Level NL			
			Average		Total (million)	
Categories / Scenarios	-1	without	-1	without	-1	without
Quality Certifications						
Superhost Indicator	-0.5868		-1.7270		-1.2232	
Verification Accounts	-7.0991	-25.9670	-9.5921	-40.6014	-6.7941	-28.7583
Consumer Reviews						
Ratings Average	-5.7275	-51.9953	-6.1465	-54.4118	-4.3536	-38.5403
Negative Reviews	1.2637	3.5131	1.6230	3.6848	1.1496	2.6100
Seller Texts						
Positive Adjectives	2.2300	6.3000	1.8513	5.1584	1.3113	3.6537
Location Words	-1.7784	-6.3085	-0.8378	-3.0022	-0.5934	-2.1265

The welfare measure is compensating variation which takes the following general form.

$$CV_i = \frac{1}{\alpha}(CS_i^{after} - CS_i^{before}) \tag{2.20}$$

where α is the price coefficient or the marginal utility of income. CS_i^{after} and CS_i^{before} represent consumer surplus after and before the counterfactual experiments, respectively. The expressions for consumer surplus for nested logit models are

$$CS_i^{NL2} = log[1 + \sum_{g=1}^{G}(\sum_{k \in g} exp[\delta_{jt}/(1-\sigma)])^{1-\sigma}] \tag{2.21}$$

$$CS_i^{NL3} = log[1 + \sum_{g=1}^{G} exp(I_g)]$$

where the inclusive values for three level nested logit models are presented below for the purpose of easier stepwise understanding and actual computations.

$$I_{hg} = (1-\sigma_1) * log \sum_{k=1}^{J_{hg}} exp[\delta_{kt}/(1-\sigma_1)] \tag{2.22}$$

$$I_g = (1-\sigma_2) * log \sum_{h=1}^{H_g} exp[I_{hg}/(1-\sigma_2)]$$

There are clear limitations in the counterfactual experiments of this paper. Due to the difficulty in supply side modeling, I cannot produce a complete description of market equilibrium before and after the counterfactual scenarios including changes in prices, quantities, and product offerings. I leave this task to future research with more data on the heterogeneous Airbnb rental unit owners. Instead, I generate counterfactual prices by an OLS hedonic regression results, following Hausmann and Leonard (2002). The price responses for a unit change in 'Superhost Indicator', 'Verification Accounts', 'Ratings Average', 'Negative Reviews', 'Positive Adjectives', and 'Location Words' are \$8.3759, \$2.4799, \$13.8010, -\$0.4440, -\$3.1582, and \$0.6599, respectively.

Table 2.12 reports compensating variations from two counterfactual scenarios. First is a unit reduction (-1) in each information variables. Second is comparing situations with and without one of the information contents. The latter approach is for controlling the different measurement scales of each information variables, and following Lewis and Zervas (2016)'s study on the impacts of reviews in hotel markets. The induced changes in price using the estimates from the hedonic price regressions are reflected in calculating CS_i^{after} together.

The 'dollar metric' from counterfactual experiments confirms the hypothesis of this paper that trustworthy information on product quality is important in sustaining a market with a high degree of information asymmetry. Enforced quality certifications and verifiable ex-post review contents turn out to be more influential than non-verified seller side 'Cheap Talk'. Specifically, the host identity verification measures ('Verification Accounts') show a greater dollar impact on purchase decisions in either case of unit reduction or complete absence ($9.5921/$40.6014). The lower value for 'Superhost' ($1.7270) is originating from the fact that only about 11% of rental units get affected by the counterfactual scenario.

CV for consumer review ratings also show greater impacts on consumer choices than those of seller side textual voluntary disclosures. Compensating variations for 'Ratings Average' are $6.1465/$54.4118, both higher than those of 'Positive Adjectives'($1.8513/$5.1584) and 'Location Words' ($0.8378/$3.0022) from advertisement texts. CVs for 'Negative Reviews' ($1.6230/$3.6848) from review texts do not show particularly more dominant impacts.

Given that there were 708,308 reservations in the sample during the time periods of our empirical analysis, the aggregate dollar values of consumer welfare from each information contents would be quite huge. For the case of a unit reduction, the welfare impacts are $1.2232, $6.7941, $4.3536, and $1.1496 million for 'Superhost', 'Verification Accounts, 'Ratings Average', and 'Negative Reviews', respectively. On the seller side, $1.3113 and $0.5934 million are for 'Positive Adjectives' and 'Location Words'. In the case of total absence, they are $1.2232, $28.7583, $38.5403, $2.6100, $3.6537, and $2.1265 million in the same order.[13]

One caution for the interpretation of positive CV signs for 'Negative Reviews' and 'Positive Adjectives' is that they represent the increase in consumers' demand for rentals with a unit lower 'Negative Reviews' and 'Positive Adjectives'. In fact, 'Positive Adjectives' seems to be a strong

[13] Lewis and Zervas (2016) estimated the welfare impacts of online review ratings for hotel markets over five U.S. states over 10 years of time periods as about $546 million with hedonic price regression adjustments for the counterfactual case of total absence.

indicator for cheap and low quality Airbnb rental units, showing negative signs on coefficients both for demand and hedonic models. On the other hand, 'Location Words' such as '5 min walk to Central Park' and 'A Walking Distance from Grand Central' are usually verifiable instantly on the Google Maps API on each Airbnb listing webpages, which is more credible and hence attracts more consumers.

2.5 Conclusion

This paper investigates how the sharing economy platform Airbnb could overcome adverse selection due to information asymmetry. The risk of adverse selection for P2P markets is expected to be higher than online retail outlets for material goods because the accommodation service transactions among anonymous non-professional individuals imply a higher degree of information asymmetry and more than just monetary losses. To test the insight from information economics that enforced and ex-post verifiable information on product quality is more influential for a consumer's decision making process, demand models were estimated.

Predominant identification challenges due to high dimensionality in attribute space were partly resolved using the variable selection by a lasso variant and exact post selection inferences. However, the model selection was unstable once endogeneity is involved and the results show that an appropriate econometric (structural) modeling approach designed to capture actual consumers' decision making principles is essential to produce more accurate utility parameters. Unstructured text information on product quality was incorporated in the model and showed nonnegligible impacts.

The results confirm our hypothesis, with quality certifications and consumer review ratings showing greater impacts on rental choices than non-verified seller side voluntary disclosures via textual advertisements.

CHAPTER 3

ESTIMATION FOR THE DISTRIBUTION OF RANDOM COEFFICIENTS WITH HETEROGENEOUS AGENT TYPES: MONTE-CARLO SIMULATION

3.1 Introduction

Since Berry, Levinsohn, and Pakes (1995), Nevo (2001), and Petrin (2002), random coefficients logit models to capture heterogeneous consumer preferences have been one of the most popular frameworks for demand research. But the estimation routine is highly nonlinear, computationally burdensome, and in some cases the convergence is not guaranteed. Even if individual choice data is available, (simulated) maximum likelihood estimation for random coefficients usually incurs too much calculation costs, which is not an attractive option to applied researchers working with more than millions of transaction records.

Fox, Kim, Ryan, and Bajari (2011, henceforth FKRB) proposes an alternative, that is nonparametric, computationally simple, easy to program, and easy to combine auxiliary methods due to its least squares format. To give a concrete idea, consider a simple logit choice probability given the binary outcome y_{ij}, attributes vector x_{ij}, and the random coefficients β_i, where i and j are indicies for individual consumers and products, respectively.

$$Pr(y_{ij} = j|x) = \int \frac{exp(x_{ij}^{'}\beta_i)}{1 + \sum_{j'=1}^{J} exp(x_{ij'}^{'}\beta_i)} dF(\beta_i) \tag{3.1}$$

Assuming there are $r = 1, \ldots, R$ types of consumers, i.e., R fixed preference parameters $\beta_1, \ldots, \beta_R$, the choice probability of choosing product j can be expressed as an weighted average with the probability tuple $\theta = (\theta_1, \ldots, \theta_R)$.

$$Pr(y_{ij} = 1|x) = \sum_{r=1}^{R} \theta^r \frac{exp(x_{ij}^{'}\beta^r)}{1 + \sum_{j'=1}^{J} exp(x_{ij'}^{'}\beta^r)} \tag{3.2}$$

Then the parameters enter the estimating moments linearly, and the main interest is to estimate the tuple θ. From the estimated tuples $\hat{\theta}$, a practitioner can also estimate the empirical joint and marginal distributions of random coefficients β. The inequality constraints for θ is also simple, required as a natural condition for a probability vector: $\sum_{r=1}^{R} \theta^r = 1$ and $\theta^r \geq 0$ for all r.

FKRB (2011) demonstrates that estimation for θ and $F(\beta)$ using the reparametrization specified as in equation (3.2) is consistent. The estimator can be applicable to a wide range of nonlinear models, but this paper focuses on the multinomial logit demand case. For more rigorous theoretical discussion, see Fox, Kim, Ryan, and Bajari (2012) and Fox, Kim and Yang (2016). FKRB(2011) is closely related to latent class models in discrete choice literature (Green (1976) and Train (2003)).

One possible weaknesses of FKRB (2011) is that, as demonstrated in their Monte-Carlo simulation results, the approximating performances of $\hat{F}(\beta)$ can deteriorate as the number of consumer types R grows. Also, there is a possibility that there are some 'nuisance' consumer types that can cause poor estimation results for $\hat{F}(\beta)$. It is a similar environment where there are too many irrelevant regressors in linear regressions. One can expect that appropriate dimensionality reduction techniques can improve the approximation performances, along with significant gains in computation speed.

To examine such a possibility, this paper tries to reduce the dimensionality in consumer heterogeneity by introducing high-dimensional metrics (Belloni and Chernozhukov (2013)). The baseline estimator based on equation (3.2) can be expressed as a linear regression with a design matrix of size $NJ * R$, where N is the number of observations (consumer choices) and J is the number of choice alternatives. I apply the lasso variant first to reduce R, and with $R^*(\leq R)$, construct a new design matrix of size $NJ * R^*$ and compare the performance metrics to measure the distances between the true CDF $F_0(\beta)$ and $\hat{F}(\beta)$. I also try the original lasso formulation by Tibshirani (1996) with 10 folds cross validation.

The lasso variant developed by Belloni and Chernozhukov (2013) guarantees the asymptotic approximation performances of post-lasso least squares estimators. It is one of the first high-dimensional metrics or machine learning application that started to be accepted in economics, with application areas including demand estimation, treatment/policy impacts, and general linear models (Belloni, Chernozhukov, and Wang (2014), Chernozhukov, Hansen, and Spindler (2015), and Chernozhukov et al. (2018)). In statistics, post-selection estimators and inference using popular machine learning methods other than lasso such as ridge regression, elastic net, and tree based models (boosting) have been actively investigated.

In Monte-Carlo experiments (Section 3.4), post-lasso estimators show better approximation performances compared to the baseline estimator. It is stable once the number of consumer types R exceeds 36. The estimated CDFs of β produced by post-lasso estimators also track the (simulated) 'true' distributions better, and this can be attributed to the variable selection process 'killing' nuisance variables so that $\hat{F}(\beta)$ does not take extreme values. Hence the combination of the baseline inequality constrained least squares and high-dimensional metrics can be a good alternative to estimate random coefficients logit demand models with 'Big Data' in various marketplaces.

The rest of this paper is organized as follows: Section 3.2 briefly introduces the multinomial random coefficients logit model, the baseline estimator from FKRB (2011), and the lasso variant by Belloni and Chernozhukov (2013). Section 3.3 outlines the Monte-Carlo designs to compare baseline estimator and post-lasso estimator. Section 3.4 reports estimation results and figures for the marginal empirical distributions of β_1 to compare the approximation performances obtained using baseline and post-lasso estimators.

3.2 Model

3.2.1 Multinomial Random Coefficients Logit Demand Model

This section lays out the multinomial random coefficients logit demand model, which is one of the key motivations for FKRB (2011).

$$u_{ij} = x'_{ij}\beta^r + \epsilon_{ij} \tag{3.3}$$

$$g_j(x_i, \beta^r) = \frac{exp(x'_{ij}\beta^r)}{1 + \sum_{j'=1}^{J} exp(x'_{ij'}\beta^r)} \tag{3.4}$$

$$Pr(y_{ij} = 1|x_i) = \sum_{r=1}^{R} \theta^r g_j(x_i, \beta^r) \tag{3.5}$$

where ϵ_{ij} is Type I extreme value error, and x_{ij} is the K observed characteristics for each pair of agents $i = 1, ..., N$ and products $j = 1, ..., J$. $\theta = (\theta^1, ..., \theta^R)$ represents the probability or share of consumer types $r = 1, ..., R$ in the population. The primary interest is to estimate the tuple θ, and hence the distribution of random coefficients (CDFs) of β.

The actual estimation is simple OLS, with $i = 1, ..., N$ observations on (x_i, y_i) and the following moment condition.

$$E[y_{ij} - Pr(y_{ij} = 1|x_i)|x_i] = 0 \tag{3.6}$$

Letting the $R \times 1$ vector $z_{ij} = (z_{ij1}, ..., z_{ijR})'$ with individual elements $z_{ijr} = g_j(x_i, \beta^r)$, if one fixes or simulates the observation pairs (x_i, y_i), z_{ijr} is a fixed regressor. Equation (3.6) gives a consistent OLS estimator for θ.

$$\hat{\theta} = arg \min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} (y_{ij} - z'_{ij}\theta)^2 \tag{3.7}$$

Defining Y as the $NJ \times 1$ vector stacking y_{ij}'s and Z as the $NJ \times R$ matrix stacking z_{ij}, the estimator is $\hat{\theta} = (Z'Z)^{-1}Z'Y$. Solving equation (3.7) can be easily done as a constrained minimization using linlsq in Matlab. The two constraints for θ naturally required as a probability vector are $\sum_{r=1}^{R} \theta^r = 1$ and $\theta^r \geq 0$ for all $r = 1, ..., R$.

Once θ is estimated, one can construct the estimated CDFs for the random coefficients.

$$\hat{F}_N(\beta) = \sum_{r=1}^{R} \hat{\theta}^r 1[\beta^r \leq \beta] \tag{3.8}$$

where $1[\beta^r \leq \beta] = 1$ when $\beta^r \leq \beta$.

3.2.2 High-Dimensional Metrics

As an extension of FKRB (2011), the main interest is to examine the performance of the baseline estimator when there are too many consumer types r = 1, ..., R to consider. In other words, this paper shows the approximating performance of the estimator $\hat{F}(\beta)$ to $F(\beta)$ when the dimensionality of grid of points R is reduced by two lasso variants, namely the original plain lasso and Belloni and

Chernozhukov (2013)'s lasso with sparsity assumption (henceforth cv and hdm lasso, respectively), to the OLS minimization problem as specified in equation (3.7).

$$arg \min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} (y_{ij} - z'_{ij}\theta)^2 + \lambda||\theta||_1 \tag{3.9}$$

$$arg \min_{\theta} \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} (y_{ij} - z'_{ij}\theta)^2 + \frac{\lambda^*}{NJ}||\hat{\Psi}\theta'||_1 \tag{3.10}$$

Minimization problem (3.9) is the formulation for cv lasso with the shrinkage parameter λ and the absolute value norm $|| \cdot ||_1$. Minimization (3.10) is for hdm lasso with sparsity assumption, and the data driven penalty loadings $\hat{\Psi}$ and λ^* defined to guarantee the asymptotic approximation performance of post-selection OLS estimators.

$$\hat{\psi}_r = \sqrt{\frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} (z_{ijr}^2 \hat{\epsilon}_{ij}^2)} \tag{3.11}$$

$$\lambda^* = 2c\sqrt{NJ}\Phi^{-1}(1 - \gamma/(2R)) \tag{3.12}$$

where $\hat{\epsilon}_{ij}$ is the residuals, Φ is the CDFs for standard normal distribution, and c and γ are conditioning parameters preset at 1.1 and 0.1 for heteroskedastic error structure.

Hence, given the agent type probabilities $\theta = (\theta^1, \dots, \theta^R)$, a researcher first reduces the dimensionality of θ, for example, $\theta^* = (\theta^{1*}, \dots, \theta^{R^*})$ with $R^* \leq R$. In the Monte-Carlo experiment, this paper picks a fixed grid of points for θ of dimension R using Halton draws, and then use lasso variants to select a grid of points for θ^* with a smaller dimensionality R^*. Corresponding x_{ij}'s and the coefficients β^r's are generated from fixed distributions. With this reduced dimensionality of the new tuple θ^*, the baseline least squares (equation (3.7)) is estimated to construct the estimated CDFs for β.

