Essays on Demand Estimation, Financial Economics and Machine Learning

Pu He

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Essays on Demand Estimation, Financial Economics and Machine Learning

Pu He

In this era of big data, we often rely on techniques ranging from simple linear regression, structural estimation, and state-of-the-art machine learning algorithms to make operational and financial decisions based on data. This calls for a deep understanding of practical and theoretical aspects of methods and models from statistics, econometrics, and computer science, combined with relevant domain knowledge. In this thesis, we study several practical, data-related problems in the particular domains of sharing economy and financial economics/financial engineering, using appropriate approaches from an arsenal of data-analysis tools. On the methodological front, we propose a new estimator for classic demand estimation problem in economics, which is important for pricing and revenue management.

In the first part of this thesis, we study customer preference for the bike share system in London, in order to provide policy recommendations on bike share system design and expansion. We estimate a structural demand model on the station network to learn the preference parameters, and use the estimated model to provide insights on the design and expansion of the system. We highlight the importance of network effects in understanding customer demand and evaluating expansion strategies of transportation networks. In the particular example of the London bike share system, we find that allocating resources to some areas of the station network can be 10 times more beneficial than others in terms of system usage, and that currently implemented station density rule is far from optimal. We develop a new method to deal with the

endogeneity problem of the choice set in estimating demand for network products. Our method can be applied to other settings, in which the available set of products or services depends on demand.

In the second part of this thesis, we study demand estimation methodology when data has a long-tail pattern, that is, when a significant portion of products have zero or very few sales. Long-tail distributions in sales or market share data have long been an issue in empirical studies in areas such as economics, operations, and marketing, and it is increasingly common nowadays with more detailed levels of data available and many more products being offered in places like online retailers and platforms. The classic demand estimation framework cannot deal with zero sales, which yields inconsistent estimates. More importantly, biased demand estimates, if used as an input to subsequent tasks such as pricing, lead to managerial decisions that are far from optimal. We introduce two new two-stage estimators to solve the problem: our solutions apply machine learning algorithms to estimate market shares in the first stage, and in the second stage, we utilize the first-stage results to correct for the selection bias in demand estimates. We find that our approach works better than traditional methods using simulations.

In the third part of this thesis, we study how to extract a signal from option pricing models to form a profitable stock trading strategy. Recent work has documented *roughness* in the time series of stock market volatility and investigated its implications for option pricing. We study a strategy for trading stocks based on measures of their implied and realized roughness. A strategy that goes long the roughest-volatility stocks and short the smoothest-volatility stocks earns statistically significant excess annual returns of 6% or more, depending on the time period and strategy details. Standard factors do not explain the profitability of the strategy. We compare alternative measures of roughness in volatility and find that the profitability of the strategy is greater when we sort stocks based on implied rather than realized rough-

ness. We interpret the profitability of the strategy as compensation for near-term idiosyncratic event risk.

Lastly, we apply a heterogeneous treatment effect (HTE) estimator from statistics and machine learning to financial asset pricing. Recent progress in the interdisciplinary area of causal inference and machine learning has proposed various promising estimators for HTE. We take the R-learner algorithm by [73] and adapt it to empirical asset pricing. We study characteristics associated with standard factors, size, value and momentum through the lens of HTE. Our goal is to identify sub-universes of stocks, "characteristic responders", in which size, value or momentum trading strategies perform best, compared with the performance had they been applied to the entire universe. On the other hand, we identify subsets of "characteristic traps" in which the strategies perform the worst. In our test period, the differences in average monthly returns between long-short strategies restricted to "characteristic responders" and "characteristic traps" range from 0.77% to 1.54% depending on treatment characteristics. The differences are statistically significant and cannot be explained by standard factors: a long-short of long-short strategy generates α of significant magnitude from 0.98% to 1.80% monthly, with respect to standard Fama-French plus momentum factors. Simple interaction terms between standard factors and ex-post important features do not explain the alphas either. We also characterize and interpret the characteristic traps and responders identified by our algorithm. Our study can be viewed as a systematic, data-driven way to investigate interaction effects between features and treatment characteristic, and to identify characteristic traps and responders.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisers, Professor Paul Glasserman and Professor Fanyin Zheng, for guiding me through the PhD program and showing me how to be a researcher. I have benefited a great deal from our weekly/bi-weekly discussions. Paul has been my academic role model for his dedication to research, innovative thinking, and depth and breadth of knowledge. In addition to technical skills, I admire his ability to describe complicated ideas and identify key issues in research, which I am still trying to emulate. Fanyin has shown me the way to empirical research in operation management and economics. Working with the two of them, I've developed great economic intuitions behind econometric models and empirical results, which has benefited me not only in academic research, but also during my industry experience. I also want to thank them for being a constant source of support and motivation for me throughout the years. I am very grateful to be advised by both of them.

I am also thankful to my other committee members, Professors Costis Maglaras, Karan Girotra, and Mark Broadie, for their helpful comments on this thesis. I have benefited benefit a lot from discussions with Costis about research and life in general throughout the years. I would like to express my gratitude to him for keeping the door open for me. I would also like to thank Professors Omar Besbes, Carri Chan, and Yash Kanoria for checking my progress periodically and keeping me on track to finish the degree. I owe my heartfelt gratitude to Professor Fangruo Chen and Professor Mei Xue for their support and advice during the past 9 years.

I have been fortunate enough to work with the most talented classmates and fellow

PhD students here at the DRO devision of Columbia Business School. I am grateful for their support and friendship during the journey. In particular, I want to thank Amine Allouah and Francisco Castro for the mutual support especially during the first year of the PhD program. Special thanks go to Zhe Liu, Kai Yuan, Lijian Lu, Seungki Min, Dongwook Shin, and many others for making the 4th floor of Uris such a fun place to work.

I owe a great deal to my parents Lin He and Yurong Wang. I want to thank them for the sacrifices they've made for me. They are the best parents a son or daughter could possibly have. They raised me up with all the love, caring, and kindness in the world. They have taught me to be a man of dignity, to stand for what is right, and to be grateful to those who have helped me. Their kind souls, attitude, perseverance, and integrity have deeply inspired me. Without them, I wouldn't be where I am today.

Last but not least, I am grateful to Lu Zheng for her love, patience and support when I work on research and writing this thesis.

To my family

谨献给: 王玉荣，贺林，郑璐，贺书晗

Introduction

Part I Demand Estimation: Methodology and Applications

The sharing economy and online platforms have experienced a boom in recent years that has come with many interesting and challenging questions for their operations. Part I of this thesis studies demand estimation problems in this domain with the goal of improving managerial decision making for companies and operation managers. In particular, Chapter 1 of Part I focuses on bike share system design for London, where we estimate a demand model for commuters and use it to provide guidance on system design and expansion for the operating company. Chapter 2 of Part I propose new estimators for classic demand estimation problems. Our proposal outperforms existing solutions when the number of features is large and sales data exhibit a long-tail pattern, that is, when a significant portion of products have zero or very small sales. High-dimensional features and long-tail pattern are increasingly common nowadays, especially with the rise of online platforms where millions of products are being offered at the same time, and more and more feature data are collected. The work in Part I is conducted under the supervision of Prof. Fanyin Zheng.

Customer Preference and Station Network in the London Bike Share System Bike share systems have rapidly expanded across major cities. The type of bike share systems we study is the dock station system where operating companies

install stations at certain locations and customers can pick up a bike at any station with available bikes, and drop it off at any station with empty docks. Studies have found that bike sharing systems have several benefits, such as public health and environmental improvements. From the managing company or the local government's perspective, however, a number of challenges have arisen. One of the key questions is the design and expansion of the docking station network. We could view this problem as a long-term capacity planning one because once installed stations cannot be easily moved. Chapter 1 of this thesis focuses on station network design and expansion in the particular example of the London bike share system. It is an important question because if the operating managers or local government want to promote use and adoption of the system, understanding how the station network affects customer demand and where to expand the network and install new stations is crucial.

Because bike sharing programs are a relatively new phenomenon, few studies are available, and in practice, policies that managing companies and local governments have largely relied on are very ad-hoc. For example, London has been implementing a 300-meter density rule, which basically requires one docking station being installed roughly every 300 meters. However, little evidence suggests this policy is optimal or reasonably good. In Chapter 1, we take an empirical approach to study the problem, using data from the London bike share system. We estimate a structural demand model for customers and utilize the model estimates to analyze different counterfactual expansion strategies to provide guidance on the network design and expansion. We provide novel solutions to endogenous choice set issues and computation bottlenecks encountered during model estimation. By endogenous choice set problem, we mean that whether a product is in stock or not is not random but correlated with how popular or attractive that product is. Failure to correct for this endogeneity leads to biased estimates and sub-optimal policy recommendations. We propose an instrumental variable (IV) approach to solve this problem in our empirical exercise,

which could potentially be used in other contexts with network products.

Machine Learning in Demand Estimation with Long-Tail Data The state-of-the-art structural demand estimation method has been around for 20 years and has been the workhorse model for demand estimation in differentiated products markets in economics. However it has severe drawbacks when dealing with sales data with long tails, where a significant portion of products have zero or very few sales. In particular, the most popular demand estimation method proposed by Berry et al. (1995) (BLP) does not allow products or services with zero sales to be included in the estimation. Standard practice in empirical applications is to either throw away low-share products or aggregate products into larger categories. However, neither solution is ideal in that they will bias estimates, and, more importantly, the biased estimates will lead to sub-optimal policy recommendations and managerial decisions.

In Chapter 2, we propose a new two-stage estimator that corrects for the bias caused by long-tail data in BLP estimates with the help of deep learning algorithms. In the first stage, we propose a novel machine learning procedure, adapted from deep learning in computer vision and natural language processing, to predict market shares for all products including the zero-sales ones. The challenge here is that a product's market share not only depends on its own characteristics, but also on other competing products' characteristics. A naive application of off-the-shelf predictors is not likely to work well in predicting market shares. Our proposed deep learning model in the first stage solves this issue by incorporating the structure of the consumer choice problem. In the second stage, we utilize the predicted shares from the first stage to correct for the bias in the estimation, by essentially re-weighting different observations. We find our proposal has better performance than existing solutions in simulation experiments, especially when the number of features is large. Our method of correcting for the bias is analogous to the bias correction often used in

the classic literature of estimating treatment effects. We use the predicted shares in the first stage to construct weights to correct for the bias in the second stage of the estimation. This approach is analogous to using an estimated propensity score to construct weights or matches to correct for bias in treatment effects estimation.

Part II Financial Engineering and Machine Learning Models in Asset Pricing

Part II of this thesis studies applications of financial engineering models and machine learning techniques to empirical asset pricing. The field of empirical asset pricing has been traditionally dominated by approaches and ideas from financial economics. Traditional financial engineering models tackle issues such as derivatives pricing and risk management. In Chapter 3, we make connections between the two areas and study a stock trading strategy based on signals from a particular type of option pricing models in financial engineering, namely, rough volatility models. Thanks to the breakthrough of computing power and machine learning algorithms, a large amount of interests and opportunities have arisen in applying ideas from machine learning to finance in both industry practice and academic research. In Chapter 4, we study factor models in empirical asset pricing through the lens of heterogeneous treatment effects (HTE), an interdisciplinary subject involving causal inference and machine learning. The work in Part II was conducted under the supervision of Prof. Paul Glasserman.

Buy Rough, Sell Smooth A recent line of research has found evidence that stock price volatility is *rough*, in the sense that a *fractional* Brownian motion (fBM), instead of an ordinary Brownian motion, drives the dynamics of volatility, which could yield rougher volatility evolution paths than models based on ordinary Brownian mo-

tion. Two primary pieces of empirical evidence support rough volatility: the time series behavior of realized volatility, and an empirical regularity of option-implied volatility at short maturities that turns out to be well explained by roughness. In Chapter 3, we connect rough volatility models to empirical asset pricing and study the question of whether the stock market cares about and responds to roughness in volatility. In particular, we focus on a trading strategy of buying stocks with the roughest volatility and shorting stocks with the smoothest volatility using two different measures of roughness, namely, realized and implied roughness. We find the proposed buy-rough-sell-smooth strategy based on implied roughness generates significantly positive annual returns of 6% or more, which standard factors cannot explain. We further investigate why the strategy works. We conclude that it is not coming from roughness being able to predict future earnings surprises. Rather, we attribute the profitability to near-term idiosyncratic event risk, based on studying the impact of corporate earnings announcements on the strategy performance.

Heterogeneous Treatment Effects in Asset Pricing The intersection of causal inference, econometrics, and machine learning has been an exciting research area lately, thanks to an increasing amount of available data and the advances in black-box predictive algorithms. The key issue here is that we want to utilize machine learning algorithms to relax parametric assumptions in part of our model without sacrificing the ability to conduct valid statistical inference on the causal parameters we care about. Chapter 2 falls into this literature as well in the sense that we utilize machine learning algorithms to solve one challenge in the identification of key causal parameters of interests. In Chapter 4, we focus on another problem tackled by researchers in causal inference and machine learning, HTE estimation, and apply one particular HTE estimator proposed by [73] to empirical asset pricing. Our work in this chapter is a novel empirical study applying HTE methods in finance, which

complements the theoretical studies that are mainly aimed at domains such as personalized medicine and program evaluations. We bring newly developed techniques from causal inference and machine learning to provide an interesting new perspective to empirical asset pricing. Our angle can be viewed as a middle ground between the traditional linear approach and simple applications of black-box machine learning algorithms to predicting stock returns.

In particular, in Chapter 4, we cast the traditional linear regression model that studies how firm characteristics predict future stock returns in the framework of causal inference and treatment effect estimation. The analogy is that we select one focal firm characteristic as the treatment variable, and its effect or impact on future stock returns is the treatment effect we are going after. With the help of HTE estimation, we relax the homogeneous treatment effect assumption that the treatment characteristic affects future stock returns exactly the same way for all stocks, which is implicitly assumed by the traditional regression approach. Specifically, we use the R-learner algorithm coupled with the state-of-the-art machine learning algorithm gradient boosted trees to estimate HTE for each stock. Based on our fitted HTE, we characterize two subsets of stocks: we call one subset the characteristic responders, which have the largest treatment effect, and we call the other subset the characteristic traps, which consist of stocks whose future returns are the least responsive to the treatment characteristic. Evaluating how accurate our fitted HTE is and whether or not the resulting characteristic responders and traps are useful in practice is difficult because we can never observe the true treatment effects in the data. To overcome this challenge, we design a long-short of long-short test to evaluate the effectiveness of our estimates. In our empirical study, we focus on the most standard characteristics value, size, and momentum as treatment variables. We set one characteristic at a time as the treatment variable and use the long-short of long-short test to confirm that the effect of treatment variable on future stock returns in the characteristic

responders subset are indeed larger than the effect in characteristic traps during our test period. The differences in the treatment effects between the two subsets are statistically and economically significant. Unlike black-box predictors, we are able to provide some nice interpretations of the HTE estimates and generate insights on return predictability of firm characteristics.

Part I

Demand Estimation: Methodology and Applications

Chapter 1

Customer Preference and Station Network in the London Bike Share System

1.1 Introduction

Bike share systems have rapidly expanded across major cities of the world. Cities such as New York, London, and Paris have all introduced this new shared transport service in the past few years. There are many benefits associated with the bike share system. Studies have found that bike sharing systems have positive effects on public health by creating a large cycling population [36, 88]. Researchers have also shown that there are significant environmental gains from introducing the systems [36]. From the managing company or the local government's perspective, however, there are a lot of challenges and room for improvement in managing the bike share systems [68]. One of the main challenges is the design and expansion of the docking station network [71, 86]. For example, if the manager's goal is to maximize bike usage on the network and capture as well as possible the benefits of the new transport service, it is important to know where to expand the network and where to install new stations.

Since bike sharing programs are a relatively recent phenomenon, few studies have focused on the network design and expansion of the stations. In practice, managing companies and local governments have largely relied on ad hoc rules and policies. For example, the city of London has been implementing a 300-meter density rule, which states that two neighboring docking stations should be, at most, 300 meters away from each other across the city [86]. Given the potential customer demand variation

across the city, it is not clear whether the uniform density rule is optimal. In New York city, the Citi bike network focused primarily on building a high-density network in the downtown area in the first few years since its introduction. After many major expansions of the network, the system still covers only the downtown and midtown areas in Manhattan, and very few places in uptown Manhattan and Brooklyn.

In this chapter, we provide guidance on the network design and expansion question of the bike share system using the example of the system in London. The analysis is conducted in two steps. First, we estimate customer demand on the station network using system usage data. We take the structural estimation approach for the following reason. Since the objective is to provide network design and expansion recommendations to managers, we need a model of customer behavior to recover the preference parameters in order to evaluate different counterfactual expansion strategies. Without a structural model, simple regression analyses do not recover customer preference parameters, and therefore we would not be able to use the estimated parameters to evaluate counterfactural expansion and design experiments of the network.

To model customer demand, the natural way to go is to treat the docking stations as products and to assume that customers choose the stations based on the utility they gain from using the bikes at those stations [63, 81]. This approach is problematic, however, because it leaves out the important network structure between stations. Customers will choose the origination station only if the destination station is also attractive. In other words, customers are choosing the route or the link on the network between stations instead of the individual stations. For instance, if there is only one station in the network, demand is going to be very low because customers will be able to make only trips originating and ending at the same station. When a node has a lot of links on the network, and the links are attractive routes to the customer, however, the demand at that station is going to be substantially higher. This is what we refer to as network effects in this chapter. Those effects can be captured only if

we take the entire network into account, instead of treating stations as independent, when modeling the choices of the customers.

In light of the importance of the network effects in the current setting, we estimate customer demand for each origination and destination station pair instead of for individual stations. In other words, we study customers' preferences on the routes generated by a station network. This creates two challenges in the estimation. The first one is the endogeneity problem of the choice set. The choice set of the customer is endogenous because whether a station is in the choice set depends on whether it has bikes or docks available. The availability of bikes and docks at a station can be correlated with unobserved characteristics of the station, which then give rise to the endogeneity issue. The problem is particularly difficult to solve in a network setting in which the conventional instrumental variable approach does not apply. Using reduced-form regression evidence, we first show that this problem leads to biased estimates and unreasonable policy recommendations. Then, we propose a novel instrumental variable solution to this problem and show that our solution removes the bias in the parameter estimates and provides reasonable policy recommendations. The second challenge in estimation is the computational difficulty given the extremely high number of routes in the network. We reduce the computational burden by dividing the coverage area into blocks and model demand on the block level. We argue that this approximation is reasonable given the objective of evaluating long-term expansion strategies.

In the second step, we use the estimated model and customer preference parameters to provide insights into the design and expansion of the docking station network. This is done through three counterfactual analyses. In the first counterfactual, we evaluate a particular expansion proposal by the local government and predict network usage increase after the expansion. It demonstrates the practical insight that our study can provide to managers of such bike share systems. In the second and

third counterfactual, we generalize the insight and highlight the importance of network effects in studying customer demand on bike share systems. We compute the effect of adding stations and adding bikes or docks to different parts of the network in the two counterfactuals, respectively. First, we show that increasing density in the city center leads to usage increase ten times as high as that from increasing the scope of the network. This shows that the 300m density rule in the London system is far from ideal. Second, we decompose the variation in usage increase at different locations of the network. We identify two types of network effects and show that network effects play a key role in understanding demand variation across the network and evaluating the expansion strategies.

Despite the importance of the design and expansion of the station network, there are very few empirical studies on this topic. The closest to our paper is [63], who study the demand of the bike sharing system at the station level in Paris. The main difference is that our paper focuses on route-level demand, which brings the network effect into the analysis. This allows us to evaluate the changes to the network of stations by usage throughout the entire system instead of focusing on usage at individual stations. Our paper is also closely related to several studies analyzing the local demand and rebalancing of bikes in the Citi bike system in New York [74, 81]. The main difference is that we model customer behavior and recover preference parameters in the structural model, which allows us to compute counterfactual predictions and provide prescriptive recommendations to managers. Our work also contributes to the broader literature of ride sharing and car sharing, with a focus on the spatial network structure of customer demand [61].

There are many studies in the transportation literature that estimate origination-destination traffic demand [8, 15, 11, 85]. Some of the studies use similar discrete choice modeling tools to study customer behavior [8, 15]. However, those studies have mainly focused on real time traffic predictions, instead of recovering causal

customer preference parameters and applying the estimated model to provide long term policy recommendations. In addition, our model is much more flexible as we allow for unobserved route characteristics and customer heterogeneity, compared with the relatively simple models commonly used in this literature [8, 15].

The main estimation method we use in the analysis is based on the classic demand estimation framework introduced by [20] and the MPEC algorithm in [83]. The new method we develop to account for the endogeneity problem of the choice set relies on the network structure. The intuition of this method is related to that of the identification strategy in [23]. The method also contributes to the literature on estimating demand with endogenous stock-out products studied by [69] in operations management and by [33] in economics.

The main contribution of this chapter is threefold.

1. Practical guidance Our analysis provides important practical guidance to managers and the local government in evaluating network expansion strategies in London. With similar data from other cities, the framework can also be easily applied to other bike share systems to understand customer demand and evaluate expansion proposals.

2. General insights Our analysis highlights the importance of network effects in studying customer demand on a transportation network or products with a spatial aspect. We provide strong evidence to show that treating bike stations as individual products is far from sufficient. The structure of the network–i.e., where the connecting nodes are and what the weight is on each link–plays a significant role in determining the demand on the each node.

3. Methodology We illustrate the problem of endogenous choice set using regression analyses. We show that the endogeneity problem leads not only to biased estimates, but, more importantly, to unreasonable policy recommendations. We provide a novel instrumental variable approach to address the problem in a spatial network setting. The method can be easily applied to studying demand for other network or

spatial products.

1.2 Background and Data

1.2.1 Background

The bike share system in London, "Boris Bikes", was introduced in 2010. Like many bike share systems in major cities around the world, "Boris Bikes" went through a few major expansions after the initial launch.[1] There are calls and proposals to expand the station network further–for example, to central and south Islington and Hackney, the borough of Southwalk. "Boris Bikes" has been an important part of the local government's urban development program. In the 2016 mayoral election, both Labour and Conservative party candidates pledged to expand "Boris Bikes" and make London more bike-friendly.[2]

1.2.2 Data

The data we use in the analysis consist of four parts.

1.2.2.1 Stations and trips

First, we have data on the stations of the London bike share system and the trips made by customers on the system in 2014. There are 724 bike stations on the network. For each station, we observe the exact location (longitude and latitude coordinates) and the size of the station (total number of docks). The system covers an area of about 8 kilometers by 16 kilometers at the center of the city.[3] For the trip data, we

[1]Source: `https://tfl.gov.uk/info-for/media/press-releases/2013/december/mayor-launches-huge-expansion-of-flagship-barclays-cycle-hire-scheme`

[2]Source: `https://www.theguardian.com/environment/bike-blog/2016/apr/14/london-mayoral-election-qa-on-cycling-policy-with-the-main-candidates`

[3]We provide a map of the Greater London administrative area with the system coverage area in the Online Appendix.

observe the location (longitude and latitude coordinates) of the origination station and the destination station of each trip, as well as the starting and ending time of the trip. There are about nine million trips in total. We find that there is huge variation in the number of trips across routes, where a route is defined as a directional link between two stations. Some routes have thousands of trips, while others have only one trip.[4] It appears that, similar to retail sales data, the distribution of usage has a long tail, where many routes have very low usage, while a few routes have very high usage. There are also many routes with zero usage throughout the year. In the subsequent analysis, we follow the standard practice of including only routes with positive usage. In the structural estimation, in particular, we aggregate routes across nearby stations, which alleviates the long tail issue.

1.2.2.2 Availability Snapshots

Second, we have station snapshot data for each station every five minutes throughout 2014. The snapshots contain information about the number of bikes and docks available at each station. To demonstrate the variation in the snapshot data, we compute the following availability measures during the busiest time windows for the system: weekday morning rush hour (5:30am-9:30am) and evening rush hour (4:00pm-8:00pm). We define the bike availability measure as the percentage of time a station has at least five bikes available–i.e.

$$bike\_avail = \frac{\text{num of 5-min intervals with } >= 5 \text{ bikes}}{\text{Total num of intervals}}.$$

Similarly, we define the dock availability measure as

$$dock\_avail = \frac{\text{num of 5-min intervals with } >= 5 \text{ docks}}{\text{Total num of intervals}}.$$

Similar to [63], we use five as a threshold to allow for the possibility of broken bikes and docks. We present the summary statistics in Table 1.1. For all four time windows

[4]See the Online Appendix for detailed summary statistics.

and two availability measures, there is substantial variation in bike and dock availability across stations. This shows that availability conditions vary a lot at different locations, which is important to take into account in the subsequent analysis.

Time Window	Availability	Mean	sd	Min	25%	median	75%	Max
morning rush hour	bike availability	0.73	0.21	0.10	0.58	0.77	0.90	1.00
evening rush hour	bike availability	0.74	0.16	0.19	0.63	0.76	0.88	1.00
morning rush hour	dock availability	0.80	0.14	0.27	0.71	0.83	0.91	1.00
evening rush hour	dock availability	0.84	0.13	0.28	0.77	0.87	0.94	1.00

Table 1.1: Summary Statistics for Long-term Availability Measure

1.2.2.3 Google Data

We collect four sets of data from Google API.

This first set is the biking distance data. We use ij to denote a route from station i to station j. For each route ij, we collect the biking distance data from Google Maps using the best biking route from location i to j according to Google. We refer to this distance as the biking distance for the rest of the chapter.

Second, in addition to the biking distance, we also collect the data on the change of elevation along each route ij. A standard measure for the degree of inclination in transportation is the slope grade. Google Maps data allows us to divide each route ij into many short segments, and compute the grade for each segment. We define two features for each route: 1) $AvgAscendGrade_{ij}$: the average grade among the ascending segments of route ij, and 2) $AscendPercentage_{ij}$: the proportion of ascending segments on routes ij. We provide calculation details in Appendix A.1.

Third, we collect the data on other travel options in the city, in order to understand the outside options the customers face when choosing whether to use the bike share systems. We collect the data on two features, the distance and the travel time, for each of the two alternative travel options, driving and public transportation. We find

that the travel time is highly correlated with the distance in the data, so we only use the distance feature in the analysis from now on.

Fourth, we collect Google Places data in London. This dataset provides the longitude and latitude coordinates of 97 categories of places that are identified on Google Maps, including subway stations, government office buildings, schools, restaurants, etc.[5] We group 97 Google Place categories into ten general categories, food, health, religion, entertainment, stores, government offices, transportation, education, finance and others, and use these general categories in our analysis. Detailed definition of the groups and the summary statistics are presented in the Online Appendix.

We divide the coverage area of the bike share system into 200 by 200 meter squares and count the number of each Google place category in each square. Since many categories have zero count for more than 40% of the squares, it can be more informative to use the total counts instead of the counts for each category in the analysis. In addition to this observation, we also calculate the correlation between each pair of category counts, including the total counts. [6] We find that the average pairwise category correlation among the ten categories of Google places is 0.29, and the total Google place count is highly correlated with any of the ten category counts, with an average correlation of 0.56. This observation implies that using total Google place count in the analysis can be a good approximation of using all ten category counts.

1.2.2.4 Census

Finally, we use demographic data from the 2011 Census, which is more complete than the data from more recent years. The data includes population, income, age, gender,

[5]Both the data content and the category used in Google Place data change over time. Unfortunately, historical data are not available. The set of the data we use was scraped in February 2017.

[6]We present the correlation matrix in the Online Appendix.

and related demographic information, measured for each Lower Super Output Area (LSOA) in London.[7] LSOA is the smallest census unit with accurate data, and there are, in total, 4835 LSOAs in the London city. Our 724 stations in the bike share system cover 430 LSOAs located in the center of the city.[8]

1.2.3 Preliminary evidence

To motivate our analysis, we present several pieces of preliminary evidence about the usage of the system before going into details about the model and the estimation. The evidence comes directly from the data and therefore is model-free.

First, we find that 75% of the total usage of the London bike share system is on weekdays. Since weekday and weekend usage patterns can be very different, and from the local managing company's perspective, the local population's weekday usage matters more, we focus on weekday usage for the analysis. Within the weekday usage data, we find that 60% of the total trips occur during the morning rush hour (5:30am to 9:30am) and the evening rush hour (4pm-8pm) on weekdays. Moreover, the local government clearly makes the commuters the main beneficiaries when discussing plans for expanding the network [54]. For these reasons, we restrict our analysis to rush hour usage.

Next, we provide some evidence on the spatial pattern of the system and its usage. Figure 1.1 shows the usage pattern at 9:30am on a weekday morning in London. Each circle represents a docking station. The shade of the circle corresponds to the number of bikes available divided by the total number of docks at the station, ranging from black, which indicates that the station is full of bikes, to white, which indicates that the station is completely out of bikes.

[7]The data were downloaded from the local government's website, `https://data.london.gov.uk/dataset/lsoa-atlas`

[8]We provide the summary statistics for both the 430 covered LSOAs and all 4835 LSOAs in the Online Appendix.