This dimension reduction is in fact conducted on the regressor vector Z with rank R, with individual elements z_{ijr}, or the individual choice probabilities $g_j(x_i, \beta^r)$. The selected choice probabilities $g_j(x_i, \beta^r)$ form a new rank-reduced regressor vector $Z^* \in R^{NJ \times R^*}$.

3.3 Monte-Carlo

3.3.1 Parameter and Settings

For the Monte-Carlo experiment, this paper tries six combinations. For each $N = 2,000$ and $5,000$ observation pair set, three gaussian mixtures distributions for generating $\beta = (\beta_1, \beta_2)$ were used. There are $J = 10$ choice alternatives, $K = 2$ observed attributes, and $R = t^2$, $t = 3, 4, \dots, 22$ (9, 16, ... , 484) consumer types. The two dimensional grid of points for θ with dimensionality R are drawn from $[-10, 10] \times [-10, 10]$ using Halton draws. For estimated CDFs, the $S = 10,201$ (101×101) grid of points on which both the actual (simulated) and estimated CDFs will be evaluated are uniformly drawn from also $[-10, 10] \times [-10, 10]$. The number of Monte-Carlo repetition M is 100.

Two kinds of approximating performance metric were used, RMISE (Root Mean Integrated Squared Error) and IAE (Integrated Absolute Error).

$$RMISE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}[\frac{1}{S}\sum_{s=1}^{S}(\hat{F}_m(\beta_s) - F_0(\beta_s))^2]} \tag{3.13}$$

$$IAE = \frac{1}{S}\sum_{s=1}^{S}|\hat{F}_m(\beta_s) - F_0(\beta_s)| \tag{3.14}$$

β_s represents the two dimensional ($K = 2$) coefficients for observed attributes x_{ij}'s at one of the grid points $s = 1, \dots, S$. $\hat{F}_m$ is the estimated CDFs at the m-th repetition and F_0 is the 'true' CDFs for random coefficients β generated using $N = 10,000$.

Each $x_{ij} \in R^2$ is drawn from $N(0, 1.5^2)$, and the true $F_0(\beta)$ are drawn from three different mixtures normal distributions with $\Sigma_1 = \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.4 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}$. In other words, there are three designs for each N, namely gaussian mixtures distributions of two, four, and six normal distributions (Equation (3.15), (3.16), and (3.17)).

$$0.4 * N([3, -1],\ \varSigma_1) + 0.6 * N([-1,\ 1],\ \varSigma_2) \tag{3.15}$$

$$\begin{aligned} &0.2 * N([3,\ 0],\ \Sigma_1) + 0.4 * N([0,\ 3],\ \Sigma_1) \\ &+ 0.3 * N([1,\ -1],\ \Sigma_2) + 0.1 * N([-1,\ 1],\ \Sigma_2) \end{aligned} \tag{3.16}$$

$$\begin{aligned} &0.1 * N([3, 0],\ \Sigma_1) + 0.2 * N([0,\ 3],\ \Sigma_1) + 0.2 * N([1,\ -1],\ \Sigma_1) \\ &+ 0.1 * N([-1,\ 1],\ \Sigma_2) + 0.3 * N([2,\ 1],\ \Sigma_2) + 0.1 * N([1,\ 2],\ \Sigma_2) \end{aligned} \tag{3.17}$$

For lasso methods, the selection results (R^*) could differ over the random draws of x's and β's. Also, to compare the approximating performances between the baseline OLS and post-selection OLS, I fixed a grid of points R, and conducted lasso selection over 10 different random draws for x's and β's producing 10 different reduced grid of points of dimensionality R^*. For example, if a Halton draws of two dimensional grid of points $R = 256$ is at hand, hdm lasso selection method was applied 10 times to each set of x's and β's to produce 10 reduced grid of points R^* with the dimensionality varying from 20 to 22 for the design of $N = 5,000$ with the number of mixtures at six (Table 3.3).

The post-cv lasso uses 10 folds cross validation, and $g_j(x_i,\ \beta^r)$'s were selected using the λ values achieving the minimum RMSE (Root Mean Squared Error) from the penalized regressions. The post-hdm lasso was applied with the default setting as specified by the R package '**hdm**' for the heteroskedastic error case.

3.4 Results and Discussion

3.4.1 Performance Metrics

Table 3.1, 3.2, and 3.3 report the Monte-Carlo simulation results. Each table contains the performance metrics (RMISE and IAE) for the baseline and post-lasso estimators using 10 folds cross validation and Belloni and Chernozhukov (2013), for each combination of N (2,000 and 5,000), mixtures distributions (two, four, and six) and R from 16 to 484. For each R, the reduced dimensionalities R^* produced by both post-cv and post-hdm lasso are reported, along with the number of positive weights estimated. RMISE and IAE results for post-cv and post-hdm lasso estimators are average values computed over the 10 different reduced grid points.

The results shows the following: (1) For both the baseline and post-lasso inequality constrained OLS, RMISE and IAE decrease in N and R but only until R reaches a certain level (144 or 169). (2) Even R is relatively high, the number of non-zero basis functions (non-zero θ^r's) stays low, about up to 11 for the most complex case ($N = 5,000$ and $R = 484$ with mixtures of six normals). (3) RMISE and IAE are lower than the baseline for R values above certain level (≥ 49) with the reduced grid R^* using either post-cv or post-hdm lasso. (5) It is hard to compare post-cv and post-hdm lasso in terms of RMISE and IAE across all the combinations of N, R, and distribution mixtures. (6) Post-hdm lasso selects fewer variables than post-cv lasso with 10 folds cross validation. The mean and maximum number of dimensionality in the post-selection grid R^* are higher for post-cv lasso. So are the number of positive weights (θ^r's).

3.4.2 Marginal Distributions of Coefficients

Figure 3.1 through Figure 3.6 depict $F_0(\beta_1)$ and $\hat{F}(\beta_1)$, namely the (simulated) true marginal distribution of β_1 and the estimated marginal distributions using the baseline, post-cv lasso, and post-hdm lasso estimators from the Monte-Carlo designs of N = 5,000 over various R's. The marginal distributions were calculated from the estimated joint CDFs $\hat{F}(\beta_1, \beta_2)$.

Figure 3.1 and Figure 3.2 compare $\hat{F}(\beta_1)$'s produced by the baseline and post-lasso estimators with relatively low levels of R ranging from 16 to 49. For R values of 16 and 25, there seems to be no clear visual confirmation that the approximation performances of post-lasso estimators are better than the baseline. As R exceeds 36, post-lasso estimators start to show better fits, with post-cv lasso performing better at tail areas than post-hdm lasso.

Figure 3.3 and Figure 3.4 depict the analogous comparisons with an increase in R values of 81 to 144. The fit for $\hat{F}(\beta)$ of post-lasso estimators improves more clearly and stays consistent, as demonstrated by the RMISE and IAE values in Table 3.3. Post-cv lasso hits the best fit at $R = 121$, and over R values of 81 and 144, post-cv lasso tracks the (simulated) true $F_0(\beta_1)$ better than post-hdm lasso though the differences are small.

Figure 3.5 and Figure 3.6 show $F_0(\beta_1)$ and $\hat{F}(\beta_1)$ for relatively high R values of 169 and 529. Both post-cv and post-hdm lasso track the true CDFs of β_1 very well, while post-cv lasso still performs slightly better than post-hdm lasso. But the computation speed is much faster when using post-hdm lasso, because the reduced dimensionality R^* for cv lasso is much greater than hdm lasso.

Figure 3.7 shows the (simulated) true joint distribution of β_1 and β_2. One thing to note is that by coincidence, the mixtures distributions become smoother as the number of mixtures increase. Though the mixtures of two normals contain more inflection points, the fit of post-lasso estimators are still excellent. The results in Figure 3.1 through 3.7 are similar for β_2.

3.5 Conclusion

This paper explores the potential gains of high-dimensional metrics to the nonparametric least squares estimator for the distribution of random coefficients in multinomial logit demand case, developed by FKRB (2011). It is easy to program and highly flexible enough to be combined with auxiliary techniques, such as lasso and other machine learning methods for dimension reduction. Post-lasso regression results shows better approximating performances to the joint mixtures distributions and faster computation speed. Without resorting to the existing nonlinear estimation methods, our post-lasso estimator successfully captures heterogeneity in consumer preferences.

Table 3.1: Monte-Carlo Results (1) (Number of Mixtures: 2)

	RMISE			IAE			dim(R*)		# of Pos. Weights		
	base	cv	hdm	base	cv	hdm	cv	hdm	base	cv	hdm
R		mean			mean		mean (min, max)			mean	
N = 2,000											
16	0.0676	0.0688	0.0705	0.0283	0.0289	0.0289	7.9 (7, 9)	5.9 (5, 7)	6.36	6.12	5.35
25	0.0695	0.0714	0.0786	0.0290	0.0300	0.0328	11.5 (9, 15)	5.1 (4, 6)	6.67	6.42	4.94
36	0.0702	0.0681	0.0767	0.0302	0.0283	0.0329	19.2 (9, 25)	7.2 (5, 9)	7.51	6.48	5.13
49	0.0695	0.0666	0.0684	0.0297	0.0277	0.0278	19.0 (11, 29)	8.0 (7, 9)	7.43	6.95	5.14
64	0.0713	0.0669	0.0655	0.0306	0.0275	0.0269	23.5 (10, 49)	8.0 (7, 10)	8.29	6.56	5.57
81	0.0711	0.0673	0.0633	0.0306	0.0283	0.0260	24.1 (13, 36)	9.7 (6, 14)	8.36	7.38	5.57
100	0.0720	0.0674	0.0656	0.0311	0.0288	0.0273	18.5 (9, 33)	8.9 (6, 12)	8.00	7.54	6.01
121	0.0722	0.0653	0.0641	0.0312	0.0272	0.0264	23.9 (12, 63)	9.3 (8, 11)	8.21	6.90	5.65
144	0.0745	0.0692	0.0649	0.0322	0.0283	0.0265	19.6 (14, 36)	9.6 (7, 12)	8.84	6.89	5.80
169	0.0755	0.0658	0.0655	0.0329	0.0273	0.0269	22.3 (13, 42)	10.1 (8, 14)	8.29	7.19	5.74
196	0.0755	0.0697	0.0670	0.0331	0.0293	0.0273	18.3 (14, 22)	10.1 (7, 12)	8.35	7.37	5.57
225	0.0771	0.0672	0.0672	0.0337	0.0278	0.0275	29.7 (18, 66)	11.9 (10, 15)	8.35	7.10	5.58
256	0.0780	0.0706	0.0668	0.0341	0.0297	0.0274	30.7 (16, 74)	10.4 (7, 13)	8.74	7.94	6.29
289	0.0778	0.0698	0.0664	0.0339	0.0292	0.0271	34.8 (18, 78)	11.9 (7, 15)	9.09	7.14	5.38
324	0.0791	0.0683	0.0663	0.0343	0.0287	0.0268	27.0 (24, 30)	11.5 (8, 14)	8.42	7.54	5.30
361	0.0788	0.0706	0.0677	0.0341	0.0293	0.0277	30.1 (18, 43)	11.9 (10, 15)	8.48	7.52	5.64
400	0.0797	0.0677	0.0666	0.0348	0.0282	0.0273	26.9 (19, 34)	13.3 (10, 17)	8.84	7.16	5.49
441	0.0799	0.0688	0.0670	0.0349	0.0290	0.0275	32.5 (22, 52)	12.9 (10, 16)	8.70	7.32	5.61
484	0.0801	0.0696	0.0660	0.0350	0.0292	0.0269	27.9 (15, 37)	13.1 (9, 21)	8.64	7.55	5.40
N = 5,000											
16	0.0693	0.0347	0.0299	0.0288	0.0379	0.0398	11.2 (8, 13)	5.2 (5, 6)	6.24	7.36	5.00
25	0.0695	0.0609	0.0607	0.0284	0.0254	0.0248	23.4 (22, 24)	7.6 (6, 9)	6.72	9.44	7.03
36	0.0652	0.0602	0.0603	0.0271	0.0259	0.0251	25.8 (22, 33)	8.8 (8, 11)	7.40	10.59	7.63
49	0.0641	0.0508	0.0520	0.0262	0.0198	0.0200	27.4 (14, 35)	10.0 (9, 11)	7.62	10.80	8.32
64	0.0759	0.0493	0.0507	0.0320	0.0191	0.0193	32.2 (26, 36)	11.6 (10, 13)	8.17	11.59	8.93
81	0.0668	0.0392	0.0371	0.0284	0.0155	0.0145	26.0 (18, 32)	11.6 (9, 16)	8.61	12.61	9.06
100	0.0671	0.0366	0.0401	0.0284	0.0151	0.0156	30.6 (15, 71)	10.2 (9, 13)	8.74	12.33	8.22
121	0.0675	0.0325	0.0352	0.0284	0.0132	0.0140	25.4 (20, 43)	12.0 (9, 14)	8.72	11.96	8.17
144	0.0688	0.0330	0.0358	0.0290	0.0129	0.0140	20.8 (20, 21)	11.0 (11, 11)	8.94	12.23	7.51
169	0.0695	0.0324	0.0321	0.0294	0.0133	0.0121	27.6 (23, 32)	11.4 (11, 13)	9.05	12.11	9.10
196	0.0699	0.0340	0.0345	0.0295	0.0139	0.0128	25.0 (22, 36)	12.2 (12, 13)	9.25	12.93	9.10
225	0.0697	0.0353	0.0373	0.0291	0.0142	0.0159	36.2 (27, 40)	13.0 (13, 13)	9.06	14.89	7.61
256	0.0713	0.0326	0.0330	0.0296	0.0140	0.0128	37.6 (20, 47)	15.2 (14, 17)	9.13	14.90	10.26
289	0.0715	0.0314	0.0283	0.0297	0.0121	0.0106	26.4 (19, 36)	14.2 (12, 17)	9.06	12.60	10.21
324	0.0714	0.0326	0.0331	0.0297	0.0131	0.0127	31.6 (25, 42)	16.2 (12, 18)	9.10	13.60	9.18
361	0.0714	0.0321	0.0326	0.0297	0.0142	0.0125	35.8 (28, 42)	16.2 (14, 18)	9.24	13.82	9.56
400	0.0717	0.0292	0.0355	0.0298	0.0117	0.0137	24.2 (20, 31)	17.0 (14, 21)	9.17	11.96	9.63
441	0.0733	0.0309	0.0348	0.0307	0.0124	0.0130	49.8 (31, 104)	16.6 (14, 21)	9.54	14.54	9.40
484	0.0716	0.0332	0.0368	0.0298	0.0137	0.0142	46.6 (37, 65)	18.8 (16, 21)	9.10	15.34	10.56

Table 3.2: Monte-Carlo Results (2) (Number of Mixtures: 4)