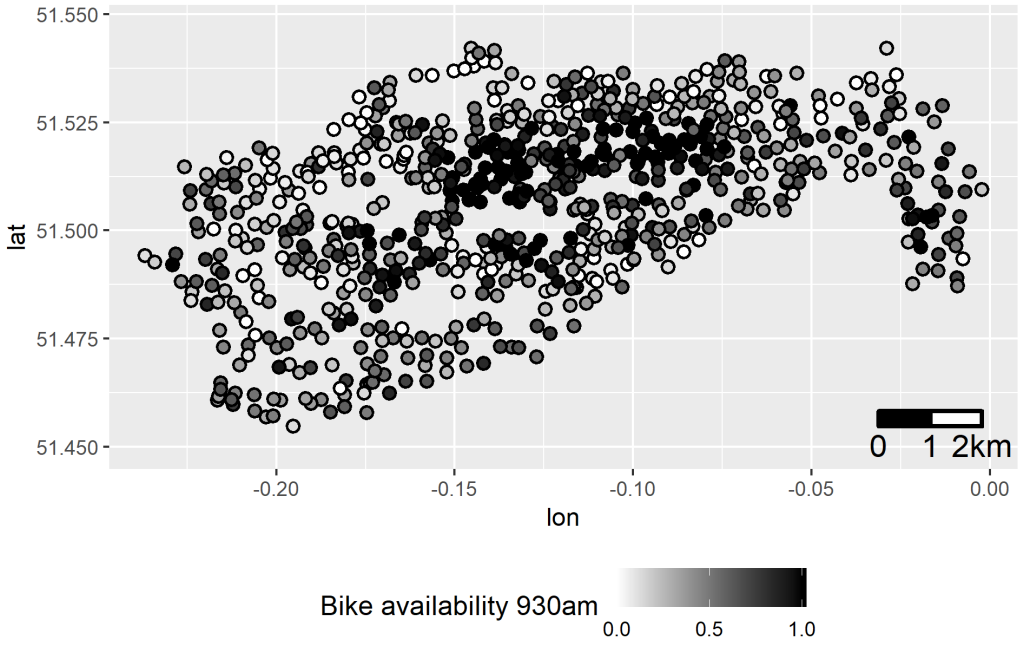

Figure 1.1: End of morning rush hour system status

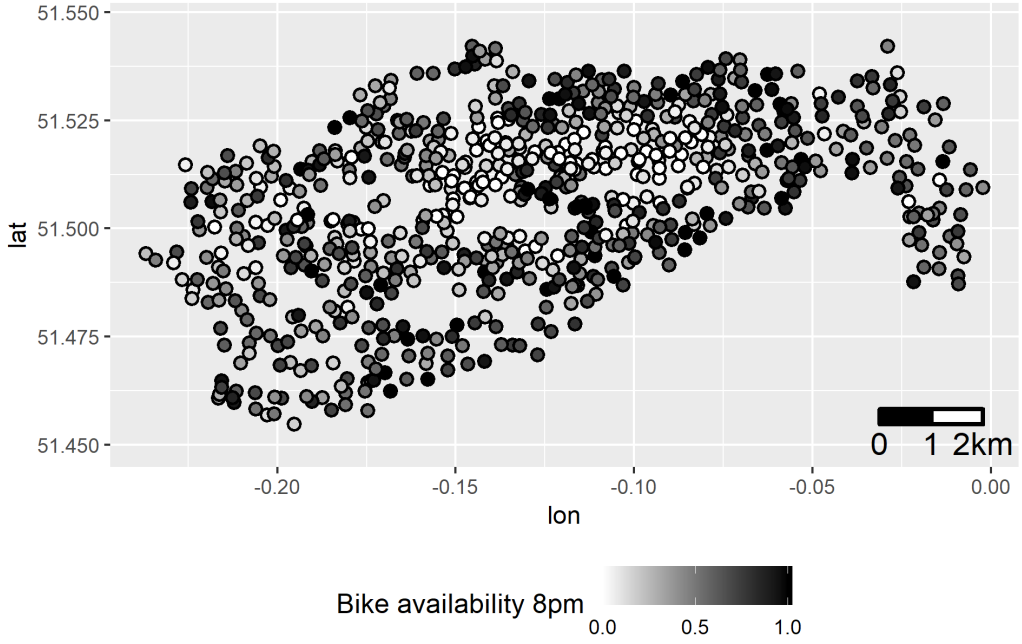

Figure 1.2: End of evening rush hour system status

Figure 1.1 shows that, towards the end of the morning rush hour, many stations around the city center are full, and many of the stations in the more residential areas in the outer part of central London are empty. This implies that people pick up bikes from where they live in the morning, commute to work, and return bikes near where they work in the center of the city. The pattern of traffic during morning rush hour is generally from the outer part of central London to the very center of the city. We see the opposite direction in the usage pattern during evening rush hour, captured in Figure 1.2. This is a snapshot of the system around 8pm on the same day. Figure 1.2 shows that a lot of bikes have been picked up in the very center of the city and returned to the more residential areas around the outer part of central London after the evening rush hour commute. This directional pattern of traffic on the network during rush hours will also show up in our estimation results presented later. More importantly, we will rely on the directional pattern to construct our instruments in order to identify the preference parameters in the structural model.

In Figures 1.1 and 1.2, another pattern in the data concerns the density of bike stations. The station density is slightly higher in the very center of the city but is more or less uniform otherwise. This is partly due to the 300m rule between stations imposed by the local transportation department, Transport For London (TFL). We will discuss whether the policy is reasonable, as well as its implications for how to expand the network when discussing the counterfactual analyses.

The last piece of evidence concerns the relationship between usage and one key route characteristic that we find in the data: route distance. We plot the average usage per route against route distance in Figure 1.3, which shows that, at first, usage increases very fast as distance increases. This is because as distance increases, more potential customers would prefer biking to walking. The peak usage occurs around 1.5 kilometers (km). As distance increases further, usage decreases quickly, as more potential customers would prefer other forms of transportation to biking. As shown

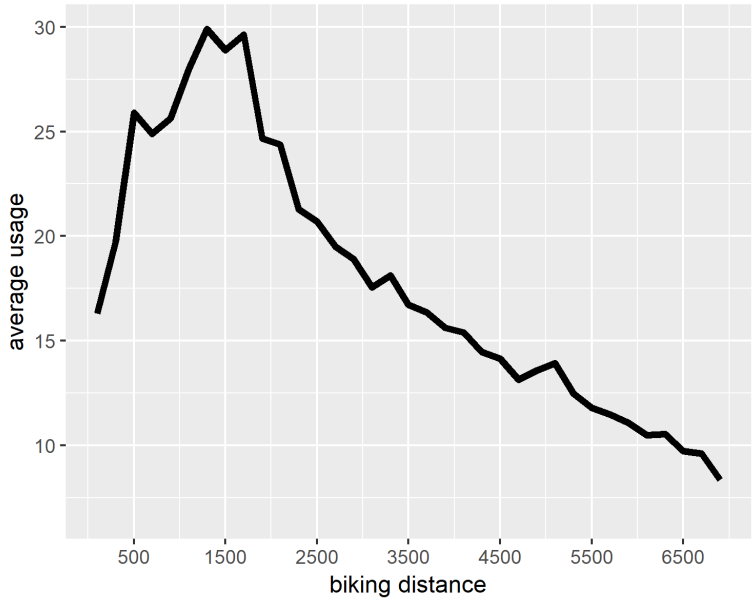

Figure 1.3: Average usage per route vs. route distance
Usage is measured in number of trips per route, and route distance is measured in meters. Peak usage around 1km.

in the estimation results in Section 1.4, route distance is a key characteristic that determines usage in the bike share system. The increasing and then decreasing usage suggests a non-linear preference pattern for route distance, which will prove to have important implications for network design and expansion. We will revisit this point in the counterfactual discussion.

The above three pieces of model-free evidence provide guidance to the modeling choices in our structural estimation analysis. Note that this evidence alone is not sufficient to provide prescriptive policy recommendations we are interested in. They are merely correlation patterns in the data. To understand customer preferences and therefore predict their choices under different counterfactual scenarios where the network is expanded, we need a choice model and an estimation procedure to recover the preference parameters when describing the model.

1.3 Choice model

In this section, we present the structural choice model. We do not argue that structural estimation is the only approach to understanding customer demand. But the reason that a structural model is necessary in our setting is as follows. The goal of the analysis is not just to understand customer demand, but, more importantly, to provide the managing company and local government with prescriptive policy recommendations in terms of station network design and expansion. Therefore, we would like to provide out-of-sample predictions that might be far from the observed data. To achieve this goal, we need to rely on a behavioral model that describes how potential customers make commuting decisions, to recover the preference parameters in the model, and to use the model to compute counterfactual predictions. Without a structural model, the parameters estimated from reduced-form regressions are not interpretable and can not help us make counterfactual predictions.

The model consists of two parts. The first part is a classic multinomial logit model that describes how customers choose between biking on a set of routes and using an outside commuting option. The second part consists of feeding the derived choice probabilities from the multinomial logit model into a density model that captures the differences in the number of potential customers in different geographic areas. We allow all parameters in the model to have different values for customer demand during the morning rush hour (MR) and the evening rush hour (ER). But for simplicity, we do not carry the MR and ER subscripts in the model.

We now discuss the model in more detail. Let U_{ijkl} be the utility that a customer who wishes to travel from location k to location l gets from using the bike share system to cycle from station i to station j, if bikes are available at station i, and docks are available at station j. Then, we have

$$U_{ijkl} = X'_{ijkl}\beta + \xi_{ij} + \varepsilon_{ijkl} \qquad \forall ij \in C_{kl} \tag{1.1}$$

X_{ijkl} is a vector of characteristics of route ij for customer traveling from k to l. It includes three types of characteristics in addition to a constant. The first type includes two terms: $\log d(i,j)$, the logarithm of the biking distance between stations i and j, and $\max\{\log d(i,j) - \log b, 0\}$, which captures the potentially non-linear distance preferences for long distance routes. We define long distance routes as those longer than 1.5 km–i.e. $b = 1.5$. This is natural in light of the fact that 1.5 km is the peak point in Figure 1.3. We provide robustness checks for other values of b in the Online Appendix. The second type of covariates are the two elevation features of route ij, $AvgAscendGrade_{ij}$ and $AscendPercentage_{ij}$ as defined in Section 1.2.2.3. They capture how physically demanding it is to bike from i to j. Third, X_{ijkl} also includes two walking distance variables: $d(k,i)$, the walking distance between the customer's origination location k and the starting station i, and, similarly, $d(j,l)$, the walking distance between the ending station j to the customer's destination location l. Although walking distance is not the focus of our study, we include them since other studies in the literature show that it is an important factor for demand [63]. Notice that unlike [63] who study station level demand and, therefore, only have the walking distance between the origination location of the customer and the starting station in their model, we include walking distance on both ends of the route.

ξ_{ij} are unobserved route characteristics, which can include whether route ij is bike-friendly, whether there is a bike lane, or other features which make ij more or less attractive. ε_{ijkl} are idiosyncratic error terms that are independent and identically distributed and follow extreme value distribution.

Since we differentiate customers only by the origination location k and the destination location l, we index customers by kl. Customer kl chooses from a set of possible routes ij and an outside option to maximize her utility. The choice set of all biking routes for customer kl is denoted by C_{kl}, which includes all routes with origination station i within walking distance of k and destination station j within

walking distance of l–if i and j have bikes and docks available. Therefore, the choice set is kl-specific. We specify C_{kl} rigorously in Section 1.4. We define the utility of customer kl's outside option as

$$U_{0kl} = X'_{0kl}\theta + \varepsilon_{0kl}, \tag{1.2}$$

where X_{0kl} includes the characteristics of customer kl's alternative transportation options. In the main specification, we include the distance of driving and the distance of public transportation from Google Maps data, as described in Section 1.2.2.3. The choice probability of kl choosing $ij \in C_{kl}$ is then

$$P_{ijkl}(X, \beta, \theta) = \frac{exp(X'_{ijkl}\beta + \xi_{ij})}{exp(X'_{0kl}\theta) + \sum_{i'j' \in C_{kl}} exp(X'_{i'j'kl}\beta + \xi_{i'j'})}. \tag{1.3}$$

For $ij \notin C_{kl}$, we have $P_{ijkl} = 0$. Let q_{ij} be the total number of trips on route ij predicted by the model. Then

$$q_{ij}(X, W, \alpha, \beta, \theta) = \int P_{ijkl}(X, \beta, \theta) dD(W_{kl}, \alpha), \tag{1.4}$$

where $D(W_{kl}, \alpha)$ is a density function that measures the number of potential customers traveling from location k to location l. It is a function of W_{kl}, the characteristics of the commuter origination and destination pair. In W_{kl}, we include the population density and the number of Google place counts at both k and l. Given that the objective of our analysis is to understand long term average customer demand and to evaluate different network expansion strategies which mainly involves comparing usage levels across locations, we do not model the high-frequency variation of usage over time.[9] Instead, we use Equation (1.4) to describe the average usage during the morning and evening rush hours, allowing all the parameter values to be different for the two rush hour time windows. We do not model demand variations

[9]Another reason for not using the temporal variation in the data is that we do not have exogenous covariates that vary over time, which introduces difficulties in estimating the parameters consistently. In fact, if we consider the morning rush hour as an example, let t be different days or weeks in a year, none of the features X in (1.3) and (1.4) depends on t.

on the same route across time or time dimension substitutions on the same route. Therefore, taking the morning rush hour as an example, q_{ij} in Equation (1.4) should be interpreted as the average daily usage of route ij during the morning rush hour, or equivalently, the total usage in a year during the morning rush hour. Similarly $D(W_{kl}, \alpha)$ measures the travel demand from location k to location l on an average day during the morning rush hour, or equivalently, the total travel demand from k to l throughout a year.

The core of the model is multinomial logit. It is known for implying unrealistic substitution patterns or the independence of irrelevant alternatives property. However, [20] illustrate that the multinomial logit model can allow for flexible substitution patterns if one introduces random coefficients. Here, we take an alternative approach and utilize the spatial aspect of the data to generate, through the density function, flexible substitution patterns. The density function plays the role of the random coefficient in the following sense. The options chosen by nearby customers with similar W_{kl} values are closer substitutes than otherwise. This breaks the independence of irrelevant alternatives property of the multinomial logit models. See [39] and [63] for similar approaches.

The unknown parameters that we need to estimate in this model are β, θ, and α in Equations (1.3) and (1.4). Note that all three parameters are vectors. We detail the estimation procedure in the following section. With the estimates of β, θ, and α, we will be able to predict the customer choices across the city and, therefore, the overall changes in usage on the network under counterfactual station network expansions.

1.4 Estimation

In this Section, we explain the estimation of the model. We start by introducing the endogenous choice set problem. Then, we explain our proposed solution to the

problem and provide evidence showing that, without properly accounting for the endogeneity problem, the parameter values can not be recovered without bias, thus leading to unreasonable expansion policy recommendations. Finally, we provide details on the rest of the estimation procedure and present the structural estimation results.

1.4.1 Endogeneity of choice set

In a classic discrete choice model, customers choose from a set of products or service options to maximize utility. The choice set is either perfectly observed or pre-set by the researcher. In our setting, the choice set is the set of possible routes ij from which customers can choose to commute from k to l. Compared with the classic setting, the complication here is that whether a particular route ij is in the choice set of a customer, C_{kl}, depends on whether there are bikes available at station i for the customer to pick up, and whether there are docks available at station j for her to return the bike to. As shown in Table 1.1, some stations frequently run out of bikes and docks, and there is significant variation across stations in terms of bike and dock availability. Of course, whether a station has bikes or docks available is *not* randomly assigned. It depends on the usage level at the station, or, in other words, how popular it is. Indeed, more popular origination stations are more likely to run out of bikes, and more popular destinations are more likely to run out of docks. Therefore, the choice set of the customer depends on how popular or preferable the routes are. This means that the choice set is endogenous to the choice behavior itself.

We now discuss the details of the endogeneity problem and how it biases the estimation results. In our discrete choice model, the utility of customer kl choosing to bike the route ij is given by Equation (1.1). Let X, W, and ξ be the matrices $\{X_{ijkl}\}_{ij=1,\ldots,N,kl=1,\ldots,M}$, $\{W_{kl}\}_{kl=1,\ldots,M}$, and $\{\xi_{ij}\}_{ij=1,\ldots,N}$, respectively. To obtain consistent estimators of α and β, one necessary condition is that ξ is mean zero, con-

ditioning on X and W–i.e., $\mathbb{E}[\xi|X,W] = 0$. Under this condition, one can construct moment conditions $\mathbb{E}[\xi \cdot X|X,W] = \mathbb{E}[\xi \cdot W|X,W] = 0$ [20, 19]. One important implication of this condition is that ξ_{ij} is uncorrelated with C_{kl} because, although we control for X, there are always route characteristics that are observable to the customer but not to the researcher. These unobserved characteristics, or preferability factors, are captured by the ξ_{ij} term in Equation (1.1). If the choice set C_{kl} is also affected by the unobserved popularity or preferability of the routes, then C_{kl} is correlated with ξ_{ij}. This is the sense in which there is an endogeneity problem with C_{kl}. The endogeneity issue will lead to bias in the estimated parameters.

We now provide the formal reasoning behind the endogeneity problem. To simplify the analysis, we first assume that the choice set C_{kl} of each customer kl is observed. We discuss later the practical implications and feasibility of this assumption. Rewriting Equation (1.3), we have

$$P_{ijkl}(X,\beta,\theta) = \frac{\mathbb{1}_{ij} \cdot \left[exp(X'_{ijkl}\beta + \xi_{ij})\right]}{exp(X'_{0kl}\theta) + \sum_{i'j'} \mathbb{1}_{i'j'} \cdot \left[exp(X'_{i'j'kl}\beta + \xi_{i'j'})\right]}, \tag{1.5}$$

where $\mathbb{1}_{ij}$ is a variable indicating whether route ij is in customer kl's choice set C_{kl}, i.e., whether station i has bikes and station j has docks available. Since $\mathbb{1}_{ij}$ is precisely observed in the data, we can treat it as a standard route characteristic. Equation (1.4) stays the same. Then, the moment condition we are actually using is

$$\mathbb{E}[\xi|X,W,\mathbb{1}] = 0, \tag{1.6}$$

where $\mathbb{1}$ is the vector of $\mathbb{1}_{ij}$, for all ij. Now, one could use Equation (1.6) to obtain estimates for α and β. However, the estimated coefficients will be biased since 1.6 is violated. The reason is that the unobserved product characteristics ξ_{ij} are correlated with the choice set indicator $\mathbb{1}_{ij}$–i.e., $\mathbb{E}[\xi_{ij} \cdot \mathbb{1}_{ij}] \neq 0$. For example, a particular route might be easier or harder to bike, depending on whether it is uphill or downhill, whether it has bike lanes, or whether the traffic along the route is more or less friendly to cyclists. These characteristics are unobserved to the researcher and, as discussed

above, captured by ξ_{ij}. If the preferences over such characteristics are correlated across customers, then a customer arriving later in the time window is more likely to face empty stations with no bikes available, or full stations with no docks available. Therefore, the choice set indicator $\mathbb{1}_{ij}$ is correlated with ξ_{ij}. Moreover, like many other bike share systems, the managing company restocks bikes throughout the day. The reallocation decisions are not random but are optimized by the managing company [74, 46]. The more popular origination and destination stations are also more likely to receive reallocated bikes. Thus, ξ_{ij} can also be correlated with $\mathbb{1}_{ij}$ through the supply of bikes. Therefore, the parameter estimates obtained using moment condition given in Equation (1.6) will be biased due to the endogeneity problem.[10]

It is important to realize that the endogeneity problem of the choice set is not specific to our estimation context. The same logic applies to all customer choice or demand estimation settings. Whenever a product is out of stock, which happens often in practice, it automatically drops out of the customer's choice set. Products are *not* randomly out of stock: more popular products are more likely to run out. As a result, the choice set of the customer is correlated with the unobserved preferability of the products. This leads to the endogeneity problem of the choice set. Since the problem arises whenever there is variability in the choice set, it is common in the context of demand estimation. However, the problem has not been studied extensively in either the operations management or the economics literature. To the authors' best knowledge, only two studies in the operations management literature explicitly discuss this issue. [69] study the impact of stock-outs on shampoo sales. [63] look at station-level demand in the bike share system in Paris and take into account the impact of availability. In the economics literature, [33] study the vending machine demand for candies, where the customer's choice set can be restricted by stock-out events.

[10] [66] study the endogeneity problem of ξ_{ij} in a different setting: ξ_{ij} is correlated with mixed marketing activities and thus is endogenous. They show that ignoring the endogeneity problem leads to substantial bias in the parameter estimates.

We discuss shortly why the methods used in those studies are not applicable in our setting.

The endogenous choice set problem is difficult to deal with directly for the following two reasons. First, choice sets are not modeled in classic discrete choice models. Choice models have focused on deriving choice probabilities for a collection of predetermined options. The set of options is typically not part of the likelihood function or moment conditions that researchers use to estimate the demand model. Second, it is difficult to apply classic instrumental variable approaches directly to the problem. Instrumental variables are classic tools for dealing with endogeneity problems in regression analysis. However, such approaches require that the endogenous variable enters the regression equation in a linear fashion. In our context, as shown in Equation (1.1), the choice set is nonlinear in the utility function. Therefore, it is infeasible to apply the instrumental variable approach directly.

Our proposed solution consists of two steps. In the first step, we convert the discrete choice set to a continuous average availability measure. Instead of specifying which routes are available in the choice set at different points in time, we allow all possible routes to be in the choice set, and let the long term average availability vary across routes. In other words, we compute, over the one year sample period, the fraction of time a station i has at least five bikes or docks available during the rush hour window, $bike\_avail_i$ and $dock\_avail_i$. We compute the two measures separately for the morning and evening rush hours.[11] By doing so, we ignore the variation of the choice set within the time window and across different days of the year. Effectively we only utilize the cross-sectional variation in the data across different locations and average out the temporal variation. We argue that this is a reasonable restriction for three reasons. First, our model is used to understand the observed and to predict the

[11] As discussed below, we estimate the morning and evening rush hour usage separately in two models, in light of the very different usage patterns. Therefore, the availability measures are also computed separately for morning and evening rush hours.

counterfactual long-term average usage, instead of usage from one hour to the next. Besides, it is natural to assume that the long-term average usage is the key measure which managing companies and local governments care most about when designing the bike share system. Second, as shown in the preliminary evidence section, the London bike share system is commuter-dominant. Therefore, it is reasonable to focus on average availability because most customers are repeat users who care more about average availability within the rush hour commuting time window.[12] Third, we conduct a variance decomposition exercise for the availability measures and find that the cross-sectional variation is the main source of variation in our data. After converting the discrete choice sets to the continuous average availability measures, we rewrite Equation (1.1) as:

$$U_{ijkl} = \beta_1 bike\_avail_i + \beta_2 dock\_avail_j + X'_{ijkl}\beta_3 + \xi_{ij} + \varepsilon_{ijkl}, \quad \forall ij \in C_{kl}, \qquad (1.7)$$

where the choice set C_{kl} consists of the feasible routes regardless of their availability. The long-term bike availability measure at the origination station i and the long-term dock availability measure at the destination station j are included in the utility function. As discussed before, these two availability characteristics are likely to be correlated with ξ_{ij}, and, therefore, are endogenous. However, Equation (1.7) shows that the endogeneity problem of choice set can become a familiar endogenous linear characteristic problem in the utility function. The problem is exactly the same as in most demand estimation, where the price of the product or service is the endogenous characteristic, and the usual instrumental variable approach applies.

The second step consists of finding valid instruments for the availability measures. In previous studies, [33] exploit a quasi-experiment in product restocking time to deal with the endogeneity problem. However, no quasi-experiment is available in our setting. [69] use supply-side instrumental variables to account for the endogene-

[12] On the other hand, if most usage is from casual users like tourists, short term availability would be more important.

ity of product stock-outs in supermarkets. However, they find that the estimation results are the same with or without the instrumental variables. [63] use similar types of instrumental variables as in [20], i.e., the characteristics of nearby stations. But these types of instruments have been shown to have weak identification power [6]. Moreover, when one treats stations as nodes on a network instead of as independent stations, the characteristics of nearby stations might become even weaker in their ability to identify the parameter of interest. To overcome these difficulties, we propose a novel set of instrumental variables by utilizing the station network structure. We show that properly accounting for the endogeneity of the choice set is key to recovering consistent estimates of the parameters and providing reasonable policy recommendations. We discuss the details of the instrumental variables in Section 1.4.2. As we will show later in the estimation results, without the instrumental variables, the bias can be substantial, and, more importantly, it can lead to unreasonable policy recommendations.

1.4.2 Instrumental variables

In this section, we discuss the instrumental variables for the bike and dock availability measures. The main difference between our setting and a classic demand estimation setting is that our data are generated from a station network instead of from independent station observations. When the data is generated from a network, the challenge in finding valid instrumental variables is that the exogenous covariates observed in the data are likely to be correlated spatially and throughout the network. The correlations, therefore, might lead to violations of the exclusion restriction for an instrumental variable to be valid. To solve this problem, we introduce a new method of constructing instrumental variables in a network setting. In particular, we utilize the station network structure to construct instrumental variables for the availability measures. We start the discussion by explaining why the exclusion restriction is vi-

olated if we apply the commonly used instrumental variables, and we then describe our proposed solution.

The commonly used instrumental variables for the endogenous characteristic (typically, price; in our case, availability measures) in demand estimation are the exogenous characteristics of the products offered in the same market [20]. These are the so-called BLP instruments [22]. The reasoning behind the validity of the BLP instruments is that, if a product is more isolated from the other products in the product characteristic space, then it has higher margin, which leads to exogenous variations in prices. The key point for these instruments to work is that the variations in product characteristics are exogenous. Applying the same idea to the availability at station i, for example, one would use station k's characteristics S_k as instruments, where i and k are close to each other. [63] use these types of instrumental variables in a similar setting. However, when stations are not independent but are nodes on a network, the BLP instruments may not be valid anymore. For example, if k and i are spatially close, S_k and S_i can be highly correlated. This implies that, conditional on S_i, there is little variation in S_k that we can utilize to identify the coefficient of the availability at i. It could also be the case that since k and i are close to each other, kj and ij are similar routes to customers, which implies that the availability at stations k and i can be very much correlated. Then, it is even more difficult to find exogenous variations to identify the availability coefficients. Both examples indicate that the BLP instruments are not appropriate in the current setting, where spatial proximity and network structure are prominent in the data.

Next, we explain how we deal with the challenge of correlations in network data and introduce the proposed instrumental variables. First, we define two stations as "connected" (or "connected on the network") if there are customers using the route between the two stations. To be precise, stations i and j are connected means that $q_{ij} + q_{ji} > 0$. We use the bike availability at station i as an example illustrating the

construction of valid instrumental variables for the availability measures. We want the instruments to be uncorrelated with ξ_{ij} but correlated with $bike\_avail_i$. For that we look for the average of some exogenous characteristic over station $h \in \mathcal{H}_1(ij)$, denoted as $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h$, where station $h \in \mathcal{H}_1(ij)$ must satisfy the following two conditions.

First, station h must be connected to station i. In other words, customers are biking between h and i. If h and i are connected, the exogenous characteristics of station h affect the usage on route hi or ih and, therefore, the bike availability at station i. The *relevance condition* then holds for the instrument S_h:

$$\mathbb{E}[bike\_avail_i \cdot S_h] \neq 0, \tag{1.8}$$

for all $h \in \mathcal{H}_1(ij)$.

Second, station h needs to be sufficiently far away from station j, so that almost no customer bikes the route hj or jh. To be precise, we require $d(h, j) \geq D$ where D is a threshold to be specified later. If no one bikes the route hj or jh, then route ij and hi or ih are *not* substitutable for any customer–i.e., route ij and route hi or ih are potential choices of customers who are interested in traveling from and to different locations. In other words, route ij and route hi or ih are products in different markets. As a result, the exogenous product characteristic S_h is unlikely to be correlated with the unobserved product characteristic ξ_{ij}. Formally, we have

$$\mathbb{E}[\xi_{ij}|S_h] = 0, \tag{1.9}$$

for all $h \in \mathcal{H}_1(ij)$. This is the *exclusion restriction*.

If Equations (1.8) and (1.9) are satisfied, then $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h$ is a valid instrument for the bike availability at i. Formally, we can write $\mathcal{H}_1(ij) := \{h \in H : d(h, j) \geq D, q_{ih} + q_{hi} > 0\}$, where H denotes the set of all stations; d denotes the distance function; q_{ij} denote the total trip count on route ij; and D denotes a required distance threshold from station h to station j. We postpone the discussion about

the choice of the threshold until the end of this subsection. Note that we require only the sum of q_{ih} and q_{hi} to be positive–i.e., we do not require both q_{ih} and q_{hi} to be positive. As long as there are connections between the two stations, Equation (1.8) is satisfied. We use the same criterion to find instruments for the dock availability at station j–i.e., $\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} S_h$ for some station characteristics S, where $\mathcal{H}_2(ij) := \{h \in H : d(h,i) \geq D, q_{hj} + q_{jh} > 0\}$.

Next, we discuss the particular choice of the exogenous characteristics S. In practice, we find that the average station characteristics of h, $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h$, can be very weakly correlated with the bike availability at station i. The main reason is that the traffic goes in and out of station i at the same time. As a result, the average correlation between $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h$ and the availability at i is close to zero. Similarly, the correlation between $\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} S_h$ and $dock\_avail_j$ is very small in absolute value. To solve this problem, we utilize the direction of the traffic during rush hours. As shown in Figures 1.1 and 1.2, the traffic during rush hours is by large unidirectional. During the morning rush hour, the traffic goes from the outer part of central London to the less residential city center. During the evening rush hour, the opposite pattern is observed. We rely on the directions of traffic and construct directional instrumental variables to ensure a stronger correlation between our instruments and the availability measures. Take route ij in the evening rush hour as an example. We use the interaction between the number of total Google place counts around station h and the population density around station i as one of our main instruments. During the evening rush hours, commuters mainly travel from their places of work to their homes. Then, the interaction between the total number of Google places around h and the population density around i is positively correlated with the bike availability at i during this time of the day. Similarly, the interaction between Google place counts around i and the population density around h is negatively correlated with the bike availability at i. In other words, for the bike availability at station i,

our main instrumental variables are $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} GooglePlaces_h * PopDensity_i$, and $\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} PopDensity_h * GooglePlaces_i$. We follow the same logic to construct the instruments for the dock availability at the ending station j: namely, $\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} PopDensity_h * GooglePlaces_j$ and $\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} GooglePlaces_h * PopDensity_j$. Recall that $\mathcal{H}_2(ij)$ is the set of stations that satisfy the "opposite" conditions as $\mathcal{H}_1(ij)$–i.e., they are far away from station i but connected to station j. Recall that earlier in this subsection we have established that 1) $\mathbb{E}[\xi_{ij}|S_i] = 0$ and $\mathbb{E}[\xi_{ij}|S_j] = 0$, by the exogeneity of S_i and S_j; and 2) $\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h] = 0$ and $\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} S_h] = 0$, where S is $GooglePlaces$ or $PopDensity$. With these two conditions, we have that the exclusive restriction holds for the interaction instruments:

$$\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} S_h S_i'] = 0,$$

$$\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_2(ij)|}\sum_{h\in\mathcal{H}_2(ij)} S_h S_j'] = 0,$$

where $S, S' \in \{GooglePlaces, PopDensity\}$, and $S \neq S'$. For example,

$$\mathbb{E}[\xi_{ij}|\frac{1}{|\mathcal{H}_1(ij)|}\sum_{h\in\mathcal{H}_1(ij)} GooglePlaces_h * PopDensity_i] = 0.$$

The interaction instruments also satisfy the relevance condition trivially. Therefore, the interaction terms are another set of valid instrumental variables for the availability measures.