		RMISE			**IAE**		**dim(R*)**			**# of Pos. Weights**	
	base	**cv**	**hdm**	**base**	**cv**	**hdm**	**cv**	**hdm**	**base**	**cv**	**hdm**
R		**mean**			**mean**		**mean (min, max)**			**mean**	
						N = 2,000					
16	0.0634	0.0627	0.0593	0.0268	0.0264	0.0245	11.8 (10, 16)	7.6 (7, 8)	6.61	6.28	5.52
25	0.0664	0.0682	0.0731	0.0279	0.0293	0.0311	14.0 (10, 19)	8.8 (7, 10)	7.17	6.59	5.13
36	0.0713	0.0655	0.0753	0.0309	0.0276	0.0328	16.6 (11, 30)	10.0 (8, 12)	8.01	6.87	5.53
49	0.0704	0.0640	0.0662	0.0303	0.0269	0.0268	18.8 (13, 29)	9.8 (7, 14)	8.18	7.01	5.19
64	0.0708	0.0642	0.0651	0.0308	0.0267	0.0270	17.4 (12, 30)	11.2 (9, 14)	8.54	6.84	5.85
81	0.0695	0.0631	0.0608	0.0303	0.0270	0.0252	23.8 (15, 38)	12.2 (11, 13)	8.61	7.66	5.85
100	0.0706	0.0660	0.0657	0.0308	0.0294	0.0277	30.2 (14, 74)	12.6 (10, 15)	8.75	7.59	6.07
121	0.0712	0.0619	0.0617	0.0311	0.0261	0.0255	21.0 (15, 32)	13.8 (11, 17)	8.68	7.29	6.03
144	0.0733	0.0646	0.0617	0.0321	0.0269	0.0252	23.6 (21, 29)	13.2 (10, 17)	8.96	7.35	6.26
169	0.0740	0.0618	0.0624	0.0325	0.0261	0.0257	23.4 (20, 27)	12.8 (10, 18)	8.66	7.32	5.87
196	0.0730	0.0662	0.0642	0.0322	0.0283	0.0263	23.0 (18, 30)	12.8 (9, 16)	8.97	7.58	5.78
225	0.0756	0.0620	0.0636	0.0331	0.0262	0.0265	28.6 (26, 34)	14.4 (12, 16)	8.97	7.45	5.94
256	0.0761	0.0678	0.0630	0.0334	0.0293	0.0260	31.8 (27, 40)	16.0 (13, 19)	9.11	7.96	6.30
289	0.0765	0.0648	0.0623	0.0336	0.0273	0.0255	31.6 (23, 43)	13.4 (7, 17)	9.21	7.46	5.70
324	0.0770	0.0647	0.0638	0.0338	0.0277	0.0257	34.2 (21, 65)	13.4 (10, 16)	9.29	7.78	5.54
361	0.0764	0.0660	0.0628	0.0332	0.0277	0.0254	29.6 (23, 36)	16.2 (11, 19)	8.82	7.62	5.73
400	0.0781	0.0644	0.0615	0.0345	0.0271	0.0254	38.0 (32, 45)	16.4 (10, 20)	8.90	7.58	5.92
441	0.0783	0.0641	0.0631	0.0345	0.0273	0.0261	41.4 (22, 76)	15.0 (11, 19)	8.94	7.75	5.84
484	0.0786	0.0648	0.0617	0.0347	0.0277	0.0252	39.2 (27, 48)	15.6 (12, 19)	9.00	7.64	5.49
						N = 5,000					
16	0.0575	0.0549	0.0569	0.0242	0.0248	0.0259	14.4 (12, 16)	8.0 (7, 9)	7.00	8.98	7.55
25	0.0610	0.0514	0.0550	0.0254	0.0228	0.0255	16.6 (15, 22)	12.0 (11, 13)	7.52	10.70	9.46
36	0.0605	0.0516	0.0617	0.0260	0.0228	0.0282	20.0 (13, 27)	11.4 (10, 13)	8.52	11.01	8.30
49	0.0580	0.0406	0.0460	0.0248	0.0168	0.0210	21.8 (15, 34)	13.2 (12, 14)	8.83	12.50	9.93
64	0.0618	0.0392	0.0447	0.0266	0.0165	0.0191	22.2 (17, 31)	15.0 (13, 17)	9.40	14.29	9.98
81	0.0610	0.0323	0.0401	0.0262	0.0144	0.0172	22.2 (18, 26)	15.2 (14, 16)	9.48	15.16	10.92
100	0.0621	0.0355	0.0439	0.0271	0.0162	0.0194	27.0 (22, 34)	14.0 (12, 16)	9.69	16.19	10.44
121	0.0622	0.0341	0.0416	0.0271	0.0149	0.0181	25.6 (23, 31)	17.2 (15, 19)	9.56	16.07	11.16
144	0.0643	0.0356	0.0414	0.0280	0.0158	0.0182	36.4 (27, 62)	16.4 (15, 18)	9.87	17.50	11.42
169	0.0650	0.0404	0.0424	0.0283	0.0173	0.0179	29.8 (24, 42)	16.6 (14, 19)	9.74	16.04	11.39
196	0.0679	0.0368	0.0379	0.0295	0.0159	0.0159	33.4 (29, 43)	16.2 (14, 18)	9.98	16.38	11.48
225	0.0672	0.0345	0.0405	0.0292	0.0147	0.0175	34.4 (29, 43)	17.2 (13, 25)	10.10	16.75	10.86
256	0.0682	0.0365	0.0407	0.0295	0.0158	0.0178	41.0 (25, 74)	20.4 (14, 29)	9.83	17.32	11.65
289	0.0685	0.0340	0.0409	0.0295	0.0150	0.0177	39.4 (34, 47)	20.2 (16, 25)	9.97	17.56	10.52
324	0.0683	0.0354	0.0423	0.0294	0.0155	0.0187	50.8 (29, 77)	17.6 (17, 18)	9.90	17.69	10.24
361	0.0685	0.0347	0.0423	0.0295	0.0152	0.0188	45.0 (14, 80)	23.0 (17, 30)	10.14	17.86	10.91
400	0.0689	0.0361	0.0432	0.0298	0.0159	0.0191	45.6 (35, 79)	19.2 (14, 22)	10.05	17.85	10.74
441	0.0683	0.0352	0.0418	0.0295	0.0148	0.0183	43.8 (37, 61)	19.6 (14, 26)	10.07	16.41	10.09
484	0.0700	0.0373	0.0433	0.0303	0.0163	0.0197	74.8 (39, 111)	20.8 (17, 24)	10.22	17.61	10.56

Table 3.3: Monte-Carlo Results (3) (Number of Mixtures: 6)

	RMISE			**IAE**			**dim(R*)**		**# of Pos. Weights**		
	base	**cv**	**hdm**	**base**	**cv**	**hdm**	**cv**	**hdm**	**base**	**cv**	**hdm**
R		**mean**			**mean**		**mean (min, max)**			**mean**	
						N = 2,000					
16	0.0651	0.0684	0.0746	0.0267	0.0288	0.0299	11.0 (9, 13)	7.4 (7, 8)	6.21	5.6	4.16
25	0.0657	0.0675	0.0763	0.0269	0.0281	0.0305	12.8 (9, 16)	8.0 (6, 10)	6.58	6.12	4.33
36	0.0620	0.0640	0.0705	0.0262	0.0264	0.0290	15.0 (11, 17)	9.8 (8, 11)	7.57	6.71	5.45
49	0.0616	0.0627	0.0637	0.0258	0.0261	0.0254	19.6 (13, 37)	12.2 (10, 14)	7.92	7.01	5.88
64	0.0646	0.0632	0.0593	0.0270	0.0259	0.0235	18.0 (16, 20)	12.0 (8, 14)	7.88	7.06	5.96
81	0.0655	0.0651	0.0594	0.0275	0.0270	0.0237	27.8 (20, 41)	13.0 (11, 15)	7.99	7.68	6.10
100	0.0662	0.0622	0.0590	0.0279	0.0258	0.0236	27.2 (23, 31)	12.8 (11, 14)	8.01	7.37	6.12
121	0.0660	0.0624	0.0601	0.0278	0.0259	0.0240	27.6 (19, 47)	12.8 (10, 15)	8.10	7.16	5.63
144	0.0683	0.0670	0.0616	0.0287	0.0273	0.0246	28.8 (19, 45)	13.4 (12, 16)	8.2	7.18	5.86
169	0.0694	0.0635	0.0621	0.0297	0.0261	0.0248	32.2 (21, 48)	15.2 (11, 19)	8.59	7.43	5.97
196	0.0705	0.0665	0.0632	0.0303	0.0278	0.0250	35.0 (28, 49)	14.0 (12, 15)	8.66	7.62	5.90
225	0.0709	0.0659	0.0641	0.0305	0.0270	0.0252	28.8 (23, 41)	16.6 (11, 23)	8.65	7.25	6.08
256	0.0721	0.0666	0.0639	0.0310	0.0275	0.0252	35.2 (28, 46)	18.4 (16, 22)	8.74	7.46	6.32
289	0.0713	0.0680	0.0639	0.0304	0.0285	0.0252	38.2 (22, 58)	15.0 (11, 19)	8.77	7.66	5.89
324	0.0732	0.0652	0.0622	0.0311	0.0271	0.0247	33.0 (25, 45)	17.2 (16, 19)	8.44	7.55	6.03
361	0.0734	0.0683	0.0659	0.0313	0.0283	0.0266	39.2 (30, 55)	17.6 (13, 21)	8.60	7.47	6.07
400	0.0734	0.0644	0.0651	0.0313	0.0268	0.0259	34.6 (28, 41)	16.0 (10, 20)	8.54	7.31	5.78
441	0.0736	0.0668	0.0642	0.0314	0.0278	0.0255	41.8 (25, 82)	18.0 (13, 22)	8.62	7.71	6.04
484	0.0735	0.0676	0.0638	0.0314	0.0281	0.0254	36.6 (31, 39)	18.6 (12, 22)	8.86	7.46	5.93
						N = 5,000					
16	0.0647	0.0568	0.0687	0.0261	0.0238	0.0295	13.0 (11, 16)	7.4 (7, 8)	6.14	9.27	6.72
25	0.0644	0.0569	0.0684	0.0258	0.0241	0.0298	18.2 (12, 22)	9.2 (9, 10)	6.51	10.51	7.66
36	0.0561	0.0387	0.0424	0.0232	0.0176	0.0205	26.2 (26, 27)	12.8 (12, 13)	8.66	12.47	10.09
49	0.0568	0.0330	0.0354	0.0234	0.0141	0.0152	24.8 (18, 33)	13.2 (12, 16)	8.66	13.44	10.48
64	0.0581	0.0371	0.0341	0.0242	0.0159	0.0148	30.0 (17, 39)	17.6 (14, 19)	9.14	16.01	12.67
81	0.0600	0.0297	0.0329	0.0253	0.0122	0.0141	25.2 (18, 29)	14.8 (10, 19)	9.13	15.86	10.39
100	0.0607	0.0306	0.0357	0.0252	0.0129	0.0160	29.8 (23, 34)	14.6 (14, 15)	9.45	17.07	10.25
121	0.0602	0.0286	0.0344	0.0250	0.0117	0.0141	28.8 (23, 34)	15.4 (15, 17)	9.49	16.18	9.95
144	0.0621	0.0347	0.0335	0.0259	0.0150	0.0137	37.0 (27, 52)	16.6 (15, 20)	9.61	18.10	11.56
169	0.0626	0.0287	0.0333	0.0262	0.0117	0.0138	32.4 (26, 37)	17.0 (15, 18)	9.60	17.01	11.47
196	0.0655	0.0309	0.0345	0.0278	0.0132	0.0151	32.8 (31, 34)	17.4 (16, 21)	9.47	16.70	11.50
225	0.0646	0.0338	0.0383	0.0271	0.0148	0.0160	63.8 (35, 95)	17.8 (16, 19)	9.69	20.65	10.26
256	0.0669	0.0319	0.0386	0.0284	0.0137	0.0163	38.8 (35, 45)	20.8 (20, 22)	9.74	17.60	11.45
289	0.0665	0.0346	0.0377	0.0279	0.0148	0.0163	39.4 (37, 47)	20.2 (18, 25)	9.82	16.64	11.50
324	0.0681	0.0348	0.0373	0.0287	0.0146	0.0156	40.8 (33, 62)	20.8 (17, 24)	9.58	17.32	11.9
361	0.0681	0.0330	0.0381	0.0287	0.0141	0.0154	46.2 (42, 52)	25.8 (21, 28)	9.58	16.64	11.57
400	0.0686	0.0338	0.0378	0.0287	0.0142	0.0162	33.8 (31, 40)	20.2 (17, 21)	9.94	15.52	11.37
441	0.0684	0.0333	0.0344	0.0289	0.0140	0.0148	57.8 (41, 97)	20.8 (19, 24)	9.64	17.91	12.17
484	0.0686	0.0346	0.0364	0.0290	0.0148	0.0147	86.4 (42, 160)	23.8 (20, 26)	9.72	19.10	11.37

Figure 3.1: $\hat{F}(\beta_1)$: Base vs. Post-cv Lasso (N=5,000, Mix 6, R = 16, ..., 49)

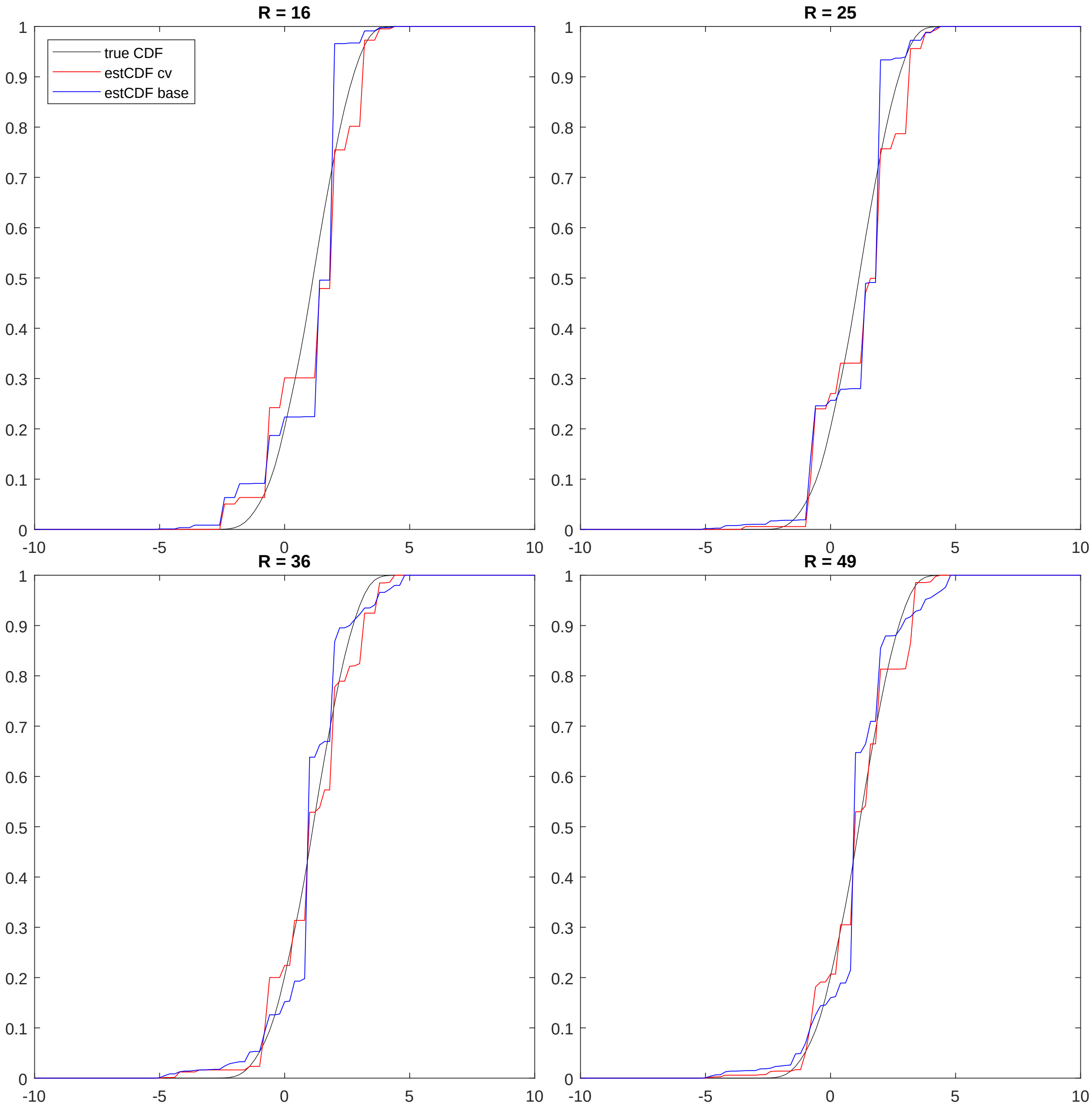


Figure 3.2: $\hat{F}(\beta_1)$: Base vs. Post-hdm Lasso (N=5,000, Mix 6, R = 16, ..., 49)