To complete the definition of our instruments, we need to specify a radius R for calculating $GooglePlaces$ and $PopDensity$ of the instrument station h, and the focal stations i and j. To be precise, $PopDensity_h^R$ is the population density integrated over a disk with radius R centered at station h, and $GooglePlaces_h^R$ is the total number of Google places within R meters from station h. The superscript R denotes the radius of the characteristic calculation. To make set $\mathcal{H}_1(ij)$ and $\mathcal{H}_2(ij)$ explicitly depend on threshold D as well we add a superscript D to them.

We next discuss the choice of the two hyperparameters R and D. We set $D = 7000m$, which is the 94% quantile of the distribution of route distance for routes with positive trip count in the data. It is a reasonable choice for two reasons. First, it is long enough to ensure the exclusive restriction, and, therefore, the validity of the instrument. Moreover, it is not too long, which avoids situations in which $\mathcal{H}_1^D(ij)$ or $\mathcal{H}_2^D(ij)$ has none or very few stations for many routes ij. On the other hand, we do not have strong views about the exact values to use for R. Thus, we use a range of values. In our main results, we choose $R = 600, 800, 1000m$. Our main results are robust to perturbing the two hyperparameters D and R. The robustness check results are presented in the Online Appendix.

So far, our proposed set of instruments are:

$$\begin{aligned}
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} GooglePlaces_h^R * PopDensity_i^R,\\
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} PopDensity_h^R * GooglePlaces_i^R,\\
&\frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} PopDensity_h^R * GooglePlaces_j^R,\\
&\frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} GooglePlaces_h^R * PopDensity_j^R,\\
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} GooglePlaces_h^R,\\
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} PopDensity_h^R,\\
&\frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} GooglePlaces_h^R,\\
&\frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_2^D(ij)} PopDensity_h^R.
\end{aligned} \tag{1.10}$$

In addition to this set of instruments, we also utilize the elevation characteristics which are directional by themselves to construct additional instruments. For example, if the route hi is difficult to bike because it requires heavy climbing, the

bike availability at station i is more likely to be low. This is the relevance condition. The exclusive condition can be verified using the same argument as for the interaction instruments proposed before–i.e., if h is far away from j, the elevation characteristics of routes hi and ih should be uncorrelated with ξ_{ij}. As a result, $\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{hi}$, $\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{ih}$, and $\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AscendPercentage_{hi}$ can be used as instruments for the bike availability at the starting station i of route ij.[13] Similarly, we construct three additional instruments for the dock availability at ending station j. We have six additional instruments given by:

$$\begin{aligned}
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{hi},\\
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{ih},\\
&\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AscendPercentage_{hi},\\
&\frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{hj},\\
&\frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AvgAscendGrade_{jh},\\
&\frac{1}{|\mathcal{H}_2^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AscendPercentage_{hj}.
\end{aligned} \tag{1.11}$$

Equations (1.10) and (1.11) specify the complete set of instruments. We use the exact same set of instruments in Equations (1.10) and (1.11) for evening and morning rush hour. Although the validity of the instruments follow the same reasoning, the correlation between our instruments specified in Equation (1.10) and the availability measures is reversed. For example, $\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} GooglePlaces_h^R * PopDensity_i^R$

[13] We do not use $\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AscendPercentage_{ih}$, since it is the same as $1-\frac{1}{|\mathcal{H}_1^D(ij)|}\sum_{h\in\mathcal{H}_1^D(ij)} AscendPercentage_{hi}$.

is positively correlated with $bike\_avail_i$ in the evening rush hour, but is negatively correlated with $bike\_avail_i$ during the morning rush hour.

1.4.3 Digression: reduced-form regressions

To illustrate the choice set endogeneity problem with data and to show the validity of our instruments, we take a digression from the structural estimation in this Section, and provide model-free evidence using reduced form regressions. The reduced form regression results show that ignoring the endogeneity problem leads to biased estimates and unreasonable policy recommendations, and that our proposed instrumental variable approach solves these problems.

1.4.3.1 Regression equation

We run the regressions separately for the morning and evening rush hours. We describe the regression equation using evening rush hour[14]. The dependent variable, $log(\#trip_{ij})$, is the log of total trip count during the evening rush hour in 2014, starting from station i and ending at station j. Then,

$$log(\#trip_{ij}) = \psi_1 bike\_avail_i + \psi_2 dock\_avail_j + Z'_{ij}\phi + \epsilon_{ij}, \tag{1.12}$$

where $bike\_avail_i$ and $dock\_avail_j$ are the availability measures during evening rush hour defined as before. Z_{ij} contains exogenous characteristics for route ij. It has 1) the Google Place counts described in Section 1.2.2.3, within $R = 600, 800, 1000m$ around both i and j; 2) the population density within $R = 600, 800, 1000m$ from both i and j; 3) a piece-wise linear function of the biking distance capturing the first increasing and then decreasing pattern of the commuter preference for the biking distance: $\log d(i,j)$ and $\max\{\log d(i,j) - \log b, 0\}$, where the kink b is set at

[14]The analysis for the morning rush hour is the same. For notational simplicity, we do not include an ER superscript for each variable in the regression equation

1.5 km. We choose 1.5 km for two reasons. First, Figure 1.3 clearly indicates that the peak of average usage is around 1.5 km. Second, when we vary the kink point, $b = 1.5$ km explains the most variations in log trip counts. We present the results with different b values in the Online Appendix; 4) the elevation characteristics $AvgAscendGrade_{ij}$ and $AscendPercentage_{ij}$ as described in Section 1.2.2.3. In total, we have $1 + 10 \times 2 \times 3 + 1 \times 2 \times 3 + 2 + 2 = 71$ exogenous characteristics in Z_{ij}. We also have two endogenous covariates for which we need instrumental variables: bike availability and dock availability. Recall that, as shown in (1.10), for each value of $R = 600m, 800m, 1000m$, we have eight instrumental variables. Thus, we have $8 * 3 = 24$ instruments from (1.10). We also have six instruments from (1.11) which makes it 30 instrumental variables in total.

	Morning rush hour		Evening rush hour	
	OLS	IV	OLS	IV
O. Bike Avail.	0.05	0.69	0.13	1.37
%	(0.03)	(0.21)	(0.03)	(0.15)
D. Dock Avail.	0.50	1.47	-0.46	1.89
%	(0.04)	(0.16)	(0.03)	(0.13)
Dist. 1	0.12	0.10	0.14	0.15
In log km	(0.02)	(0.02)	(0.03)	(0.03)
Dist. 2	-0.37	-0.36	-0.42	-0.39
>1.5km, in log km	(0.03)	(0.03)	(0.03)	(0.03)
R^2	0.20	-	0.33	-
Cragg-Donald Statistic	-	83.64	-	99.37

Table 1.2: Reduced-form regression results

1.4.3.2 Regression results

In Table 1.2, we present the results of ordinary least squares (OLS) regressions without instrumenting availability measures and the results of the instrumental variable (IV) regressions using our proposed instruments. Table 1.2 includes the main coefficient

estimates and their standard errors in parentheses. For the OLS regressions, we also report the adjusted R^2. For the IV regressions, we estimate Equation (1.12) using the generalized method of moments (GMM). Moreover, to test the strength of the instruments, we perform the weak instrument test introduced by [82] and report the Cragg-Donald statistics. As discussed above, in our main specification, we set $R = 600, 800, 1000m$ for the station characteristics and $D = 7000m$ for the distance threshold. We check that our results are robust to perturbing values of R and D values and details are left in the Online Appendix.

Since the results for the morning and evening rush hour are similar, we discuss only the results for the evening rush hour, which are presented in columns 3 and 4 of Table 1.2. There are four main insights. First of all, in the OLS regression presented in column 3, the ending station dock availability coefficient is negative and significant. This means that commuters get higher utility if it is more difficult to find a dock available at the destination station. Moreover, the policy implication of the result is that, if the managing company or local government wanted to increase the usage of the system, they should make it harder for people to dock their bikes at the destination stations. Obviously, this cannot be the case. But if we acknowledge the endogeneity issue of the availability measures, this result is easy to explain: more popular destinations are more likely to run out of docks-hence the negative coefficient on dock availability. This is the classic "reverse causality" problem in econometrics. On the other hand, when we run the IV regression with our proposed instruments in column 4, the negative coefficient becomes positive, and it is statistically significant. The interpretation is that higher dock availability at destination stations is preferable to commuters. Therefore, the usage would be higher, if the dock availability improved at the destination stations.

Second, we compare the coefficient on the origination station bike availability in the OLS and IV results. Both of the estimated parameters are positive. One

might argue that, in this case, using the IV regression does not benefit the analysis substantially. However, the magnitudes of the two coefficients are very different. The estimated impact of having higher bike availability at the origination station on log trips in the IV regression is more than ten times as high as that in the OLS regression. This difference is expected since the "reverse causality" predicts a negative correlation between bike availability at i and usage on ij. The OLS result combines the "reverse causality" and the correct positive relationship, and, therefore, is smaller than the IV estimate. This comparison shows how important it is to take into account the endogeneity issue of the choice set, and how significant its implications are for policy recommendations derived from the estimation results.

Third, the distance coefficients do not change much between the OLS and IV results. The first distance parameter is positive, which captures the route usage increasing in route distance when the distance is below 1.5 km, as shown in Figure 1.3. The second distance parameter is negative and much bigger in magnitude than the first distance parameter. This is consistent with the pattern in Figure 1.3. The same pattern is also observed in the structural estimation result, which is detailed in Section 1.4.4.

Fourth, the Cragg-Donald statistics from both the morning and evening rush hour are very high. The critical value corresponding to our setting, in which there are 30 instruments and two endogenous covariates, is 20.86 (for the relative bias level of 0.05, see [82] for details of the calculation). Our statistics, 83.64 and 99.37, are more than four times bigger than the critical value. Therefore, we can strongly reject the weak instrument null hypothesis.

1.4.4 Structural estimation

In this section, we discuss the structural estimation procedure and present the results. We use the generalized method of moments with the proposed instruments to recover

the parameters in Equations (1.3) and (1.4). However, estimating the parameters in the structural model creates computational challenges. The methods introduced in [20] can accommodate non-linear parameters, but they are computationally feasible for only relatively low number of products or services in the choice set [83]. Since there are over 500,000 routes in the data, the estimation is computationally infeasible, even using the MPEC algorithm developed by [83].

To make the estimation feasible, we divide the coverage area into blocks and estimate the model on routes between station blocks instead of routes between stations. Similar methods have been used in the literature studying demand and supply of taxi cabs [24, 45] and demand predictions of the bike share system in New York [81]. In the block model, customers choose the commuting routes between station blocks instead of the specific route between stations. This reduces the computational burden substantially and makes the estimation possible. Of course, it imposes restrictions on the model and, therefore, on consumer behavior captured by the model. We argue that the modification is reasonable for the following reasons. First, it preserves the average substitution patterns across routes between different blocks. Since the counterfactual predictions we are interested are not the exact longitude and latitude of the location of the stations to be built, but which neighborhoods more stations should be added to or which new areas the network should expand into, understanding the average usage and substitution patterns across blocks is sufficient. Second, only three percent of the total trips in the data are between two stations within a block. As a result, we exclude very few data points by estimating the model at the block level. In other words, summarizing trips across blocks provides a good approximation of the general usage patterns in the network.

Next, we discuss the details of the block model. We divide the coverage area into uniform $1000m \times 1000m$ blocks. The stations within each block are treated as a single representative station. The location of this representative station is defined as the

center of the stations in that block and used as the location of the station block. We denote the starting station block by capital letter I and the ending station block by J. The availability measures of a station block, ba_I or da_J, are calculated as the average availability measures for all stations within that block. The route distance of IJ is defined as the average distance across all routes from any station in I to any station in J. The other covariates in the utility function are defined similarly. Compared with the route-level model, we include two additional covariates in the utility function, $\log SC_I$ and $\log SC_J$, which are the log total number of station counts in block I and J. These covariates capture the variation of the number of route options across station blocks.

The origination and destination locations, which commuters are interested in traveling from and to, are modeled as points of a grid. We divide the coverage area into $200m \times 200m$ squares and take the center point of each square as potential origination and destination locations of customers. In total there are 3263 such locations and therefore $3263^2 - 3263$ possible origination-destination pairs kl considered in our model. For commuters traveling from location k to location l, walking distance $d(k, I)$ is defined as the average distance between location k and all stations $i \in I$. $d(J, l)$ is similarly defined. We define C_{kl}, the choice set of commuters kl, as any routes starting from the four closest station blocks to k based on $d(k, I)$, to the four closest station blocks to l based on $d(J, l)$. There are thus 16 block-level routes in C_{kl} for customer kl. Since we do not observe the walking distance directly from the data, we also conduct the estimation using the same model without walking distance. Our conclusions do not change.[15]

The utility of commuter kl choosing route IJ is given by

$$U_{IJkl} = X'_{IJ}\beta + X'_{IJkl}\gamma + \xi_{IJ} + \varepsilon_{IJkl}, \quad \forall IJ \in C_{kl}. \tag{1.13}$$

X_{IJ} includes two sets of covariates. First, it includes a set of variables similar to the

[15]We provide the details of this robustness check in the Online Appendix.

ij-level covarites: an intercept, the two endogenous covariates ba_I, da_J, the distance characteristics $\log d(I,J)$ and $\max\{\log d(I,J) - \log b, 0\}$, and the elevation features $AvgAscendGrade_{IJ}$ and $AscendPercentage_{IJ}$. Second, it also includes the log station count for both the starting and the ending clusters, $\log SC_I$, $\log SC_J$. We define X_{IJ} separately from the rest of the covariates because they depends only on IJ but not kl. In other words, β is the vector of linear parameters. X_{IJkl} includes the walking distance variables $d(k,I)$ and $d(J,l)$. Since X_{IJkl} depends on both IJ and kl, γ is the vector of nonlinear parameters.