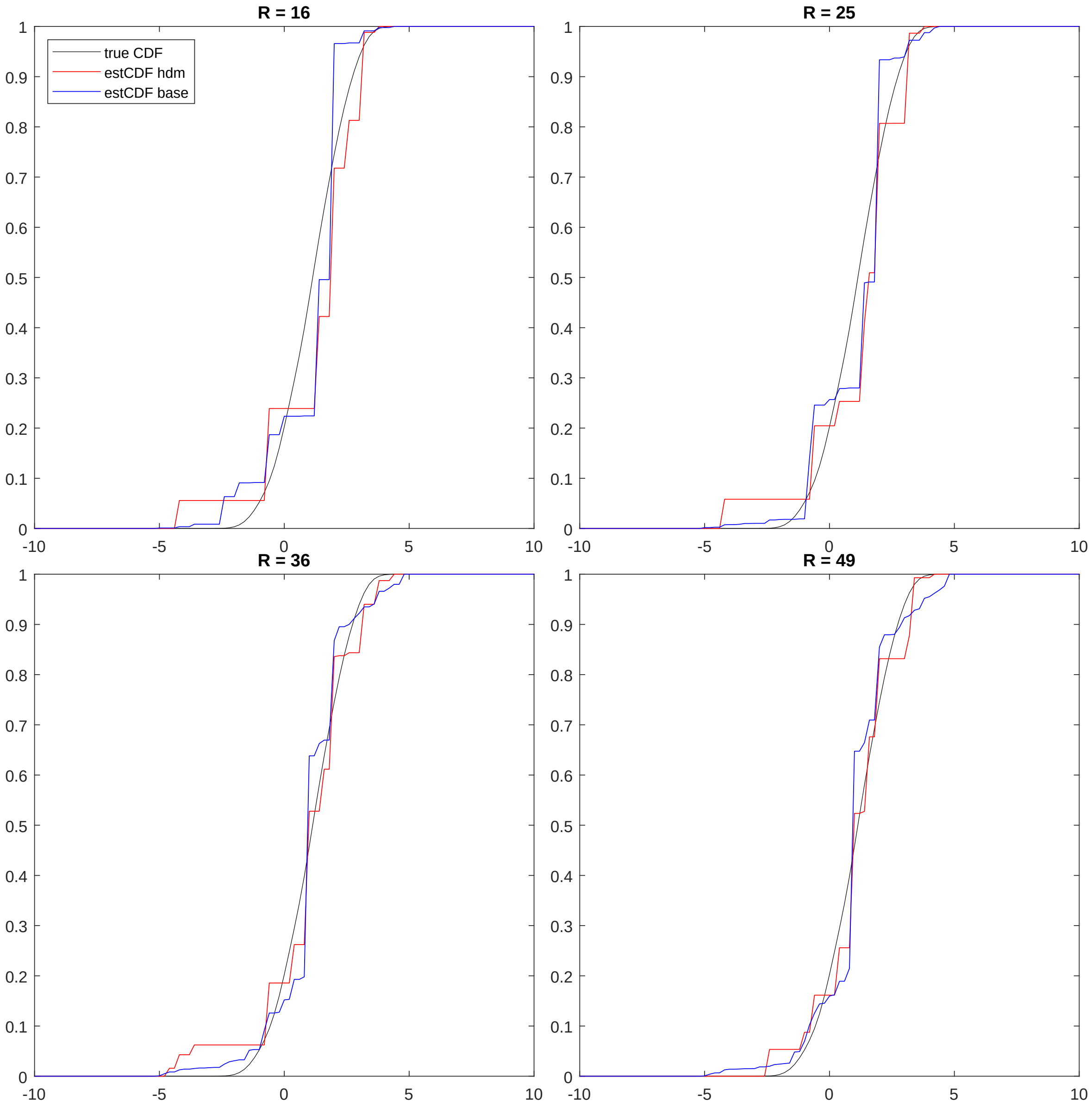


Figure 3.3: $\hat{F}(\beta_1)$: Base vs. Post-cv Lasso (N=5,000, Mix 6, R = 81, ..., 144)

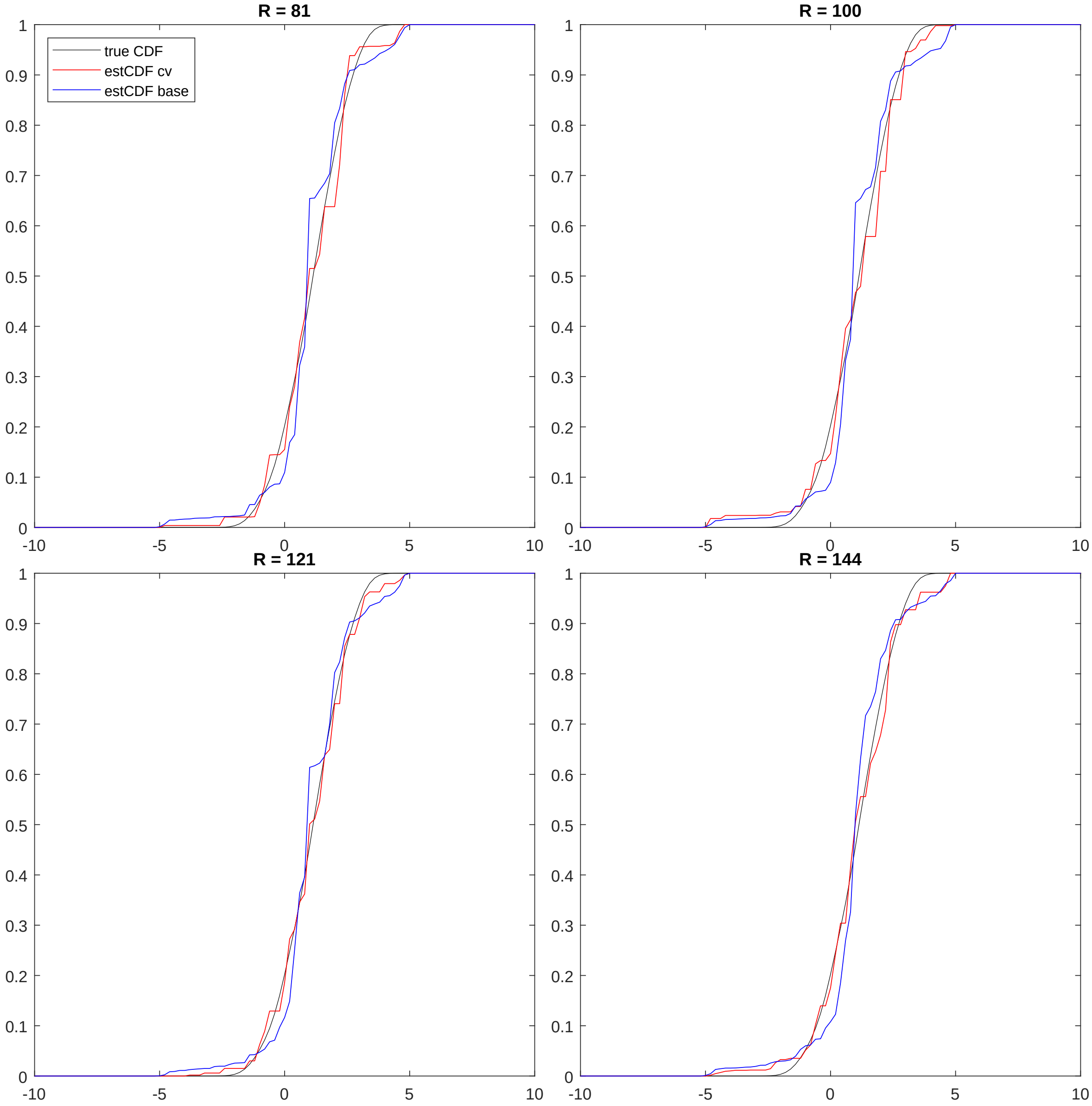


Figure 3.4: $\hat{F}(\beta_1)$: Base vs. Post-hdm Lasso (N=5,000, Mix 6, R = 81, ..., 144)

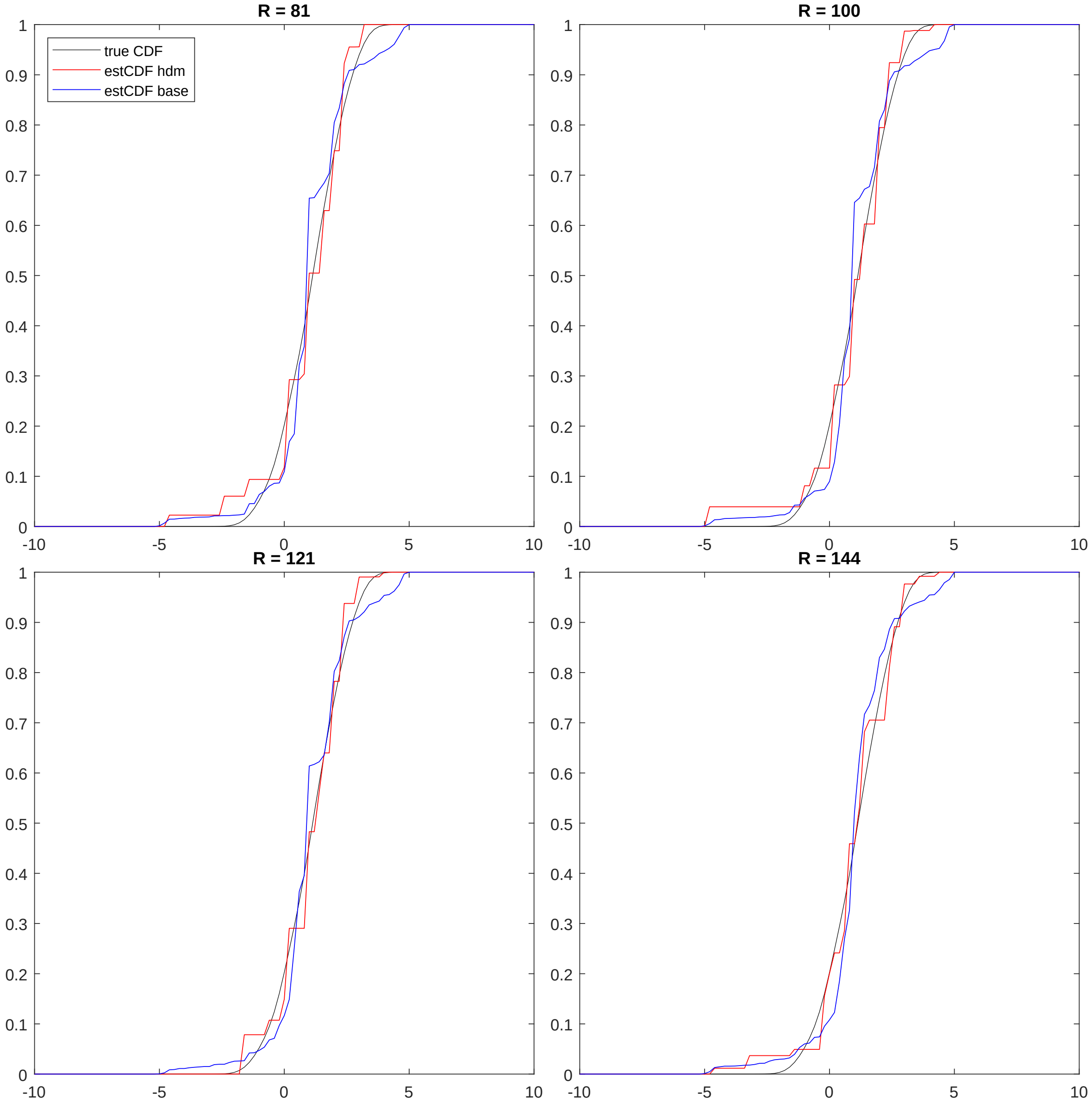


Figure 3.5: $\hat{F}(\beta_1)$: Post-cv vs. Post-hdm Lasso (N=5,000, Mix 6, R = 169)

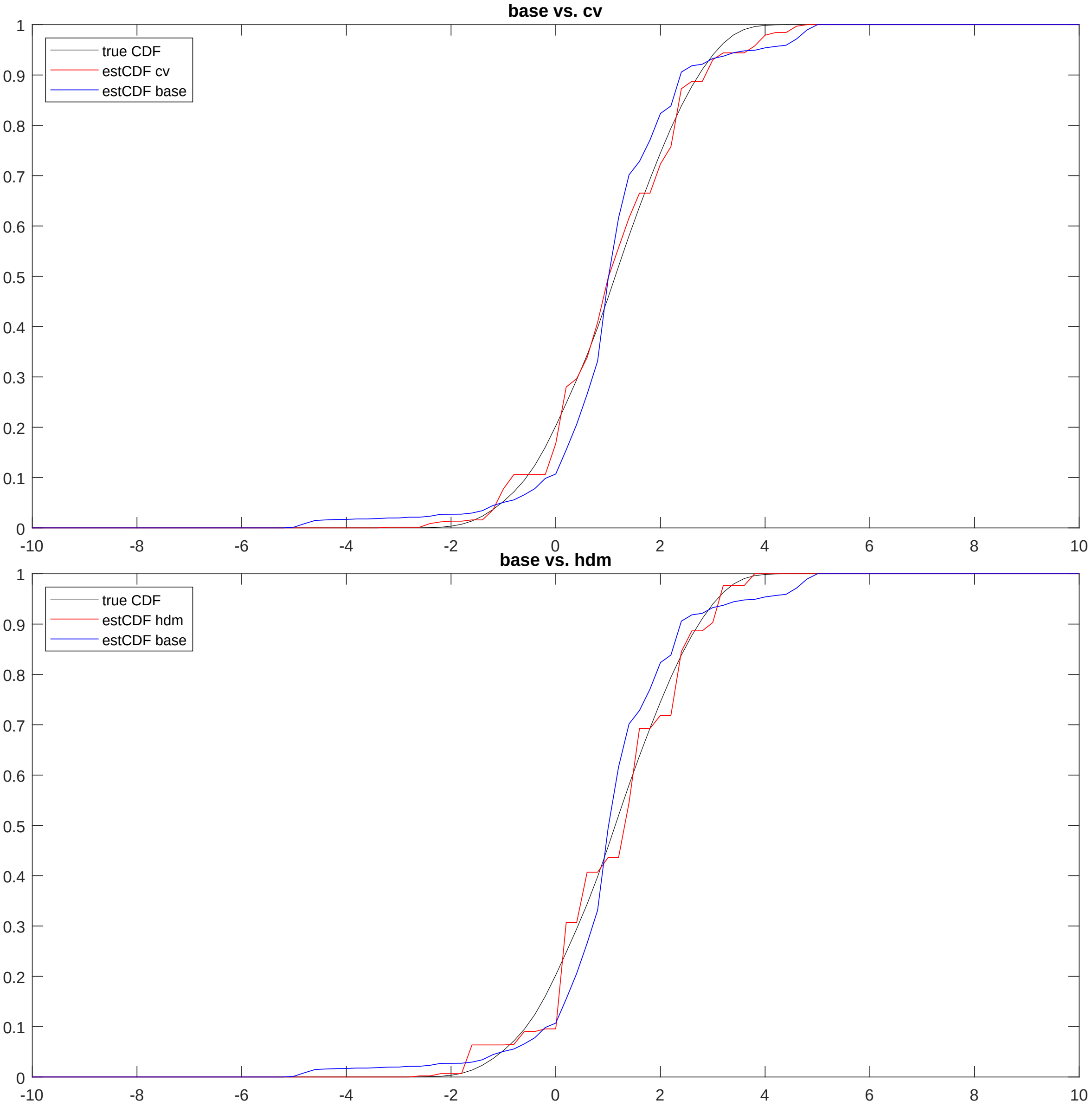


Figure 3.6: $\hat{F}(\beta_1)$: Post-cv vs. Post-hdm Lasso (N=5,000, Mix 6, R = 529)