The choice probabilities are calculated similarly to Equation (1.3) in the ij-level model:

$$P_{IJkl}(X,\beta,\gamma,\theta) = \frac{exp(X'_{IJ}\beta + X'_{IJkl}\gamma + \xi_{IJ})}{exp(X'_{0kl}\theta) + \sum_{I'J' \in C_{kl}} exp(X'_{I'J'}\beta + X'_{I'J'kl}\gamma + \xi_{I'J'})}. \tag{1.14}$$

Similar to Equations (1.4), the model predicted total number of trips on route IJ is:

$$q_{IJ}(X,W,\alpha,\beta,\gamma,\theta) = \int P_{IJkl}(X,\beta,\gamma,\theta) dD(W_{kl},\alpha). \tag{1.15}$$

We follow the MPEC algorithm [83] and minimize the GMM loss function while matching the observed usage on the block-level routes with the predicted q_{IJ} to recover the parameters α, β, γ, and θ. We have 112 station blocks in total and thus 112^2 potential routes.[16] We use the same set of instruments for the availability measures as in the reduced-form regressions, specified by Equations (1.10) and (1.11).

We present the structural estimation results for the morning and evening rush hours in Table 1.3 and 1.4, respectively. We start the interpretations of the results by the linear utility parameters. First, in both the morning and evening rush hour results, the availability measures have the expected signs. The higher the average

[16]The dimension of the product space is reduced from around 761^2 in the route-level model to around 112^2 in the current model.

Non-linear para.		Linear para.	
O. Pop. Density	0.021	Intercept	-6.941
(Ppl per hec)	(0.015)		(1.211)
D. Pop. Density	-0.038	S. Station Count	0.899
	(0.019)	(In log)	(0.055)
O. Google Plc.	0.019	E. Station Count	0.756
(Num per 4 hec)	(0.016)		(0.031)
D. Google Plc.	0.286	S. Bike Avail.	1.730
	(0.107)	(%)	(0.263)
O. Walking Dist.	-3.145	E. Dock Avail.	1.790
(1km)	(0.764)		(0.206)
D. Walking Dist.	-0.927	Route Dist. 1	1.610
	(0.514)	(In log)	(0.267)
O.D. Driving Dist.	0.254	Route Dist. 2	-2.953
(1km)	(0.052)	(>1.5km, in log)	(0.362)
O.D. Transit Dist.	0.039	Avg. ascend grade	-5.549
(1km)	(0.056)	($\tan\alpha$)	(2.953)
		% segments ascending	-1.253
		(between 0 and 1)	(0.196)
Pseudo-R^2	0.667		

Table 1.3: Demand estimates: morning rush hour

availability, the higher utility the commuters gain from biking the route. Moreover, on average, the commuters seem to care more about the bike and dock availability during the morning rush hour than the evening rush hour. This is consistent with the fact that commuters have a tighter schedule in the morning than in the evening, and, therefore, value availability more on the way to work in the morning than after work in the evening. Second, the biking distance parameter is positive when the route length is less than 1.5 km, and negative when the route length exceeds 1.5 km. This is consistent with both the model-free preliminary evidence directly observed in the data and the reduced form regression results. The coefficients of both origination and destination station counts are positive and statistically significant. This implies that commuters get higher utility when there are more station options in an area. Third, the average ascending grade and the ascending percentage both have negative coeffi-

Non-linear para.		Linear para.	
O. Pop. Density	-0.045	Intercept	-3.494
(Ppl per hec)	(0.004)		(0.586)
D. Pop. Density	0.005	S. Station Count	0.897
	(0.007)	(In log)	(0.024)
O. Google Plc.	0.182	E. Station Count	0.924
(Num per 4 hec)	(0.024)		(0.052)
D. Google Plc.	0.037	S. Bike Avail.	0.978
	(0.011)	(%)	(0.189)
O. Walking Dist.	-2.036	E. Dock Avail.	0.822
(1km)	(0.288)		(0.282)
D. Walking Dist.	-4.174	Route Dist. 1	2.096
	(0.724)	(In log)	(0.201)
O.D. Driving Dist.	0.151	Route Dist. 2	-4.083
(1km)	(0.031)	(>1.5km, in log)	(0.267)
O.D. Transit Dist.	0.075	Avg. ascent grade	-7.375
(1km)	(0.032)	($\tan\alpha$)	(2.694)
		% segments ascend	-1.223
		(betweem 0 and 1)	(0.149)
Pseudo-R^2	0.792		

Table 1.4: Demand estimates: evening rush hour

cients and are mostly significant. This implies that the routes with higher ascending grade or higher share of ascending segments are less attractive to commuters which is consistent with our intuition.

Next, we discuss the non-linear parameters in the density model and the utility of the outside option. First, all walking-distance coefficients are negative and significant, except for the walking distance from the ending station block to the destination location in the morning rush hour. This implies that commuters prefer docking stations closer to their origination and destination locations of interest. However, when we compare the pseudo-R^2 of the results to those where we exclude walking distance from the model, the difference is less than .01. In other words, including walking distance does not improve the explanatory power of the model.[17] Second,

[17]The detailed results are presented in the Online Appendix.

for the morning rush hour, the population density coefficient is much bigger at the origination location k than at the destination location l. The reverse is true for the Google place count coefficients. This indicates that the commuters are more likely to travel from the residential areas of the city (where they live) to the city center (where they work). The direction of the traffic is consistent with what we observe directly from the data in Figure 1.1. We observe the opposite pattern in the magnitudes of those coefficients for the evening rush hour. This result shows that our model captures the directions of the traffic nicely, which would have been impossible to achieve had we treated stations as independent instead of as nodes linked to each other on a network. Third, the driving distance and transit distance coefficients indicate that commuters find biking a less attractive option when the trip is longer.

Finally, the pseudo-R^2 of both the morning and evening rush hour results, 0.669 and 0.793, respectively, shows that our model fits the data well. Moreover, as shown in the Online Appendix, the pseudo-R^2 remains high with different model specifications. This suggests that our model fit is robust to model specifications.

1.5 Counterfactuals

Using the estimated model, we conduct three counterfactual experiments related to network design and expansion. In the first counterfactual, we evaluate a specific plan to expand the network into the Islington and Hackney areas, which the local community proposed in 2012. Using our model and estimates, we show that although the expansion benefits the local community in Islington and Hackney, the magnitude of the overall usage increase in the network is not substantial. Our analysis provides the managing company with important insights and guidance for the expansion of the network. In the second counterfactual, we generalize the insights from the first counterfactual and investigate the best locations to add stations in the current net-

work. In particular, we compute the marginal effect of adding one station at different locations in the network. We show that adding stations to the city center leads to ten times more usage increase than adding stations to the peripheries. More importantly, we identify two types of network effects in our findings. Our results highlight the significance of network effects in understanding and evaluating the expansion of the network. In the third counterfactual, we investigate a different type of expanding strategy than in the second counterfactual. Instead of adding stations to the network, we keep the current network and study the best locations to add bikes and docks. Similar to the second counterfactual, we show that the network effects play an important role in determining the usage increase when adding bikes and docks to different locations. Moreover, by comparing the differences between adding bikes and docks, we highlight the interplay of network effects and the direction of commuting traffic in the results.

1.5.1 Expansion to Islington and Hackney

Like many bike sharing programs in major cities, the "Boris Bikes" went through several major expansions since they were first introduced.[18] The city has been expanding the network continuously, [19] and there are calls and proposals from the local government and communities for further expansions. One of the many proposals is to expand the bike share system to Islington and Hackney.[20] In this counterfactual, we use our estimated model to evaluate this particular expansion proposal.

We add four new blocks of stations covering the Islington and Hackney areas. Each block contains four stations, which is the same station density as in the neigh-

[18]Source:https://tfl.gov.uk/info-for/media/press-releases/2013/december/mayor-launches-huge-expansion-of-flagship-barclays-cycle-hire-scheme

[19]Source:http://www.cityam.com/268035/lycra-ready-south-london-santander-cycles-scheme-expanding

[20]Source: http://www.islingtongazette.co.uk/news/environment/plea-for-boris-bikes-to-be-wheeled-out-across-islington-1-1454024
https://www.change.org/p/transport-for-london-and-islington-council-roll-out-the-london-cycle-hire-scheme-to-the-whole-of-islington

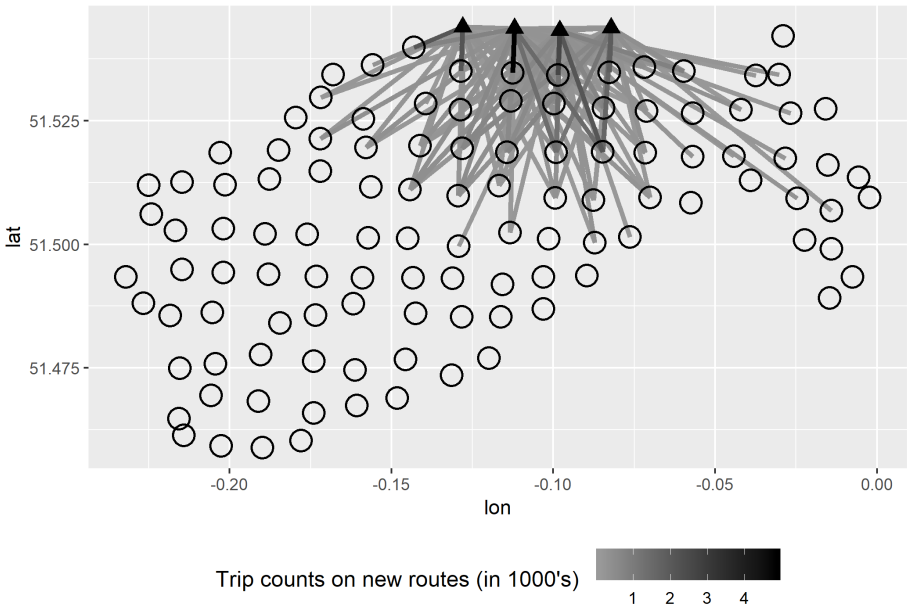

Figure 1.4: Predicted new station usage after expansion

boring blocks observed in the current network. We compute the sum of predicted usage during morning and evening rush hours (throughout a year). We assume that the new station blocks have the same average availability levels to those of the closest existing blocks. The availability of the rest of the station blocks stays the same. Given we are adding a small number of new stations to the network, we think this is a reasonable assumption. The results are presented in Figure 1.4.

The links indicate the routes on which usage increases after the expansions. The shading of the link indicates the level of usage increase: the darker the color, the higher the level of usage. Most of the usage increase comes from trips between the new stations and two areas: one around the new stations, and the other close to the city center. This result is consistent with the intuition that customers are most likely to use bikes to travel to nearby areas or to commute to and from the city center.

However, the key point is that the magnitude of the usage increase is very limited. The total usage increase is about 61,000, which is 1.5% of the total number of trips on the network during the morning and evening rush hours before the expansion. Our

model also shows that, if we compare the trip increase per additional station, this usage increase is less than 20% of the usage increase had the new stations been added to the city center instead. It suggests that, if maximizing the overall usage is the objective, adding stations in Islington and Hackney is not optimal. We revisit this point and provide a more general discussion about the optimal expansion location in the next counterfactual.

To summarize, although the expansion would benefit the residents in Islington and Hackney, the usage increase would be much higher if the new stations were allocated, instead, to the city center. The managing company in the city of London faces a trade-off between benefiting more people citywide versus providing service access to a particular community.

1.5.2 Adding one station to the system

In the second counterfactual, we generalize the insights from evaluating the particular expansion proposal in the first counterfactual. We analyze the optimal location to expand the network and highlight the importance of network effects in the efficacy of different types of expansions. The insights are not only important to the bike share system in London, but are also general enough to be applicable to the design of other bike share systems.

In this experiment, we add one station to different locations of the network. We investigate where the optimal location is–i.e., the location that leads to the highest usage increase in the entire network. Our analysis not only takes into account the usage increase on routes originating from and ending at the new station. It also considers the change in usage in other parts of the network induced by customers substituting between routes. In other words, we calculate the change in usage for all routes on the *entire* network, which, we show, is key to evaluating the expansion of the system.

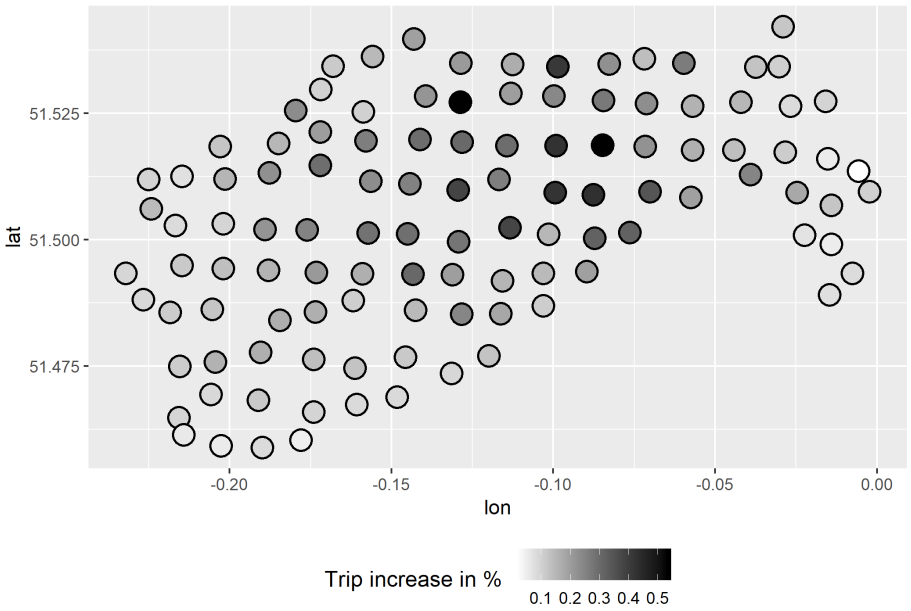

Figure 1.5: Predicted usage increase after adding one station

Adding one station is equivalent to increasing the number of stations by 0.14%. We compute the percentage increase in total usage in the network depending on the block of stations to which the new station is added. Similar to the previous counterfactual, we assume the availability level of all station blocks stays the same. Give the limited impact one station has on the entire network, we think this is a reasonable assumption. The results are presented in Figure 1.5. Each dot represents one of the 112 blocks in our coverage area. The shade of the dot indicates the total usage increase in the network when the extra station is added to the block. As shown in the figure, there is large heterogeneity in how effective the additional station is in increasing the overall network usage. The percentage usage increase varies from close to zero to around 0.5%. The usage increase when the station is added to the city center can be ten times as high as when the station is added to the peripheries. This shows that increasing *density in the city center* has much bigger positive effect on the overall network usage than increasing the *scope* of the network. The station density is too low in city center and too high in the peripheries. Thus, the city center

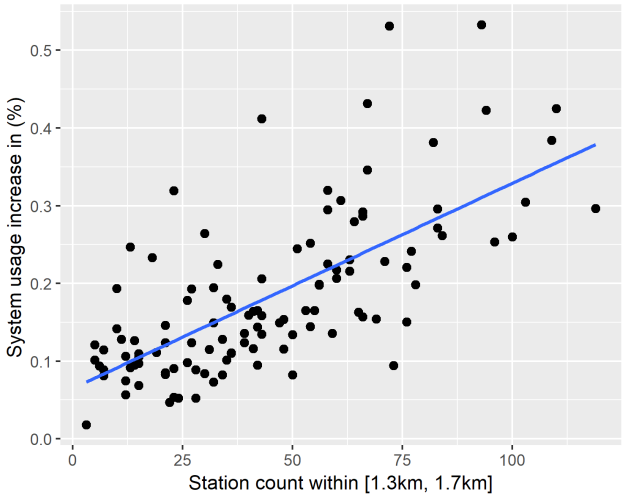

Figure 1.6: Predicted usage increase vs. number of stations 1.3km to 1.7km away

is clearly the bottleneck. This finding is partially a result of the implementation of the 300m between-station rule. Our analysis suggests that the uniform density rule is much too rigid. A redistribution of stations from the peripheries of the coverage area to the city center would make the current network far more efficient, if efficiency is judged by overall usage.

We also investigate the importance of network effects in the large heterogeneity of the predicted system usage increase, depending on where the new station is added. First, in Figure 1.6, we plot the usage increase against the number of stations that are approximately 1.5 km away from the focal station block. There is a very strong positive correlation. This suggests that adding station to *central* regions in the network, which we define next, is an important determinant of the usage increase and the effectiveness of network expansion. This motivates us to look for a precise measure of centrality that captures the network effect associated with it.

We construct a centrality measure using the number of stations approximately 1.5 km away and compare the measure to station-level usage determinants to better understand the network effect. This centrality measure is constructed by counting the number of stations between 1.3 km and 1.7 km away for each station and then

averaging over all stations within each block. We regress the total usage increase if one station is added to each block (presented in Figure 1.5), on our centrality measure of the block, the average population density of the block, the average number of Google places within the block, and the total number of stations in the block. We also compare our centrality measure with the conventional eigenvector centrality measure, where the adjacency matrix is defined by one over the distance between each pair of station clusters. We regress the predicted usage increase (when adding one station to each station block) on the three block characteristics and the two centrality measures. We study how much those five factors can explain the variation across the predicted usage increase at different station blocks. We do so by comparing the adjusted R^2 in linear regressions. Notice that the estimated coefficients do not have causal interpretations. Therefore, in the results, we report only whether the coefficients are statistically significant, and the adjusted R^2. We present the results in Table 1.5.

POPULATION DENSITY	✓	✓	✓	✓
GOOGLE PLACES	✓***	✓***	✓***	✓**
NUMBER OF STATIONS	✓***	✓***	✓	✓***
OUR CENTRALITY	-	✓***	-	✓***
EIGENVECTOR CENTRALITY	-	-	✓***	✓
ADJUSTED R^2	0.30	0.52	0.42	0.52

Table 1.5: Determinants of usage increase and network effect

In column 1 of Table 1.5, we present the baseline result in which we regress the predicted usage increase (when adding one station to each block) on three block characteristics. The result shows that the station block characteristics explain 30% of the variation in predicted usage increase across blocks. The regression in column 2 includes our constructed centrality measure as an additional explanatory variable.

Comparing column 1 and column 2, we see that including our centrality measure increases the adjusted R^2 by 73%. In other words, using our measure of centrality, being central on the network is key to locating the optimal blocks for the expansion of the network. This is the first type of network effect of relevance to the station network expansion and design, which we identify using the model and the estimation. The regression in column 3 replaces our centrality measure with the conventional eigenvector centrality measure. The comparison between column 2 and column 3 shows that our centrality measure explains the usage increase almost twice as well as the conventional eigenvector centrality measure. Finally, when we include both centrality measures, the conventional measure does not explain any additional variation in the usage increase across station blocks: the adjusted R^2 is still 0.52. Moreover, across all four regressions, the adjusted R^2 is never higher than 0.52. This means that, even when we can identify the right network centrality measure ex ante, the richness of the structural model helps us predict usage increase and identify optimal expansion locations much better than simple reduced-form regressions.

Next, we illustrate a second type of network effect by conducting the following analysis. We recompute the predicted usage increase resulting from adding one station to each block with the following change. While keeping the origination station block characteristics (population density, place counts, number of stations, and bike availability) and route distance the same as in the data, we set the destination station block characteristics (population density, place counts, number of stations, and dock availability) as the median values of those characteristics. Thus, we investigate whether the kind of stations (i.e., stations with similar characteristics) to which a station is connected matters for usage and network expansions. In other words, we examine whether a good match of origination and destination stations matters for usage. Specifically, by setting the destination station characteristics to the median level, we break the match in calculating the predicted usage increase. The difference

between the result of this calculation and the original usage increase computed above is the network effect that comes from a good match between an origination station and a destination station.

We find that, when setting the destination station characteristics as the median value, the usage increase goes down by 17%, on average, compared to the original usage increase. This indicates that not only the number of connecting stations approximately 1.5 km away matters, but that the *kind* of connected stations also matters for usage and network design. This is another type of network effect that we find relevant for evaluating network expansions in our analysis. In other words, treating stations as individual products and studying the demand for each on its own is not enough. Including the entire network of stations in the analysis is crucial to understanding customer demand and evaluating system expansion strategies.

1.5.3 Improving long-term availability in the system

Adding stations to the network is one way of expanding the bike share system. Another way of expanding the network is to add bikes or docks to the existing stations in the network. This type of expansion can be captured by improving the average bike and dock availability in our model. In this counterfactual, we analyze the optimal locations for the operating company to add bikes and docks. This is similar to the long-term effect of improving availability computed in [63], but our focus is on the optimal area to improve the availability measures in terms of long-term system usage. Moreover, we study the effects of improving both bike availability and dock availability, while [63] only considers bike availability. We present the results for the evening rush hour in this section. The analysis of the morning rush hour is done exactly the same way, and the results do not change qualitatively. Therefore, we leave the results for the morning rush hour to the Appendix.

In this experiment, we first calculate the improvement of system usage in the

evening rush hour by adding 0.05 to the average bike availability of each cluster–that is, the predicted usage increase of the entire system if the operating company makes the probability of finding an available bike for each cluster 5 percentage points higher during the evening rush hour. Similarly to the previous counterfactual, we present in Figure 1.7 the predicted system usage increase in percentage when adding 0.05 to the average bike availability of every cluster, and in Figure 1.8, the improvement when adding dock availability. Again, each dot represents one of the 112 blocks in our coverage area, and the shading of the dot indicates the total percentage usage increase in the network. We can see that there is also a wide gap in predicted trip increase between the most and least beneficial areas in which to boost availability during the evening rush hour.

The results also illustrate the interplay of the unidirectional travel pattern and the network effect. For the evening rush hour, we can see that people generally move from the city center to more residential areas and, therefore, the best places to improve bike availability are all in the very center of the network. However, for dock availability, based on the direction of commuting flow, we should focus on residential stations around the peripheries since, on average, there will be more travelers going to those stations. However, as a destination, those stations in the peripheries tend to have fewer connecting stations that people can bike from–i.e., from the network effect point of view, they are not the perfect choice. The two contrasting factors result in the best clusters for boosting dock availability. Figure 1.8 shows that the highest usage increase stations are a little more scattered around the map, not as concentrated in the very center of the network as in Figure 1.7.

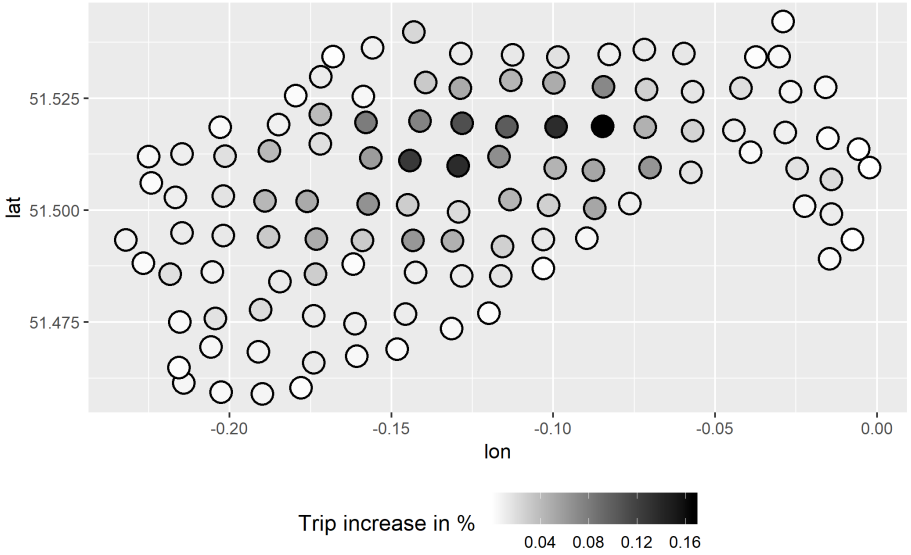

Figure 1.7: Predicted evening rush hour usage increase after improving bike availability

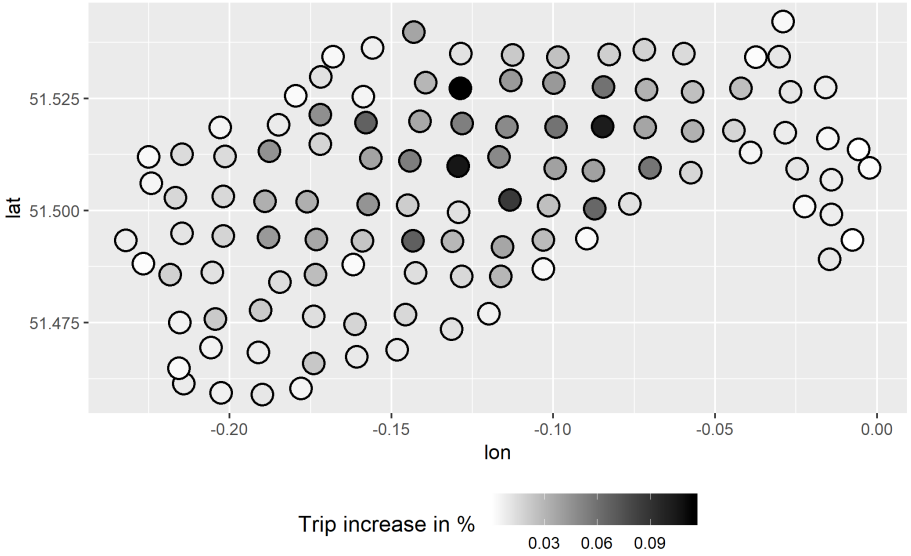

Figure 1.8: Predicted evening rush hour usage increase after improving dock availability

1.6 Conclusion

We study customer demand in the station network of the London bike share system. Using our model and the estimated preference parameters, we provide guidance on the network design and expansion of the system. We highlight the tradeoff between the density and scope of a network in increasing overall usage. We also evaluate a particular proposal of expansion to the Islington and Hackney areas of the city. Our empirical results provide important insights and policy recommendations for the managing company and the local government.

In the estimation of the structural demand model, we develop an instrumental variable approach to deal with the endogeneity problem of the choice set. We show that properly taking into account the endogeneity of choice set in the demand estimation is important for both the estimated parameters and the policy recommendations derived from the estimated model. The method can be applied to other empirical settings in which understanding customer demand is important and products and services can go out of stock.

Chapter 2

Machine Learning in Demand Estimation in Long-Tail Markets

2.1 Introduction

Structural estimation is an important tool of empirical studies in operations management, economics, and marketing. It allows us to recover parameter values with meaningful behavioral interpretations in decision models. Using those estimated models, we are able to provide descriptive policy recommendations for managers and policy makers to make better decisions. Demand estimation is a prominent example in such structural estimation methods. We can estimate customer preference parameters for various product or service features such as price using these tools. More importantly, using the estimated models, we are able to provide prescriptive policy recommendations on important managerial decisions such as pricing, assortment, introduction of new products, etc. to managers and policy makers.

Given the importance of these decisions to managers and policy makers, it is key to recover the parameter estimates without bias in empirical studies. To achieve this goal, many estimation routines have been developed in the literature [20, 3, 21]. The most commonly adopted is introduced by [20] (BLP). The method has two main advantages. First, it allows for endogenous product or service features such as price to be correlated with unobserved product or service features. Second, it allows for heterogeneity in customer tastes captured by random coefficients in the utility function. It has been widely applied in studying customer choice behavior in many

different industries in operations management, economics, and marketing [70, 84, 56]. This is the method we focus on in the current paper.

One disadvantage of BLP is that the method exhibits major limitations when the customer choice data has long tails. As discussed in detail in Section 2.2, BLP does not allow products or services with zero sales to be included in the estimation. Moreover, the method produces noisy estimates when products and services have very low sales or market shares. However, in practice, as shown in Section 2.2, customer choice data often has very long tails, i.e. many products and services offered have zero sales or very low sales at a given period of observation. The phenomenon has become increasingly common with big data and, in particular, large customer choices in settings such as online retail. We show that common practices of dropping products and services with zero and very low sales introduces bias in BLP estimates. Moreover, the bias in estimates leads biased policy recommendations and managerial decisions.

In this chapter, we propose two new estimators that correct for the bias caused by long tail data in BLP estimates. Both estimator have two stages. One is called two-stage bound estimator which builds on the bound estimator proposed by [51]. The other is called two-stage weighting estimator. The two estimators only differ in the second stage. In the first stage of the two proposed estimators, we use deep learning to predict the market shares of all products or services including those with zero or low sales. In the second stage, we utilize the predicted shares from first stage to correct for the bias in the estimation. More importantly, using the de-biased demand estimates from our proposals in counterfactual analysis, we can correct for the bias in policy recommendations and managerial decisions.

Our method of correcting for the bias in BLP estimates is analogous to the bias correction often used in the classic literature of estimating treatment effects. We use the predicted shares in the first stage to construct weights to correct for the bias in the second stage of the estimation. This is similar to using estimated propensity score

to construct weights or matches to correct for bias in estimating treatment effects.

To the best of our knowledge, there is only one recent paper in the econometrics literature that studies the same problem. [51] proposes a bound estimator to correct for the bias. The limitation of their approach is that it can only allow for a few covariates with continuous values. While the second stage of our two-stage bound estimator has similarities to [51], it does not have any limitation on the dimension of the covariates and improves on the precision of the [51] estimator. We provide detailed comparison of the two approaches and their performances later in the paper. Our two-stage weighting estimator is based on entirely different ideas than bound estimator.

The rest of the chapter is organized as follows. Section 2.2 provides examples of long tail data and explains the source of bias. Section 2.3 provides a brief description of the two-stage estimator. We describe the detailed construction of the estimator in Section 2.4. We show the simulation results in Section 2.6. We are in the process of implementing our method in real data, which is to be added to the paper next.

2.2 Bias in Demand Estimation with Zero's

2.2.1 Long-tail pattern in empirical data

As pointed out in [51], long tail patterns arise in typical demand data like the workhorse store level scanner data. Such data sets are used in a huge empirical literature to estimate consumer preference. To recall the empirical pattern, we first take another look at the Dominick's Finer Foods grocery chain data, which is also used by [51].

The data set in Table 2.1 is at a store/week/UPC level where UPC stands for universal product code. The first column shows the production categories we are aggregating demand information. The second column calculate the average number

Category	Avg num of products in a store/week pair	Percent of total sales of the top 20% products	Percent of Zero Sales
Laundry Detergents	200	65.52%	50.46%
Soaps	140	77.26%	44.39%
Toothbrushes	137	73.69%	58.63%
Beers	179	87.18%	50.45%
Crackers	112	81.63%	37.33%

Table 2.1: Selected Product Categories in the Dominick's Database [51]

of UPCs in a (store, week) pair. The third column shows presents the percentage of total sales that are from top 20% best sellers while the last column averages the percentage of zero sales product across all (store, week) pairs. We can easily see that sales are heavily concentrated on top selling products: top 20% takes nearly 80% of the market share and the percentage of zero sales product in an average (store, week) pair is quite large. Take Laundry Detergents as an example: on average 65.52% of sales are from top 20% best sellers and half of the detergent products do not sell at all for an average (store, week) pair.