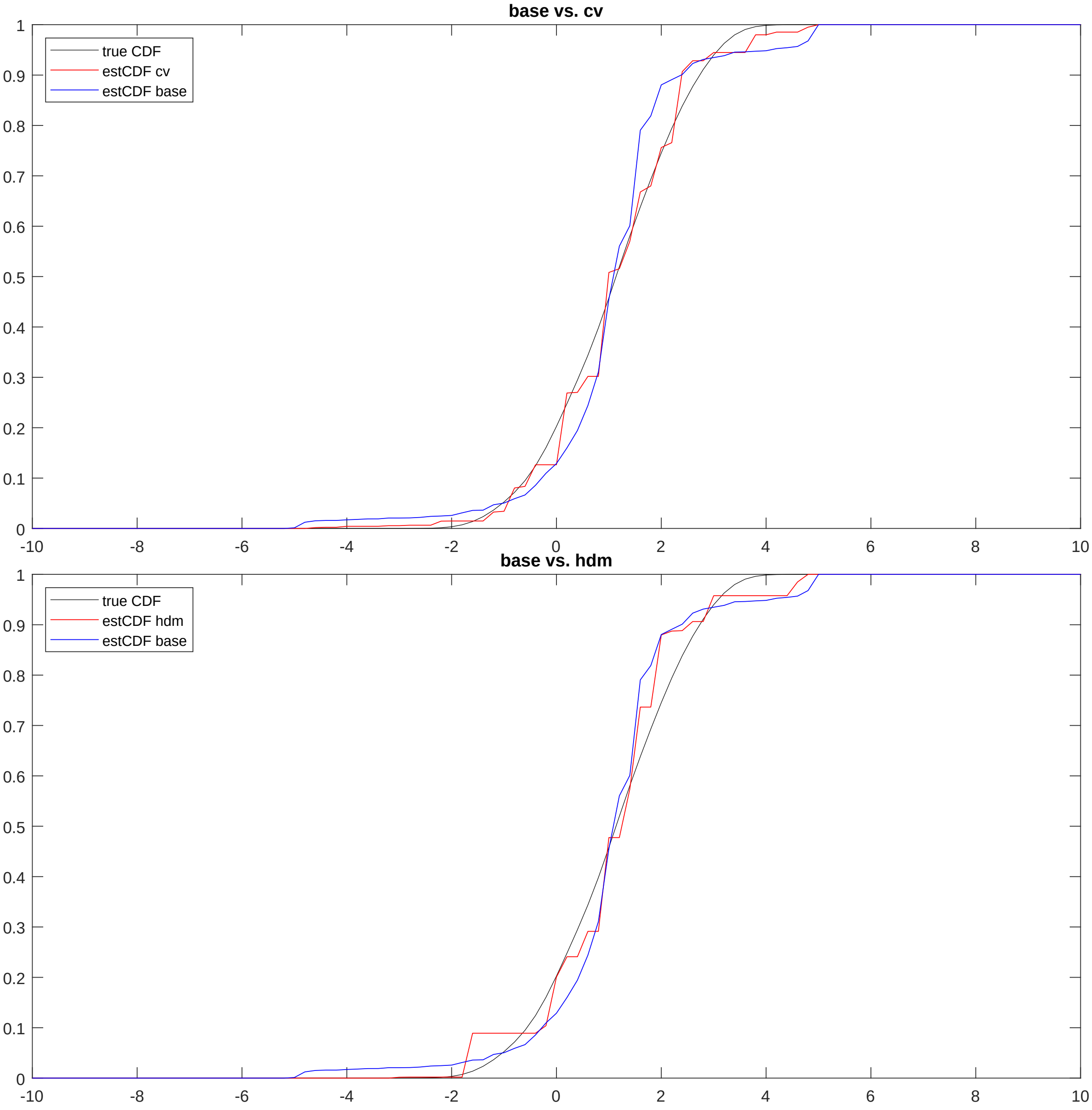


Figure 3.7: True Joint Distributions of β_1 and β_2 (1) N=5,000, Mixture of Two Normals

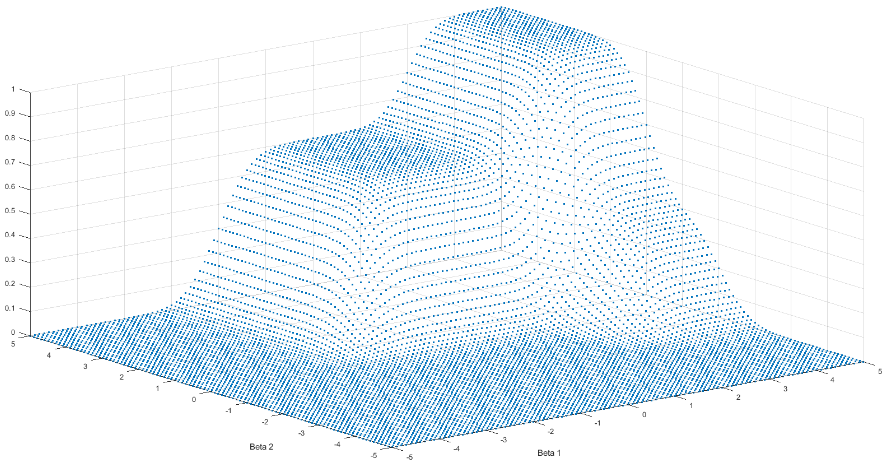


Figure 3.8: True Joint Distributions of β_1 and β_2 (2) N=5,000, Mixture of Four Normals

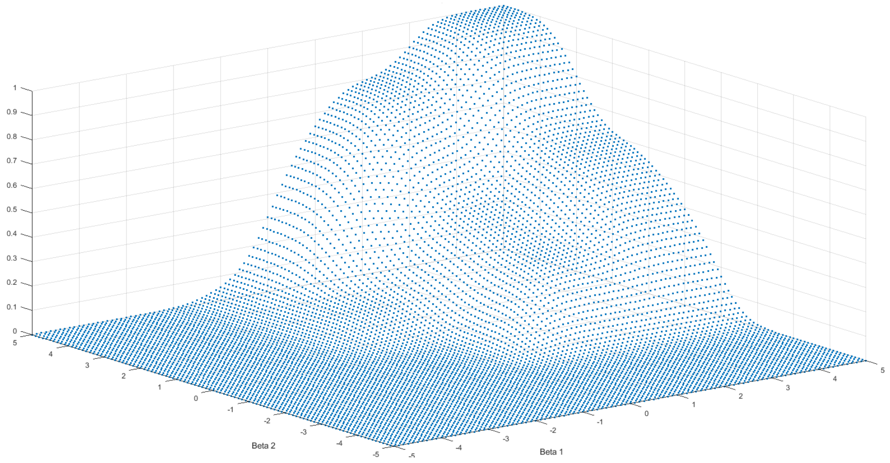
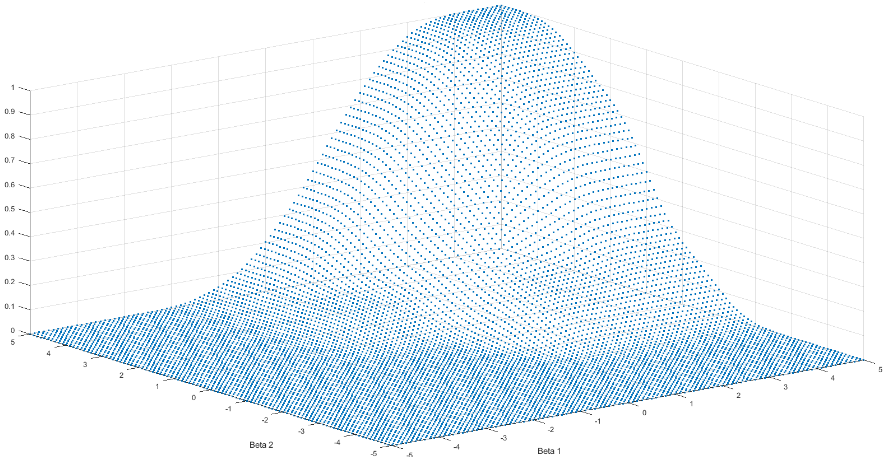


Figure 3.9: True Joint Distributions of β_1 and β_2 (3) N=5,000, Mixture of Six Normals



APPENDICES

APPENDIX A

MACHINE LEARNING AND POST SELECTION INFERENCE

A.1 Penalized Regression: Lasso, Ridge, and Elastic Net

Shrinkage methods or regularized regressions set an additional constraint on magnitudes parameter estimates can take. If there is a situation where regression coefficients can 'explode' due to multicollinearity one could employ one of shrinkage methods. Also, if there are irrelevant variables it could filter out such variables by increasing the shrinkage parameter.

A basic LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani(1996)) formulation could be stated as the following, where $PRSS(\beta_{l1})$ represents penalized residual sum of squares, where the shrinkage penalty on coefficient values (β) is given by $L1$ metric;

$$\begin{aligned} min_{\beta} PRSS(\beta_{l1}) &= \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (A.1) \\ &= (Y - X\beta)'(Y - X\beta) + \lambda ||\beta||_1 \end{aligned}$$

Ridge regression (Hoerl and Kennard (1970)) is a similar minimization but with $L2$ metric;

$$\begin{aligned} min_{\beta} PRSS(\beta_{l2}) &= \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (A.2) \\ &= (Y - X\beta)'(Y - X\beta) + \lambda ||\beta||_2 \end{aligned}$$

$$\begin{aligned} \frac{\partial PRSS(\beta_{l2})}{\partial \beta} &= -2X'(Y - X\beta) + 2\lambda\beta \qquad (A.3) \\ \hat{\beta}_{Ridge} &= (X'X + \lambda I_p)^{-1} X'Y \end{aligned}$$

λ is the tuning parameter that determines the degree of shrinkage for both LASSO and ridge regression problems. As λ approaches zero, the estimation gets closer to OLS (Ordinary Least Squares), and as λ approaches to infinity the model becomes an intercept only specification. Compared to ridge regression, LASSO tends to eliminate too many coefficients and ridge tends to leave too many variables.

Elastic net (Zou and Hastie (2005)) is a convex combination of LASSO and ridge that tries to harmonize the two methods;

$$min_{\beta} PRSS(\beta_{ElasticNet}) = (Y - X\beta)'(Y - X\beta) + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2 \tag{A.4}$$

Letting $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$ and t as an arbitrary positive real number, the solution to the elastic net regression is the following.

$$\hat{\beta}_{ElasticNet} = argmin_{\beta}(Y - X\beta)'(Y - X\beta) \tag{A.5}$$

$$s.t. (1 - \alpha)||\beta||_1 + \alpha||\beta||_2 \leq t \tag{A.6}$$

glmnet package in R implements LASSO, ridge regression, and elastic net with n-folds cross validation and RMSE (Root Mean Squared Error) criterion.

A.2 Gradient Boosting: Regression Tree Based Prediction

Given a response variable Y and predictors $X = (x_1, x_{2,}..., x_p)$ the decision tree picks a variable, pinpoints a splitting value on the selected variable and splits the predictor space X recursively. Each node contains a subset of observations for predictors and the response variable. Average value of the response in each final nodes is a tree model's prediction on Y. Splitting process stops when a loss function reaches a preset threshold. To improve prediction accuracy, a pruning process is commonly applied after fitting the tree model, $F(X)$.

Loss function choice for categorical response with J classes is Gini impurity measure $I_G(p)$, where p_j is the probability of predicting class j correctly and $1 - p_j$ is the probability of predicting class j with a wrong class at each node. In the categorical case, the decision tree is called classification tree I used for the review sentiment classification.

$$I_G(p) = \sum_{j=1}^{J} p_j(1 - p_j) = 1 - \sum_{j=1}^{J} p_j^2 \tag{A.7}$$

If the response is a continuous numeric variable, now it is a regression tree. One common choice of loss function for a regression tree is RMSE (Root Mean Square Error).

$$RMSE(Y, F(X)) = \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - F(x_i))^2} \tag{A.8}$$

At each node, the split variable and value are determined to minimize the resulting RMSE.

However, regression tree has its own weaknesses. Though a high level of prediction accuracy could be achieved, the resulting tree structure could be too complicated ('Bushy'). Then interpretation of the fitted model becomes nearly impossible and the model yields poor out of sample prediction performances. Also, if there is one variable that has a particularly strong correlation with the response, the splitting process is concentrated on the variable leading to biased estimates.

To deal with the weaknesses of plain regression tree, practitioners use model averaging techniques. First averaging method is 'Bagging' (Breiman (1996)). Bagging fits many trees on bootstrapped subsets of training data, and predicts outcome by majority vote from the estimated tree models. Second is 'Random Forests', a refined bagging approach (Breiman (2001)). Random forests method uses the same bootstrapped samples, but for each tree, a random sample of $m(< p)$ predictors is drawn and only those m features are used in the fitting processes. It tries to improve on bagging by de-correlating each trees.

This paper uses the third averaging technique, namely 'Gradient Boosting' (Friedman (2001)). Basic formulation can be stated as;

$$\widehat{Y} = F(X) + \sum_{l=1}^{L} \rho h_l(X) \tag{A.9}$$

$F(X)$ can be an initial fitted tree with predictors X. $h_l(X)$ is called a 'Weak Learner', another tree trained on the residuals from $F(X) + \sum_{l=1}^{l-1} h_l(X)$. Specifically, for the initial model $F(X)$, the residuals is $Y - F(X)$. Then $h_1(X)$ sets the residuals as a new response variable and trains another tree. The shrinkage parameter ρ is set low enough so that there would not be an overfitting problem. The reason why this approach is called as gradient boosting is from the fact that it uses residuals. Specifically, if we set an RMSE loss function, our optimization problem will be

$$min_{F(x_i)} J = \sum_{i=1}^{n} (y_i - F(x_i))^2 \tag{A.10}$$

Treating $F(x_i)$ as parameters, the derivative is,

$$\frac{\partial J}{\partial F(x_i)} = \frac{\partial \sum_i (y_i - F(x_i))^2}{\partial F(x_i)} = \frac{\partial (y_i - F(x_i))^2}{\partial F(x_i)} = 2(F(x_i) - y_i) \tag{A.11}$$

Gradient descent optimization minimizes a function by moving the function in the opposite direction of the gradient, in this case $-\frac{\partial J}{\partial F(x_i)} = y_i - F(x_i)$.

A practitioner can implement gradient boosting fitting procedure using the **gbm** package in R, also using n-folds cross validation and RMSE criterion function.

A.3 Exact Inference for OLS Estimates after Lasso Selection

The post selection inference approach used in this paper is directly from Lee, L. Sun, Sun, and Taylor (2016), implemented by R package '**selectiveInference**'. Exact inference for regression models after statistical/machine learning is an actively developing area and interested readers could benefit a lot from the recent literature written by the pioneers of machine learning in statistics (Lockhart, Taylor, J. Tibshirani, and Tibshirani (2014), J. Tibshirani, Taylor, Lockhart, and Tibshirani (2016), and Taylor and Tibshirani (2017)).

This appendix introduces a brief outline and compare two confidence intervals (C.Is) obtained from the classical OLS method and post selection inference of Lee et al. (2016). For a typical OLS regression, the objective y follows a multivariate normal distribution.

$$y \sim N(\mu, \sigma^2 I_n) \tag{A.12}$$

where μ is the mean vector modeled as a linear combination of p predictors $x_1, ..., x_p \in R^n$ and σ is the standard error. The primary goal is to get an exact distribution of coefficients obtained from OLS conducted only with the selected variables by LASSO or model M.

$$\beta^M = argmin_{b^M} E||y - X_M b^M||^2 = X_M^+ \mu = (X_M^T X_M)^{-1} X_M^T \mu \tag{A.13}$$

The LASSO selection event in fact implies one get the variables with non-zero coefficients and corresponding signs. The event of selecting the observed model $\hat{M}$ and signs $\hat{s}$ i.e., $\{\hat{M} = M,\ \hat{s} = s\}$ can be described by a polyhedron in the form of $\{Ay \leq b\}$.

$$\{\hat{M} = M,\ \hat{s} = s\} = \{A(M,s)y \leq b(M,s)\} \tag{A.14}$$

$$A(M,s) = \begin{pmatrix} A_0(M,s) \\ A_1(M,s) \end{pmatrix} = \begin{pmatrix} A_0(M,s) \\ -diag(s)(X_M^T X_M)^{-1} X_M^T \end{pmatrix}$$

$$b(M,s) = \begin{pmatrix} b_0(M,s) \\ b_1(M,s) \end{pmatrix} = \begin{pmatrix} b_0(M,s) \\ -\lambda diag(s)(X_M^T X_M)^{-1} s \end{pmatrix}$$

$$A_0(M,s) = \frac{1}{\lambda} \begin{pmatrix} X_{-M}^T (I - P_M) \\ -X_{-M}^T (I - P_M) \end{pmatrix}$$

$$b_0(M,s) = \begin{pmatrix} 1 - X_{-M}^T (X_M^T)^+ s \\ 1 + X_{-M}^T (X_M^T)^+ s \end{pmatrix}$$

where the subscript $-M$ represents variables of zero-coefficients in the LASSO selector, λ is the penalty parameter, P_M is the projection matrix toward the vector space of the selected variables, and $diag(s)$ is a diagonal matrix with the elements of s.

The next step is to get an exact distribution of individual coefficients β_j^M, conditional on the model selection event $\{Ay \leq b\}$. First the authors establish the conditional distribution of a generic linear transformation of the objective y: $\eta^T y | \{Ay \leq b\}$. With the choice of $\eta = (X_M^+)^T e_j$, one gets the conditional distribution of $\beta_j^M = e_j^T X_M^+ \mu = \eta^T \mu$. The selection event $\{Ay \leq b\}$ can once again be transformed into an interval of residuals from projecting y onto the direction of η.

$$\{Ay \leq b\} = \{\nu^-(z) \leq \eta^T y \leq \nu^+(z),\ \nu^0(z) \geq 0\} \tag{A.15}$$

$$\nu^-(z) = \max_{j:(Ac)_j < 0} \frac{b_j - (Az)_j}{(Ac)_j}$$

$$\nu^+(z) = \min_{j:(Ac)_j > 0} \frac{b_j - (Az)_j}{(Ac)_j}$$

$$\nu^0(z) = \min_{j:(Ac_j) = 0} b_j - (Az)_j$$

where A and b are as defined in the previous page, $z = (I_n - P_\eta)y$ is the residual with the projection matrix P_n onto the direction of η, and $c = \eta(\eta^T\eta)^{-1}$. Notice that z being residual, is independent of $\eta^T y$ and hence the LASSO selection event does not incur any complication to produce the conditional distribution but just imposes upper and lower limits on $\eta^T y$.

Hence, the distribution of $\eta^T y$ conditional on the model selection is a truncated normal.

$$[\eta^T y | Ay \leq b,\, z = z_0] \sim TN(\eta^T\mu,\, \sigma^2||\eta||^2,\, \nu^-(z_0),\, \nu^+(z_0)) \tag{A.16}$$

where z_0 is a realization of residual, and equation (A.16) is true for any z_0 because of the independence. The cumulative density $F^{[\nu_s^-(z),\, \nu_s^+(z)]}_{\eta^T\mu,\eta^T\eta}$ is monotone decreasing in $\eta^T\mu$ or in our specific interst, in β_j^M which gives the confidence interval $[L,\, U]$ with L and U are defined as $F^{[\nu_s^-(z),\, \nu_s^+(z)]}_{L,\sigma^2||\eta||}(\beta_j^M) = 1 - \frac{\alpha}{2}$ and $F^{[\nu_s^-(z),\, \nu_s^+(z)]}_{U,\sigma^2||\eta||}(\beta_j^M) = \frac{\alpha}{2}$ to achieve a significance level α. Thus,

$$P[\beta_j^M \in [L,\, U] | \hat{M} = M,\, \hat{s} = s] = 1 - \alpha \tag{A.17}$$

APPENDIX B

OMITTED DETAILS FOR CHAPTER 1

B.1 Exact Inference for Post Lasso Estimates

Table B.1 reports the C.Is obtained by both OLS and OLS post LASSO. For almost all variables, the C.Is from truncated normal essentially reproduce those of OLS but they are slightly wider reflecting the changes in density due to truncation. There are two exceptions to this.

First is when the signal of a variable is weak. Then parameter estimates could be near to one of the truncation endpoints, giving much wider intervals than OLS. It is the case of seller text variables, and the ratio $\frac{[L,U]^{postLASSO}}{[L,U]^{OLS}}$ 17.87 and 4.57 for 'Positive Adjectives' and 'Location Words', while the average for others (except for 'Entire Home/Apt') is 1.09.

Second is when the signal is 'too strong'. For 'Entire Home/Apt', the Z-score is 173. Then the lower end of truncation is very high and almost every value above it satisfies the significance level. The R package in this case produces the output 'inf'.

Table B.1: Confidence Intervals for OLS post Lasso

5% C.Is	OLS		Exact Truncated Normal			
obs: 75,236	C.Is		C.Is		Tail Areas	
	L	U	L	U	L	U
Quality Certification						
Superhost	0.044510	0.060512	0.044440	0.060600	0.024014	0.023888
Verification Accounts	0.014836	0.019844	0.014821	0.020556	0.024218	0.024826
Review Scores						
Cleanliness	0.045786	0.052211	0.045766	0.053591	0.024051	0.024403
Location	0.133259	0.140423	0.133232	0.140430	0.024261	0.024782
Value	-0.096141	-0.087599	-0.096158	-0.087576	0.024626	0.024383
Review Texts						
Negative Reviews	-0.008115	-0.006459	-0.008118	-0.006454	0.024613	0.024319
Positive Phrases	0.005995	0.008099	0.005986	0.008527	0.023892	0.024603
Seller Texts						
Positive Adjectives	-0.000550	0.000587	-0.020134	0.000175	0.024976	0.024988
Location Phrases	0.000841	0.001604	0.000927	0.004413	0.024842	0.024967
Accommodation Capacities						
Default Guests	0.053228	0.059414	0.053214	0.059429	0.024489	0.024443
Bathrooms	0.095556	0.111882	0.095064	0.111892	0.024449	0.024862
Bedrooms	0.104912	0.116032	0.104866	0.116618	0.024066	0.024889
Beds	-0.023948	-0.014518	-0.024029	-0.014506	0.024033	0.024716
Included Guests	0.021074	0.027366	0.021046	0.027368	0.024016	0.024923
Amenity and Service						
Air Conditioner	0.118190	0.134485	0.118118	0.134490	0.024008	0.024931
Buzzer Wireless Intercomm	0.082411	0.093107	0.082432	0.096908	0.024692	0.024774
Cable TV	0.094420	0.105542	0.090869	0.105517	0.024627	0.024924
Free Parking on Street	-0.129221	-0.112493	-0.129269	-0.111331	0.024326	0.024521
Indoor Fire Place	0.109043	0.135648	0.108097	0.135658	0.024230	0.024907
Lock on Bedroom Door	-0.076620	-0.058717	-0.076746	-0.058666	0.024509	0.024355
Cats Allowed	-0.094924	-0.074257	-0.094945	-0.072997	0.024770	0.024382
Internet	0.018999	0.032768	0.018958	0.036304	0.023992	0.024511
Shampoo	0.033656	0.044654	0.033637	0.045603	0.024591	0.024570
(Room Type)						
Entire Home/Apt	0.583494	0.596795	0.339306	inf	0	0
Shared Room	-0.185973	-0.142035	-0.186067	-0.141924	0.024510	0.024422

B.2 Manhattan vs. Other Neighborhoods

Table B.2: GMM: Manhattan and Other Neighborhoods

$obj : log(p_{it})$	Manhattan		Other Neighborhoods	
	Fixed Effects	GMM	Fixed Effects	GMM
Quality Certification				
Superhost	0.0116***	0.0124***	0.0086***	0.0084***
	(0.0030)	(0.0030)	(0.0027)	(0.0027)
Verification Accounts	0.0005	-0.0002	0.0034***	0.0061***
	(0.0010)	(0.0011)	(0.0010)	(0.0011)
Review Scores				
Cleanliness	-0.0025	-0.0003	0.0007	0.0022
	(0.0025)	(0.0030)	(0.0026)	(0.0030)
Location	0.0059*	0.0061	0.0044*	0.0087***
	(0.0034)	(0.0042)	(0.0027)	(0.0031)
Value	-0.0013	-0.0059*	-0.0061**	-0.0065**
	(0.0028)	(0.0033)	(0.0028)	(0.0031)
Review Texts				
Negative Reviews	-0.0007	-0.0025***	0.0006	-0.0022***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Positive Phrases	0.0045***	0.0042***	0.0029***	0.0031***
	(0.0009)	(0.0009)	(0.0008)	(0.0007)
Seller Texts				
Positive Adjectives	-0.0019***	-0.0014	0.0005	0.0002
	(0.0007)	(0.0011)	(0.0007)	(0.0009)
Location Phrases	0.0023***	0.0017***	0.0020***	0.0010*
	(0.0004)	(0.0006)	(0.0004)	(0.0005)
Constant		0.1875***		0.1616***
		(0.0118)		(0.0115)
$\rho : rho$		0.9631***		0.9667***
		(0.0024)		(0.0025)
Number of Obs	17,687		19,931	

***: 1% significant, **: 5% ,*: 10%, standard errors in parentheses

Table B.3: GMM: Manhattan and Other Neighborhoods (Continued from Table B.2)

$obj : log(p_{it})$	Manhattan		Other Neighborhoods	
	Fixed Effects	GMM	Fixed Effects	GMM
Accommodation Capacities				
Default Guests	0.0275***	0.0254***	0.0268***	0.0329***
	(0.0026)	(0.0048)	(0.0024)	(0.0043)
Bathrooms	0.0401***	0.0602***	0.0051	0.0135
	(0.0136)	(0.0189)	(0.0094)	(0.0148)
Bedrooms	0.0383***	0.0279***	0.0460***	0.0471***
	(0.0063)	(0.0106)	(0.0056)	(0.0105)
Beds	0.0245***	0.0260***	0.0051	0.0099**
	(0.0039)	(0.0062)	(0.0034)	(0.0049)
Included Guests	0.0163***	0.0152***	0.0150***	0.0149***
	(0.0029)	(0.0046)	(0.0024)	(0.0036)
Amenity and Service				
Air Conditioner	0.0060	-0.0043	-0.0025	-0.0095
	(0.0081)	(0.0116)	(0.0061)	(0.0086)
Buzzer Wireless Intercomm	0.0104*	0.0121	0.0145**	0.0118
	(0.0062)	(0.0077)	(0.0067)	(0.0093)
Cable TV	0.0011	0.0090	-0.0025	0.0068
	(0.0061)	(0.0071)	(0.0058)	(0.0071)
Free Parking	0.0014	-0.0020	-0.0088	-0.0067
	(0.0108)	(0.0138)	(0.0067)	(0.0089)
Indoor Fire Place	-0.0464**	-0.0234	0.0745***	0.0749***
	(0.0183)	(0.0216)	(0.0180)	(0.0217)
Lock on Bedroom Door	0.0119	-0.0047	-0.0124**	-0.0229***
	(0.0075)	(0.0101)	(0.0060)	(0.0073)
Cats Allowed	-0.0496***	-0.0397*	-0.0233**	-0.0115
	(0.0138)	(0.0218)	(0.0092)	(0.0148)
Internet	-0.0011	-0.0043	-0.0038	0.0027
	(0.0066)	(0.0082)	(0.0057)	(0.0063)
Shampoo	0.0019	0.0077	-0.0072	-0.0031
	(0.0051)	(0.0066)	(0.0047)	(0.0057)
Room Type				
Entire Home/Apt	0.1151***	0.1305***	0.1828***	0.1991***
	(0.0085)	(0.0152)	(0.0085)	(0.0164)
Shared Room	-0.0929***	-0.0571	-0.0060	0.0549
	(0.0215)	(0.0476)	(0.0228)	(0.0602)

B.3 Annual Variations in Attributes

Table B.4: Summary Statistics for Variables and Annual Variations

obs: 37,618	201712	201612	Difference (201712-201612)			
	Mean	Mean	Mean	S.D.	Min	Max
Price ($)	184.6176	181.8856	2.7319	31.0575	-550	700
Quality Certification						
Superhost Indicator	0.1761	0.1003	0.0758	0.3842	-1	1
Verification Accounts	4.4649	4.2075	0.2574	1.0407	-3	8
Review Scores						
Cleanliness	9.2026	9.1913	0.0113	0.4523	-4	6
Location	9.3968	9.3492	0.0476	0.3782	-4	7
Value	9.3077	9.2816	0.0262	0.4186	-4	7
Review Text						
Negative Reviews	4.8731	3.4239	1.4492	2.1540	-24	19
Positive Phrases	3.5182	2.5320	0.9862	1.6366	-38	14
Seller Text						
Positive Adjectives	6.6013	6.3382	0.2631	1.5587	-24	20
Location Phrases	10.6401	9.6204	1.0196	2.6019	-29	47
Accommodation Capacites						
Default Guests	2.9284	2.9287	-0.0002	0.5111	-14	14
Bedrooms	1.1092	1.1092	0.0001	0.1007	-2.5	3
Bathrooms	1.1567	1.1476	0.0091	0.2060	-6	4
Beds	1.5891	1.5709	0.0182	0.3436	-10	11
Guests Included	1.6164	1.5702	0.0462	0.4377	-11	13
Amenity and Service						
Air Conditioning	0.8842	0.8692	0.0150	0.1589	-1	1
Buzzer Wireless Intercom	0.5495	0.5480	0.0015	0.1753	-1	1
Cable TV	0.3688	0.3672	0.0016	0.1824	-1	1
Free Parking	0.1103	0.1128	-0.0026	0.1350	-1	1
Indoor Fire Place	0.0396	0.0400	-0.0004	0.0597	-1	1
Lock on Bedroom Door	0.1161	0.0945	0.0215	0.1633	-1	1
Cats Allowed	0.0675	0.0684	-0.0009	0.1004	-1	1
Internet	0.8133	0.8193	-0.0060	0.1860	-1	1
Shampoo	0.6231	0.5992	0.0238	0.2213	-1	1
(Room Type)						
Entire Home/Apt	0.5582	0.5579	0.0003	0.1369	-1	1
Shared Room	0.0141	0.0145	-0.0004	0.0497	-1	1

B.4 Evidence for Relevance of Lagged Instruments

Table B.5: Relevance Tests for Lagged Instruments

obs: 37,618	R Squared	Wald F (24, 37,618)
QUALITY CERTIFICATION		
Superhost Indicator	0.2623	13.81
Verification Accounts	0.2737	9.83
REVIEW SCORES		
Cleanliness	0.8487	19.55
Location	0.8180	22.28
Value	0.8023	23.29
REVIEW TEXT		
Negative Reviews	0.9404	44.99
Positive Phrases	0.9363	13.49
SELLER TEXT		
Positive Adjectives	0.9083	10.71
Location Phrases	0.8877	28.15

GMM method in Chapter 1 assumes that all of the attributes in X_{it} could be endogenous with the error η_{it}. Further lagged observations for attributes X_{it-2} are proposed as instruments. X_{it-2} contains rental unit attributes recorded three to four months earlier than those in X_{it-1} recorded in 2016. For example, for a rental i appearing in June, 2017 and 2016, x_{ikt-2} is the k-th attribute recorded in Febrary, 2016. Though GMM implementation does not require an explicit first stage regression, it is important that X_{it-2} satisfies the relevance condition.

Table B.5 provides evidence of relevance for each attribute $x_{ikt} \in X_{it}$, from a joint significance test for hypothesis (B.2) with $Z_{it} = (x_{i1t-2}, ..., x_{ikt-2}, ..., x_{iKt-2})$ and the corresponding coefficients vector $\Gamma = (\gamma_1, ..., \gamma_K)$. The regressor vector X_{it}^{-k} includes attributes x_{ikt}'s except for x_{ikt}, and X_{ikt-1} includes x_{ikt-1}'s for all k.

$$x_{ikt} = c + \Gamma Z_{it} + \theta_p log(p_{it-1}) + \Theta_1 X_{it}^{-k} + \Theta_2 X_{it-1} \tag{B.1}$$

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_{25} = 0 \tag{B.2}$$

The degrees of freedom for F-statistic is 25(=K) and 37,618, and the corresponding p-values for each variable are less than 0.0001. Table B.5 shows the relevance conditions for information variables, and the results are similar for amenity and service features.

APPENDIX C

OMITTED DETAILS FOR CHAPTER 2

C.1 Exact Inference for Post Lasso Estimates

Table C.1, C.2, and C.3 report the C.Is produced by classical OLS and exact post selection inference for OLS, IV and two level nested logit models, respectively. As a reminder, the first step LASSO selection was conducted separately on datasets for each estimation methods to see if the model selection varies due to endogeneity. For OLS logit, the LASSO selector chose variables from the dataset of all attributes and price, p_{jt}. For IV logit p_{jt} was replaced with the instrumented price $\hat{p}_{jt}$. For two level nested logit, $ln(\hat{s_{j|gt}})$ was included also with $\hat{p}_{jt}$. The LASSO selector, for three level nested logit with $ln(\hat{s_{j|hgt}})$ and $ln(\hat{s_{h|gt}})$ omitted a few key variables, including σ_2 for the regional level correlated preferences. It shows the selection results are susceptible and vary over endogeneity controls or econometric modeling choices.

C.Is from the exact inference reflecting the re-normalization of density due to truncation or, additional uncertainty due to LASSO selection are slightly wider than C.Is from OLS inference, but they essentially reproduce them for most of the variables, except for two cases.

The first case is when the signal of a variable is 'weak', or the correlation between the objective and covariate is small. Then either the variable is not chosen, or the parameter estimate could be near to one of the truncation endpoints, giving much wider intervals than OLS. For example, 'Location Words' was not selected by LASSO for OLS logit case. But it was selected in all the other cases, confirming the suspicion of instability of selection results due to endogeneity. 'Shared Room' showed slightly weak signals in both OLS and IV logit datasets, with $\frac{[L,U]^{Exact}}{[L,U]^{OLS}} = 1.3535$ and 1.1557 where the averages for others were about 1.0050.

'Instant Bookable' turn out to be insignificant or a weak signal in the two level nested logit case, with $ln(\hat{s_{j|gt}})$ and $\hat{p}_{jt}$. $\frac{[L,U]^{Exact}}{[L,U]^{OLS}}$ for 'Instant Bookable' is 1.8376 where the average for others is also about 1.0050.

The second case is when the signal is ’too strong’, which is the case for the two level nesting parameter σ with a Z-score of 99.67. Then the lower endpoint of truncation is set too high (0.73), and any value above it satisfies the significance level $\alpha = 0.05$. In this case, R package **selectiveInference** produces ’inf’ for the upper bound of the C.Is.

Table C.1: C.Is for OLS Logit, Classical vs. Post Selection Inference

Obs: 62,673	OLS		Exact Post LASSO		Tail Areas	
5% C.Is	L	U	L	U	L	U
Price	-0.000852	-0.000697	-0.000853	-0.000697	0.024498	0.024434
Quality Certifications						
Superhost Indicator	0.190012	0.236667	0.189803	0.236677	0.023989	0.024951
Verification Accounts	0.040741	0.055015	0.040685	0.055027	0.024116	0.024819
Consumer Review						
Ratings Average	0.180669	0.206865	0.180619	0.206938	0.024568	0.024364
Number of Reviews	0.016606	0.017638	0.016606	0.017643	0.024950	0.023989
Negative Reviews	-0.041592	-0.034697	-0.041624	-0.034695	0.023977	0.024963
Seller Texts						
Positive Adjectives	-0.016798	-0.008212	-0.016828	-0.008202	0.024202	0.024739
Location Phrases	(Not chosen by the first step LASSO selector)					
Accommodation Capacities						
Default Guests	0.089439	0.104727	0.089414	0.107401	0.024505	0.024389
Bathrooms	0.037290	0.083423	0.037103	0.083452	0.024192	0.024852
Additional Guests	-0.005437	0.013559	-0.023288	0.013201	0.024951	0.024513
Instant Bookable	0.461827	0.499565	0.461756	0.499671	0.024571	0.024361
Amenity and Service						
24 Hour Check-In	0.114049	0.149457	0.114005	0.149579	0.024715	0.024218
Hangers	0.127225	0.160103	0.127082	0.160115	0.024023	0.024915
Heating	0.033329	0.104381	0.030885	0.104490	0.024740	0.024649
Shampoo	0.126503	0.158248	0.126501	0.158395	0.024984	0.023957
(Room Type)						
Entire Home/Apt	-0.102712	-0.064644	-0.102753	-0.064505	0.024756	0.024179
Shared Room	0.157776	0.245300	0.157417	0.245352	0.024074	0.024863

Table C.2: C.Is for IV Logit, Classical vs. Post Selection Inference

Obs: 62,673	2SLS		Exact Post Selection		Tail Areas	
5% C.Is	L	U	L	U	L	U
Price (Instrumented)	-0.015932	-0.015179	-0.015936	-0.015175	0.023905	0.023868
Quality Certifications						
Superhost Indicator	0.315313	0.360276	0.315160	0.360333	0.024226	0.024708
Verification Accounts	0.077299	0.091051	0.077279	0.091096	0.024674	0.024259
Consumer Review						
Ratings Average	0.385672	0.412724	0.385632	0.412811	0.024660	0.024273
Number of Reviews	0.015625	0.016612	0.015622	0.016613	0.024174	0.024760
Negative Reviews	-0.041871	-0.035289	-0.041886	-0.035274	0.024456	0.024476
Seller Texts						
Positive Adjectives	-0.063586	-0.054530	-0.063620	-0.054521	0.024171	0.024764
Location Phrases	0.006822	0.013240	0.006799	0.013247	0.024193	0.024744
Accommodation Capacities						
Default Guests	0.366664	0.386868	0.366592	0.386891	0.024193	0.024741
Bathrooms	0.695131	0.750209	0.694847	0.750469	0.023839	0.023933
Additional Guests	0.162532	0.182525	0.162514	0.182601	0.024797	0.024139
Instant Bookable	0.314323	0.351105	0.314310	0.351265	0.024917	0.024021
Amenity and Service						
24 Hour Check-In	0.159904	0.193794	0.159750	0.193799	0.023976	0.024965
Hangers	0.125852	0.157232	0.125703	0.157394	0.023932	0.023840
Heating	0.171242	0.239436	0.170930	0.239444	0.023969	0.024972
Shampoo	0.213911	0.244531	0.213832	0.244595	0.024407	0.024525
(Room Type)						
Entire Home/Apt	1.372216	1.455404	1.372104	1.455681	0.024689	0.024244
Shared Room	-0.149253	-0.064283	-0.149284	-0.034280	0.024190	0.024747

Table C.3: C.Is for Two Level Nested Logit, Classical vs. Post Selection Inference

Obs: 62,673	2SLS		Exact Post Selection		Tail Areas	
5% C.Is	L	U	L	U	L	U
Price (Instrumented)	-0.006324	-0.005529	-0.006327	-0.005527	0.024296	0.024636
Nesting Parameters						
$\sigma(\sigma_1)$	0.735739	0.765255	0.735592	inf	0.023852	0
Quality Certifications						
Superhost Indicator	0.043312	0.086441	0.043220	0.086551	0.024708	0.024418
Verification Accounts	0.059870	0.072666	0.059849	0.072705	0.024623	0.024309
Consumer Review						
Ratings Average	0.108272	0.135667	0.108169	0.135693	0.024167	0.024784
Number of Reviews	0.004012	0.005035	0.004010	0.005038	0.024532	0.024399
Negative Reviews	-0.014863	-0.008658	-0.014875	-0.008641	0.024561	0.024370
Seller Texts						
Positive Adjectives	-0.039552	-0.031086	-0.039595	-0.031045	0.023859	0.023913
Location Phrases	0.013800	0.019769	0.013779	0.019776	0.024211	0.024723
Accommodation Capacities						
Default Guests	0.091769	0.113417	0.091673	0.113422	0.023998	0.024941
Bathrooms	0.284585	0.338250	0.284530	0.338447	0.024765	0.024170
Additional Guests	0.061140	0.080142	0.061081	0.080172	0.024298	0.024635
Instant Bookable	-0.005903	0.030520	-0.036878	0.030052	0.024883	0.023929
Amenity and Service						
24 Hour Check-In	0.040703	0.072543	0.040690	0.072679	0.024902	0.024036
Hangers	0.019869	0.049325	0.019704	0.049452	0.023948	0.024029
Heating	0.072892	0.136373	0.072865	0.136643	0.024932	0.024039
Shampoo	0.060668	0.089754	0.060639	0.089861	0.024771	0.024164
(Room Type)						
Entire Home/Apt	0.684688	0.766577	0.684392	0.766664	0.024180	0.024754
Shared Room	-0.115604	-0.036653	-0.115694	-0.024454	0.024643	0.024571

C.2 First Stage Regression for Prices and Group Shares

Table C.4: First Stage Regression Results

Obs: 62,673 / Obj:	p_{jt}	$ln(s_{j\|gt})$	$ln(s_{j\|hgt})$	$ln(s_{h\|gt})$
Lagged Base Price	0.016891***	0.000001	-0.000007	0.000008
$(p_{jt-1} - CleaningFee_{jt-1})$	(0.000805)	(0.000009)	(0.000014)	(0.000013)
Business Starting Date	-0.004411***	0.000053***	0.000041***	0.000013*
	(0.000445)	(0.000005)	(0.000007)	(0.000007)
Dates Last Scraped	-0.011947***	-0.000615***	-0.000632***	0.000018
	(0.002884)	(0.000032)	(0.000048)	(0.000046)
Availabilities				
30 Days	0.657721***	0.002714***	0.004588***	-0.001874
	(0.085911)	(0.000947)	(0.001439)	(0.001380)
60 Days	0.015415	-0.006876***	-0.007233***	0.000357
	(0.087163)	(0.000961)	(0.001460)	(0.001400)
90 Days	0.083191*	0.006025***	0.004392***	0.001632**
	(0.046007)	(0.000507)	(0.000771)	(0.000739)
365 Days	0.024166***	0.000299***	-0.000059	0.000358***
	(0.003571)	(0.000039)	(0.000060)	(0.000057)
Reviews per Month	-8.706140***	0.401024***	0.400606***	0.000418
	(0.332901)	(0.003670)	(0.005575)	(0.005347)
Cancellation Policy				
Moderate	-8.146719***	0.043266***	0.035425***	0.007842
	(0.801306)	(0.008833)	(0.013420)	(0.012870)
R Squared	0.43	0.36	0.18	0.01
Wald F Statistic: F(9, 62673)	279.57	1672.05	680.71	21.99

: 5% significant, *: 1%, and standard errors in parentheses

To instrument price p_{jt} and group shares, variables reflecting supply side decisions were chosen as the first stage regressors. Table C.4 reports the coefficients from the first stage regressions (only for instruments). To test the relevance condition of IVs, I report the F statistics for the following hypothesis for each objective variable. For price, it would be

$$p_{jt} = constant + z_{jt}\gamma_p + x_{jt}\theta_p + \epsilon_{jt} \tag{C.1}$$

$$H_0 : \gamma_{p1} = \gamma_{p2} = \cdots = \gamma_{p9} = 0$$

'Lagged Base Price' means per night rental price minus 'Cleaning Fee' recorded one year before for all rental units in the dataset. One important fact is that it indirectly reflects hosts' decisions to impose or remove the 'Cleaning Fee'. Also, there is a big difference in the means ($p_{jt} - (p_{jt-1} - CleaningFee_{jt-1}) = -\$159.1783$) with a standard deviation of $455.2833 implying heavy adjustments in prices and 'Cleaning Fee' imposing decisions. Such big changes reflect the tendency of new Airbnb hosts setting high prices at the start of business and constantly decreasing prices as they face low demand due to heavy competition.

'Business Starting Date' is the date a host started Airbnb hosting, and 'Dates Last Scraped' is the date InsideAirbnb.com last recorded the data on the host. Both are in the format of cumulative days from a starting point. For example, if a host started hosting on Jan 4, 2016, then 'Business Starting Date' is the cumulative days since Jan 1, 1900.

'Availability' variables represent how many consecutive days a host or a rental unit can provide. For example, the variable '60 Days' ranges from 0 to 60. If it is 40, then a potential guest can make a reservation request with a maximum length of 40 days. Deciding such lengths of maximum nights is mostly in the hands of rental hosts, unlike hotels.

'Moderate' cancellation policy contains a host's policy on refunds, reservation modifications or cancellation. There are six more categories including 'Flexible', 'Strict', and 'Long Term'. 'Reviews per Month' is not the cumulative number of reviews divided by months of operation up to date, but it is the average number of reviews received in a month when data scraping occurred. It was included to indirectly capture an exogenous variation in supply decisions, given that Airbnb hosts can set their business days as flexibly as imaginable.

One reason for not using BLP instruments (the isolation measure $\sum_{j \neq r} x_{jk}$) or, product characteristic IVs is that when the market is large or 'thick' meaning there are too many close substitutes, the identifying power of them could be in doubt. In this case, cost shifters or supply side instruments could provide a better identifying power (Armstrong (2016)).

C.3 NYC Service Neighborhoods

Table C.5: NYC Airbnb Service Neighborhoods

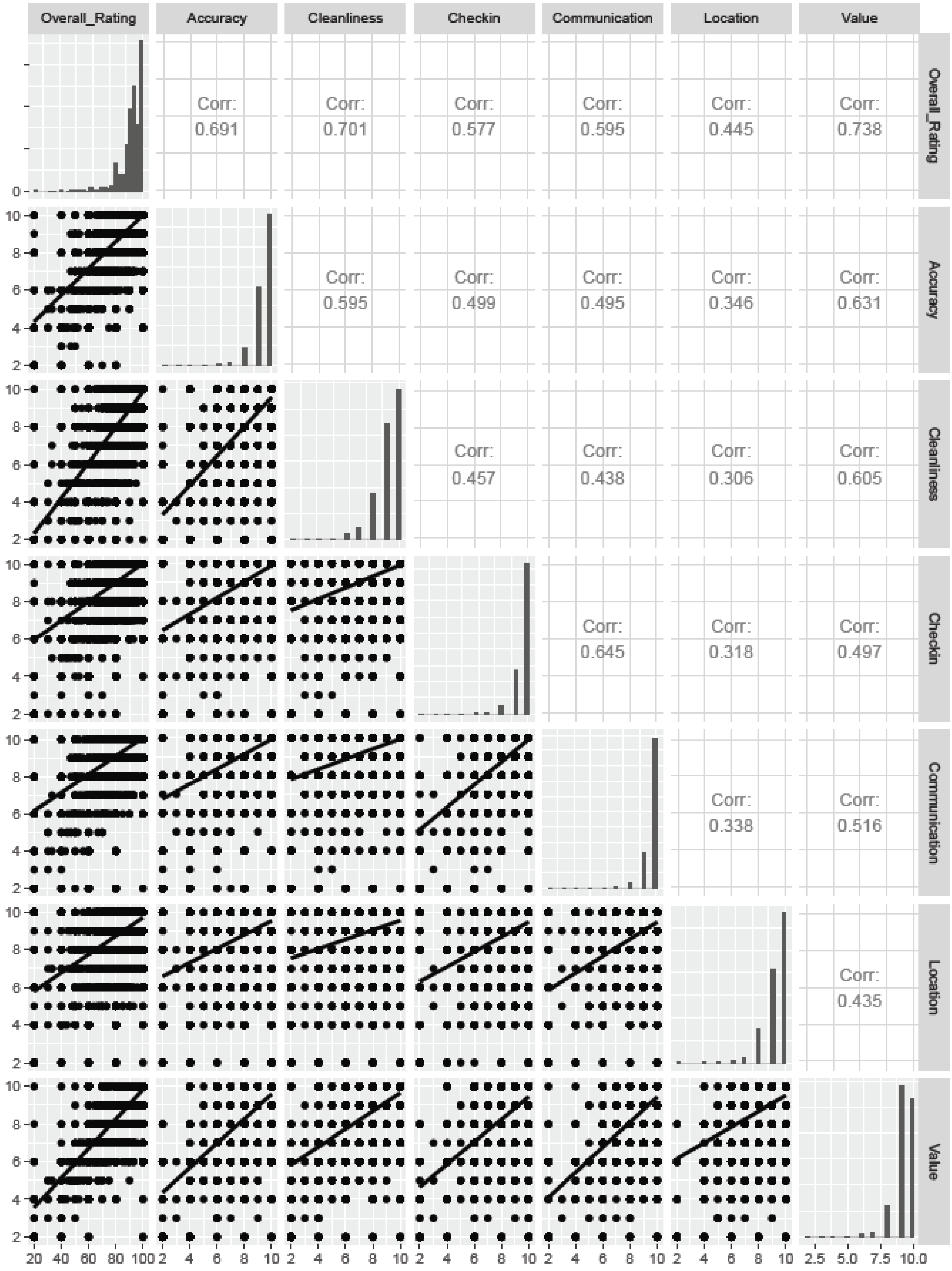
Bronx	Brooklyn	Manhattan	Queens	Staten Island
Allerton	Bath Beach	Battery Park City	Arverne	Arden Heights
Baychester	Bay Ridge	Chelsea	Astoria	Arrochar
Belmont	Bedford-Stuyvesant	Chinatown	Bay Terrace	Bay Terrace
Bronxdale	Bensonhurst	Civic Center	Bayside	Bull’s Head
Castle Hill	Bergen Beach	East Harlem	Bayswater	Castleton Corners
City Island	Boerum Hill	East Village	Belle Harbor	Charleston
Claremont Village	Borough Park	Financial District	Bellerose	Chelsea
Clason Point	Brighton Beach	Flatiron District	Breezy Point	Clifton
Concourse	Brooklyn Heights	Gramercy	Briarwood	Concord
Concourse Village	Brownsville	Greenwich Village	Cambria Heights	Dongan Hills
Co-op City	Bushwick	Harlem	College Point	Eltingville
Country Club	Canarsie	Hell’s Kitchen	Corona	Emerson Hill
East Morrisania	Carroll Gardens	Inwood	Ditmars Steinway	Fort Wadsworth
Eastchester	Clinton Hill	Kips Bay	Douglaston	Graniteville
Edenwald	Cobble Hill	Little Italy	East Elmhurst	Great Kills
Fieldston	Columbia St	Lower East Side	Edgemere	Grymes Hill
Fordham	Coney Island	Marble Hill	Elmhurst	Howland Hook
Highbridge	Crown Heights	Midtown	Far Rockaway	Huguenot
Hunts Point	Cypress Hills	Morningside Heights	Flushing	Lighthouse Hill
Kingsbridge	Downtown Brooklyn	Murray Hill	Forest Hills	Mariners Harbor
Longwood	DUMBO	NoHo	Fresh Meadows	Midland Beach
Melrose	Dyker Heights	Nolita	Glen Oaks	New Brighton
Morris Heights	East Flatbush	Roosevelt Island	Glendale	New Dorp
Morris Park	East New York	SoHo	Hollis	New Dorp Beach
Morrisania	Flatbush	Stuyvesant Town	Hollis Hills	New Springville
Mott Haven	Flatlands	Theater District	Holliswood	Oakwood
Mount Eden	Fort Greene	Tribeca	Howard Beach	Port Richmond
Mount Hope	Fort Hamilton	Two Bridges	Jackson Heights	Prince’s Bay
North Riverdale	Gerritsen Beach	Upper East Side	Jamaica	Randall Manor
Norwood	Gowanus	Upper West Side	Jamaica Estates	Richmondtown
Olinville	Gravesend	Washington Heights	Jamaica Hills	Rosebank
Parkchester	Greenpoint	West Village	Kew Gardens	Rossville
Pelham Bay	Kensington		Kew Gardens Hills	Shore Acres
Pelham Gardens	Manhattan Beach		Laurelton	Silver Lake
Port Morris	Midwood		Little Neck	South Beach
Riverdale	Mill Basin		Long Island City	St. George
Schuylerville	Navy Yard		Maspeth	Stapleton
Soundview	Park Slope		Middle Village	Todt Hill
Spuyten Duyvil	Prospect Heights		Neponsit	Tompkinsville

Table C.6: NYC Airbnb Service Neighborhoods (Continued from Table C.5)

Bronx	Brooklyn	Manhattan	Queens	Staten Island
Throgs Neck	Prospect-Lefferts Gardens		Ozone Park	Tottenville
Tremont	Red Hook		Queens Village	West Brighton
Unionport	Sea Gate		Rego Park	Westerleigh
University Heights	Sheepshead Bay		Richmond Hill	Willowbrook
Van Nest	South Slope		Ridgewood	Woodrow
Wakefield	Sunset Park		Rockaway Beach	
West Farms	Vinegar Hill		Rosedale	
Westchester Square	Williamsburg		South Ozone Park	
Williamsbridge	Windsor Terrace		Springfield Gardens	
Woodlawn			St. Albans	
			Sunnyside	
			Whitestone	
			Woodhaven	
			Woodside	
Subtotals				
49	48	32	53	44

C.4 Evidence for Review Rating Inflation

Figure C.1: Correlation Across Review Score Categories



BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Akerlof, G. A., Aug. 1970. The market for "lemons": Quality uncertainty and the market mechanism. The Quarterly Journal of Economics 83 (3), 488–500.

[2] Alaei, A. R., Becken, S., Stantic, B., Dec. 2017. Sentiment analysis in tourism: Capitalizing on big data. Journal of Travel Research, 1–17.

[3] Armstrong, T., Aug. 2016. Large market asymptotics for differentiated product demand estimators with economic models of supply. Econometrica 84 (5), 1961–1980.

[4] Bajari, P., Fruewirth, J. C., Kim, K. I., Timmins, C., Aug. 2012. A rational expectations approach to hedonic price regressions with time-varying unobserved product attributes: The price of pollution. American Economic Review 102 (5), 1898–1926.

[5] Bajari, P., Nekipelov, D., Ryan, S. P., Yang, M., May 2015. Machine learning methods for demand estimation. American Economic Review 105 (5), 481–85.

[6] Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19 (2), 521–547.

[7] Belloni, A., Chernozhukov, V., Fernandez-Val, I., Hansen, C., Jan. 2017. Program evaluation and causal inference with high-dimensional data. Econometrica 85 (1), 233–298.

[8] Belloni, A., Chernozhukov, V., Wang, L., 2014. Pivotal estimation via square root lasso in nonparametric regression. The Annals of Statistics 42 (2), 757–788.

[9] Berry, S., Levinsohn, J., Pakes, A., Jul. 1995. Automobile prices in market equilibrium. Econometrica 63 (4), 841–890.

[10] Berry, S., Linton, O. B., Pakes, A., 2004. Limit theorems for estimating the parameters of differentiated product demand systems. Review of Economic Studies 71, 613–654.

[11] Berry, S., Pakes, A., Nov. 2007. The pure characteristics demand model. International Economic Review 48 (4).

[12] Berry, S. T., 1994. Estimating discrete-choice models of product differentiation. The RAND Journal of Economics 25 (2), 242–262.

[13] Bjornerstedt, J., Verboven, F., Jul. 2016. Does merger simulation work? evidence from the swedish analgesics market. American Economic Journal: Applied Economics 8 (3), 125–164.

[14] Breiman, L., Aug. 1996. Bagging predictors. Machine Learning 24 (2), 123–140.

[15] Breiman, L., Oct. 2001. Random forests. Machine Learning 45 (1), 5–32.

[16] Brenkers, R., Verboven, F., Dec. 2010. Liberalizing a distribution system: The european car market. Journal of the European Economic Association 4 (1).

[17] Cardell, N. S., Apr. 1997. Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. Econometric Theory 13 (2), 185–213.

[18] Chen, Y., Xie, K., 2017. Consumer valuation of airbnb listings: A hedonic pricing approach. International Journal of Contemporary Hospitality Management 29 (9), 2405–2424.

[19] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., Feb. 2018. Double, debiased machine learning for treatment and structural parameters. The Econometrics Journal 21 (1), C1–C68.

[20] Chernozhukov, V., Hansen, C., Spindler, M., May 2015. Post-selection and post-regularization inference in linear models with many controls and instruments. American Economic Review 105 (5), 486–490.

[21] Chevalier, J. A., Mayzlin, D., Aug. 2006. The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research 43 (3), 345–354.

[22] Chintagunta, P. K., Gopinath, S., Venkataraman, S., Sep. 2010. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. Marketing Science 29 (5), 944–957.

[23] Ciliberto, F., Moschini, G., Perry, E. D., Dec. 2017. Valuing product innovation: Genetically engineered varieties in u.s. corn and soybeans. Center for Agricultural and Rural Develop-mentWorking Paper 17-WP 576, Available at SSRN: https://ssrn.com/abstract=3088632 or https://dx.doi.org/10.2139/ssrn.3088632.

[24] Cui, G., Lui, H.-K., Guo, X., 2012. The effects of online consumer reviews on new product sales. International Journal of Electronic Commerce 17 (1), 39–58.

[25] Dellarocas, C., Zhang, X. M., Awad, N. F., 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. Journal of Interactive Marketing 21 (4).

[26] Dhar, V., Chang, E. A., Nov. 2009. Does chatter matter? the impact of user-generated contents on music sales. Journal of Interactive Marketing 23 (3), 300–307.

[27] Draganska, M., Klapper, D., Aug. 2011. Choice set heterogeneity and the role of advertising: An analysis with micro and macro data. Journal of Marketing Research 48 (4), 653–669.

[28] Duan, W., Gu, B., Whinston, A. B., Nov. 2008. Do online reviews matter? an empirical investigation of panel data. Decision Support Systems 45 (4), 1007–1016.

[29] Edelman, B., Luca, M., Svirsky, D., Apr. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. American Economic Journal: Applied Economics 9 (2), 1–22.

[30] Ert, E., Fleischer, A., Magen, N., Aug. 2016. Trust and reputation in the sharing economy: The role of personal photos in airbnb. Tourism Management 55, 62–73.

[31] Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., Freling, T., Jun. 2014. How online product reviews affect retail sales: A meta-analysis. Journal of Retailing 90 (2), 217–232.

[32] Fox, J., Kim, K. I., Ryan, S. P., Bajari, P., Feb. 2012. The random coefficients logit model is identified. Journal of Econometrics 166 (2), 204–212.

[33] Fox, J., Kim, K. I., Yang, C., Dec. 2016. A simple nonparametric approach to estimating the distribution of random coefficients in structural models. Journal of Econometrics 195 (2), 236–254.

[34] Fox, J. T., Kim, K. I., Ryan, S. P., Bajari, P., 2011. A simple estimator for the distribution of random coefficients. Quantitative Economics 2, 381–418.

[35] Fradkin, A., 2017. Search, matching, and the role of digital marketplace design in enabling trade: Evidence from airbnb.

[36] Fradkin, A., Grewal, E., Holtz, D., 2018. The determinants of online review informativeness: Evidence from field experiments on airbnb.

[37] Friedman, J. H., Oct. 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29 (5), 1189–1232.

[38] Fu, W., Knight, K., 2000. Asymptotics for lasso-type estimators. The Annals of Statistics 28 (5), 1356–1378.

[39] Ghose, A., Ipeirotis, P. G., Oct. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering 23 (10), 1498–1512.

[40] Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., Goodwill, A., Apr. 2018. Pricing in the sharing economy: a hedonic pricing model applied to airbnb listings. Journal of Travel and Tourism Marketing 35 (1), 46–56.

[41] Green, P. E., Carmone, F. J., Wachspress, D. P., Dec. 1976. Consumer segmentation via latent class analysis. Journal of Consumer Research 3 (3), 170–174.

[42] Greif, A., 2006. Institutions and the Path to the Modern Economy. Cambridge University Press, Cambridge, UK.

[43] Grossman, S., 1981. The informational role of warranties and private disclosure about product quality. Journal of Law and Economics 24 (3), 461–483.

[44] Grossman, S. J., Hart, O. D., 1980. Takeover bids, the free-rider problem, and the theory of the corporation. The Bell Journal of Economics 11 (1), 42–64.

[45] Guttentag, D., 2015. Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. Current Issues in Tourism 18 (12), 1192–1217.

[46] Hausmann, J. A., Leonard, G. K., Sep. 2002. The competitive effects of a new product introduction: A case study. The Journal of Industrial Economics 50 (3), 237–263.

[47] Herriges, J. A., Phaneuf, D. J., Nov. 2002. Inducing patterns of correlation and substitution in repeated logit models of recreation demand. American Journal of Agricultural Economics 84 (4), 1076–7090.

[48] Hoerl, A. E., Kennard, R. W., Feb. 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

[49] Horton, J. J., Zeckhauser, R. J., 2016. Owning, using and renting: Some simple economics of the "sharing economy"NBER Working Paper No 22029.

[50] Houser, D., Wooders, J., 2006. Reputation in auctions: Theory, and evidence from ebay. Journal of Economics & Management Strategy 15 (2), 353–369.

[51] Jin, G. Z., Kato, A., 2006. Price, quality, and reputation: Evidence from an online field experiment. The RAND Journal of Economics 37 (4), 983–1004.

[52] Kim, H., Kim, K. I., Sep. 2017. Estimating store choices with endogenous shopping bundles and price uncertainty. International Journal of Industrial Organization 54, 1–36.

[53] Kim, J. B., Albuquerque, P., Bronnenberg, B. J., Nov. 2010. Online demand under limited consumer search. Marketing Science 29 (6), 1001–1023.

[54] Kreps, D. M., 1990. Corporate Culture and Economic Theory. Cambridge University Press.

[55] Kreps, D. M., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational cooperation in the finitely repeated prisoners' dilemma. Journal of Economic Theory 27 (2), 245–252.

[56] Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., 2016. Exact post-selection inference, with application to the lasso. The Annals of Statistics 44 (3), 907–927.

[57] Lewis, G., Jun. 2011. Asymmetric information, adverse selection and online disclosure: The case of ebay motors. American Economic Review 101 (4), 1535–1546.

[58] Lewis, G., Zervas, G., 2016. The welfare impact of consumer reviews: A case study of the hotel industry. Industrial Organization Workshop (Pennsylvania State University, 2016).

[59] Liang, S., Schuckert, M., Law, R., Chen, C., Jun. 2017. Be a "superhost": The importance of badge systems for peer-to-peer rental accommodations. Tourism Management 60, 454–465.

[60] Liu, Y., Jul. 2006. Word of mouth for movies: Its dynamics and impact on box office revenue. Journal of Marketing 70 (3), 74–89.

[61] Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R., 2014. A significance test for the lasso. The Annals of Statistics 42 (2), 413–468.

[62] McFadden, D., 1978. Modeling the choice of residential location. In Spatial Interaction Theory and Planning Models, ed. by A. Karlgvist, et al. Amsterdam: North-HollanCowles Foundation Discussion Papers 477, Yale University.

[63] Milgrom, P. R., 1981. Good news and bad news: Representation theorems and applications. The Bell Journal of Economics 12 (2), 380–391.

[64] Milgrom, P. R., North, D. C., Weingast, B. R., Mar. 1990. The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. Economics and Politics 2 (1), 1–23.

[65] Murdock, J., Jan. 2006. Handling unobserved site characteristics in random utility models of recreation demand. Journal of Environmental Economics and Management 51 (1), 1–25.

[66] Nevo, A., Mar. 2001. Measuring market power in the ready-to-eat cereal industry. Econometrica 69 (2), 307–342.

[67] Nowak, A., Smith, P., 2017. Textual analysis in real estate. Journal of Applied Econometrics 32, 896–918.

[68] Ogut, H., Tas, B. K. O., Aug. 2012. The influence of internet customer reviews on the online sales and prices in hotel indsutry. The Service Industries Journal 32 (2), 197–213.

[69] Petrin, A., Aug. 2002. Quantifying the benefits of new products: The case of the minivan. Journal of Political Economy 110 (4), 705–729, nBER Working Paper No. 8227.

[70] Proserpio, D., Xu, W., Zervas, G., 2016. You get what you give: Theory and evidence of reciprocity in the sharing economy.

[71] Resnick, P., Zeckhauser, R., Swanson, J., Lockwood, K., Jun. 2006. The value of reputation on ebay: A controlled experiment. Experimental Economics 9 (2), 79–101.

[72] Rosen, S., Jan. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy 82 (1), 34–55.

[73] Shin, H. S., Hanssens, D. M., Kim, K. I., Dec. 2016. The role of online buzz for leader versus challenger brands: The case of the mp3 player market. Electronic Commerce Research 16 (4), 503–528.

[74] Tadelis, S., Oct. 2016. Reputation and feedback systems in online platform markets. Annual Review of Economics 8, 321–340.

[75] Taylor, J., Tibshirani, R., Mar. 2017. Post-selection inference for l1 penalized likelihood models. The Canadian Journal of Statistics 46 (1), 41–61, special Issue on Big Data and the Statistical Sciences.

[76] Teubner, T., Hawlitschek, F., Dann, D., 2017. Price determinants on airbnb: How reputation pays off in the sharing economy. Journal of Self-Governance and Management Science 5 (4), 53–80.

[77] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B Methodological 58 (1), 267–288.

[78] Tibshirani, R. J., Taylor, J., Lockhart, R., Tibshirani, R., Aug. 2016. Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association 111 (514), 600–620.

[79] Train, K. E., 2003. Discrete Choice Methods with Simulation. Cambridge University Press.

[80] Verboven, F., 1996. International price discrimination in the european car market. The RAND Journal of Economics 27 (2), 240–268.

[81] Wang, D., Nicolau, J. L., Apr. 2017. Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb.com. International Journal of Hospitality Management 62, 120–131.

[82] Xie, K. L., Zhang, Z., Zhang, Z., Oct. 2014. The business value of online consumer reviews and management response to hotel performance. International Journal of Hospitality Management 43, 1–12.

[83] Ye, Q., Law, R., Gu, B., Mar. 2009. The impact of online user reviews and hotel room sales. International Journal of Hospitality Management 28 (1), 180–182.

[84] Ye, Q., Law, R., Gu, B., Chen, W., Mar. 2011. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Computers in Human Behavior 27 (2), 634–639.

[85] Zervas, G., Proserpio, D., Byers, J. W., Jan. 2015. A first look at online reputation on airbnb, where every stay is above average. Available at SSRN: https://ssrn.com/abstract=2554500.

[86] Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B Methodological 67 (2), 301–320.