As argued in [51] and Anderson (2006), this long-tail pattern is quite universal. It is becoming even more prevalent nowadays thanks to more granular data for customers, products and markets at finer time scale becoming available and online retailers/platforms offering much larger assortments than ever before. For example, Airbnb has over 30,000 listings for New York city in 2017[1] and Amazon has millions of products offered to customers. We give a detailed example using London's Bike Share System data from [58]. In Table 2.2, we show the percentage of zero usage biking routes, defined as (starting station, ending station) pair, from 2017 data. We aggregate routes based on distance percentile where distance is simply calculated as the shortest distance on earth between starting and ending stations. The first column

[1]http://www.crainsnewyork.com/reports/airbnb-data/#!/

shows the distance percentile buckets while the second column calculates within that bucket, the percentage of biking routes that have never been used throughout 2017. We see that clearly as distance gets longer, the percentage of zero routes increases as expected. However, even around median distance say $50\% - 60\%$, there are already 40.6% of routes have 0 usage for the entire year. If we want to study route usage at weekly or monthly level as opposed to yearly, these numbers will be even higher.

With the observations above, we formally defined long-tail pattern in our context as the phenomenon that empirical frequency of quantities of sales or market shares drop very fast as sales/market shares increase, leaving a long thin right tail. It implies that many products will not have lots of sales and that market shares are concentrated heavily in the top-selling products. In these long-tail data sets, we often observe a large fraction of products having zero sales.

Another observation is that in modern "big data" applications, not only do we often observe long-tail pattern, more and more features are available to researchers. For example, in the study of London bike share system by [58], there are more than 100 features in the original data that can be utilized in the demand estimation. It is important that demand estimation techniques can scale up well as the number of covariates grow. In this work, we use machine learning tools to ensure that our proposal to estimate demand in long-tail markets scales well with number of predictors.

2.2.2 Bias in common practice

Now that we see that long-tail pattern is very common, we describe in this subsection why long-tail market share data matters to demand estimation and why it is important to correctly account for it in order to get unbiased demand estimate and make optimal managerial decisions.

As pointed in the demand estimation literature [77, 51] , the standard structural demand estimation method BLP [20] cannot take data points with zero market shares

Distance percentile	Zero routes %
0-10%	4.7%
10-20%	5.3%
20-30%	9.7%
30-40%	17.2%
40-50%	27.0%
50-60%	40.6%
60-70%	55.2%
70-80%	68.2%
80-90%	81.9%
90-100%	94.5%

Table 2.2: Zero usage routes in the London bike share system [58]

since as shown later, the estimation algorithm breaks down when it tries to take log of the 0 market share. This breakdown makes empirical researcher must deal with zero sales data point in long-tail market. Usually there are two different ways to address long-tail pattern: one is to aggregate products up to larger categories so that the long-tail pattern is not that severe and percentage of zero sales products become very low; the other one is to directly ignore all zero sales product in the data. Sometimes the two are used together. It turns out that both solutions are problematic.

The first common practice is aggregating into larger product categories and treating those categories as the new "products". This will lose variation from more granular data. More importantly we will miss the substitution effects between products within category.

Whether or not we do further aggregation, another common practice is to throw away the zero sales data. This practice amounts to subset selection and the subset we are using the product with strictly positive (> 0) sales. Often the subset selection is even more extreme: for example we not only throw away 0 sales products but also throw all products except for the 20% top selling products in our estimation. Intuitively, this subset used in estimation is not a random subset selected from original

data, which raises the big question that whether such selection will lead to bias in our estimates. Indeed, there is a big selection bias introduced by subset selection, as pointed out by [51]. We briefly explain the bias next. The discussion below largely follows from [51] and [20].

2.2.2.1 BLP estimation

We first provide a brief description of the classic BLP estimation framework.

Denote total number of markets observed as T and in each market t there are J_t products offered. The observation unit will be (product, market) pair (j, t). There are $n_t, t = 1, 2, ..., T$ consumers in total in market t. For each jt, we model consumer i's utility of purchase jt as given by

$$u_{ijt} = x'_{jt}\beta_i + \xi_{jt} + \epsilon_{ijt}, \tag{2.1}$$

where x_{jt} of dimension d_x is the product features observed by us as researchers. Coefficients of interests β_i denote how each x_{jt} affect consumer utility. Also there are unobserved product features denoted as ξ_{jt} which can approximate things like unobserved quality, brand reputation etc. It is almost always the case that we are not able to observe all features that matter to consumers and thus it is important for the model to allow for such unobserved characteristics. Lastly we assume that ϵ_{ijt} has extreme value distribution and that consumer can also choose to purchase nothing. We call that as outside option and denote outside option at market t as $0t$. We normalize outside options' utility to be 0, i.e., $u_{i0t} = 0$ for any i, t. Every customer i in market t makes a choice among outside option $0t$ and products jt, $j = 1, 2, ..., J_t$ by maximizing her utility.

Standard random coefficients specifications assume that demeaned beta $\tilde{\beta}_i := \beta_i - \mathbb{E}[\beta_i]$ follows a normal distribution with $\tilde{\beta}_i \sim \mathcal{N}(0, \Sigma_\beta)$ where Σ_β is a d_m by d_m covariance matrix. We usually assume that it is a diagonal matrix and denote its diagonal elements by a column vector λ. Let $\beta := \mathbb{E}[\beta_i]$. Thus, the preference

parameter we want to estimate is $\theta := [\beta', \lambda']'$ and denote the true values by $\theta_0 := [\beta_0', \lambda_0']'$. Define $\delta_{jt} := x_{jt}'\beta + \xi_{jt}$.

The distribution assumption on ϵ_{ijt} leads to a closed form solution for choice probabilities of consumer i in market t choosing jt given $\tilde{\beta}_i$,

$$\mathbb{P}[i \text{ chooses } j|\tilde{\beta}_i] = \frac{e^{\delta_{jt}+x_{jt}'\tilde{\beta}_i}}{1+\sum_{k=1}^{J_t} e^{\delta_{kt}+x_{kt}'\tilde{\beta}_i}}$$

The true choice probability or market share of product j in market t is then given by

$$\pi_{jt} = \int_{\tilde{\beta}} \frac{e^{\delta_{jt}+x_{jt}'\tilde{\beta}}}{1+\sum_{k=1}^{J_t} e^{\delta_{kt}+x_{kt}'\tilde{\beta}}} d\mathbb{P}(\tilde{\beta}) \qquad \tilde{\beta}_i \sim \mathcal{N}(0, \lambda_i), \quad i = 1, 2, ..., p \tag{2.2}$$

Then the true choice probability for outside option is given by $\pi_{0t} = 1 - \sum_{j=1}^{J_t} \pi_{jt}$ for each market t.

Let $s_t := (s_{0t}, s_{1t}, s_{2t}, ..., s_{J_t t}) \quad t = 1, 2, ..., T$ be market shares observed in the data. We also call s as empirical shares. We assume that the numbers of purchases in each market denoted by $N_t, \quad t = 1, 2, ..., T$ are known to the researcher. Further assume s_t's are generated by multinomial sampling based on N_t i.i.d. customers making purchasing decisions by true choice probability π_t. To be precise,

$$s_{jt} = \frac{\sum_{i=1}^{N_t} 1_{i^{th} \text{ customer chooses } j}}{N_t} \quad \forall t = 1, ..., T, j = 0, ..., J_t \tag{2.3}$$

Note that empirical shares include the shares of outside options, which is represented by $j = 0$ for each market t. Also note that fixing any market t, as $N_t \to \infty$, empirical shares are consistent estimators for true choice probability, i.e., $s_{jt} \to \pi_{jt}$ for any $j = 0, 1, ..., J_t$ if say the number of products J_t remains the same.

Let $\delta_t \in \mathbb{R}^{J_t} \quad t = 1, 2, ..., T$ be $\delta_t := (\delta_1, \delta_2, ..., \delta_{J_t})'$ for each market t and similarly, define $\pi_t := (\pi_{1t}, \pi_{2t}, ..., \pi_{J_t t})'$ and $x_t := (x_{1t}', x_{2t}', ..., x_{J_t t}')'$. Let function $\sigma(\delta_t, x_t, \lambda)$ be the mapping from mean utility δ_t, observable features x_t in a market t to true market

share/choice probability given λ. To be precise, $\sigma(\cdot; x_t, \lambda)$ is a mapping from $\mathbb{R}^{J_t}$ to region $\left\{x \in \mathbb{R}^{J_t} : x > 0, \sum_{j=1}^{J_t} x_j < 1\right\} \subset \mathbb{R}^{J_t}$ such that for any market t and any product $j = 1, 2, ..., J_t$,

$$\sigma_j(\delta_t; x_t, \lambda) := \int_{\tilde{\beta}} \frac{e^{\delta_{jt} + x'_{jt}\tilde{\beta}}}{1 + \sum_{k=1}^{J_t} e^{\delta_{kt} + x'_{kt}\tilde{\beta}}} d\mathbb{P}(\tilde{\beta}) \qquad \tilde{\beta}_i \sim \mathcal{N}(0, \lambda_i), \quad i = 1, 2, ..., p \quad (2.4)$$

Without a closed-form formula for the integral in equation (2.4), usually we use Monte-Carlo simulations to approximate the integral or expectation with respect to $\tilde{\beta}_i$'s. Based on Berry (2006), function σ defined in equation (2.4) is invertible under general assumptions. We could define the inverse function as σ^{-1}. Note that this inversion function is even harder to compute. Originally, [20] uses a contraction mapping algorithm to compute σ^{-1}. Standard BLP assumes next that $\pi_t \approx s_t$ and $\sigma^{-1}(s_t; x_t, \lambda) \approx \sigma^{-1}(\pi_t; x_t, \lambda)$ for all $t = 1, 2, ..., T$, $x_t \in \mathbb{R}^{J_t \times p}$ and $\lambda \in \mathbb{R}^p$.

We use $z_{jt} \in \mathbb{R}^{d_z}$ to denote instruments for covariates x_{jt}, where d_z denotes the number of instruments[2]. We follow general methods of moments (GMM) framework and the moment conditions we rely on to identify parameters is $\mathbb{E}[\xi_{jt}|z_{jt}] = 0$. Let W be the weighting matrix for different moment conditions. We minimize the empirical analog of $\mathbb{E}[\xi_{jt} z_{jt}] = 0$ to find our estimates. To be precise, we minimize the following objective function to get $\hat{\theta} = (\hat{\beta}', \hat{\lambda}')'$:

$$\min_{\theta=(\beta',\lambda')'} \sum_{t=1}^{T} \sum_{j=1}^{J_t} (z_{jt}(\sigma^{-1}(s_t, x_t, \lambda) - x'_{jt}\beta))' W z_{jt}(\sigma^{-1}(s_t, x_t, \lambda) - x'_{jt}\beta) \quad (2.5)$$

As pointed out by [83], the contraction mapping inside the optimization problem 2.5 of standard BLP estimation will make the procedure numerically unstable. [83] proposes using optimization with non-linear constraints to deal with computation of σ^{-1}. Let $J := \sum_{t=1}^{T} J_t$ be the total number of products across all markets. The new optimization problem is given by,

[2]Note that in the particular case of no endogeneous covariates, we have $d_z = d_x$ and $z_{jt} = x_{jt}$

$$\min_{\beta\in\mathbb{R}^p,\lambda\in\mathbb{R}^p,\delta\in\mathbb{R}^J} \sum_{t=1}^{T}\sum_{j=1}^{J_t} \left(z_{jt}\left(\delta_{jt}-x'_{jt}\beta\right)\right)' W \left(z_{jt}\left(\delta_{jt}-x'_{jt}\beta\right)\right) \tag{2.6}$$

$$\text{s.t.} \quad \sigma\left(\delta_t; x_t, \lambda\right) = s_t \qquad \forall t = 1, 2, ..., T \tag{2.7}$$

The above formulation is often called Mathematical Programming with Equilibrium Constraints (MPEC). We can easily see that this problem is non-convex due to the nonlinear equality constraints. However, we only need to compute σ instead of σ^{-1}, which does not need the contraction mapping routine. Also note that δ_{jt}'s are additional variables introduced for the solver to optimize over together with β and λ. In the simulation experiments of this work, we focus on MPEC formulation for its numerical advantages over contraction mapping. Effectively, we are doing the same inversion step (calculating σ^{-1}) by enforcing the constraints in (2.7) that matches model predicted shares to observed shares, instead of contraction mapping proposed by [20]

It is obvious that $\pi_{jt} = \int_{\tilde{\beta}_i} \frac{e^{\delta_{jt}+x'_{jt}\tilde{\beta}_i}}{1+\sum_{k=1}^{J_t} e^{\delta_{kt}+x'_{kt}\tilde{\beta}_i}} d\mathbb{P}(\tilde{\beta}_i)$ is strictly positive since logit probability is strictly positive, i.e., $\frac{e^{\delta_{jt}+x'_{jt}\tilde{\beta}_i}}{1+\sum_{k=1}^{J_t} e^{\delta_{kt}+x'_{kt}\tilde{\beta}_i}} > 0$. Therefore $\pi_t = \sigma(\delta_t; x_t, \lambda) > 0$ element-wise for any values of δ_t, x_t and λ. This means that in the random coefficient model, model predicted market shares are strictly positive. Therefore, when we try to invert observed market shares from data, if for some (j,t), $s_{jt} = 0$, then corresponding $\delta_{jt} = \sigma^{-1}(s_t; x_t, \lambda)$ has to be $-$Inf and this the exact reason why BLP estimation becomes infeasible when there are 0's in market shares data s. The same problem occurs regardless of whether we are using BLP contraction mapping or MPEC formulation to evaluate σ^{-1}. Indeed, even if we are using MPEC, the constraints in (2.7) with $s_{jt} = 0$ will be infeasible for any values of the input variables. The most common way to deal with this minus infinity issue is to just throw away (j,t) with $s_{jt} = 0$, which will cause biased estimates in a way similar to the classic selection bias problem in economics. We explain this bias in the next subsection.

2.2.2.2 Source of the bias using selected subset

The source of the bias with subset selection in our estimation is similar to the classic selection bias problems as described in [90]. In this subsection we assume that there is only one covariate and it is exogeneous to illustrate the intuition of the bias. The exposition here follows largely from [51]. Under these simplifying assumptions we have $d_x = d_z = 1$ and $x_{jt} = z_{jt}$. The identification of our preference parameters comes from the moment condition

$$\mathbb{E}\left[\xi_{jt}|x_{jt}\right] = 0$$

If, however, we apply some threshold M on s_{jt} and only take the subset $(j,t) : s_{jt} > M$ in the estimation, then the moment condition will be violated in the following sense

$$\mathbb{E}\left[\xi_{jt}|x_{jt}, s_{jt} > M\right] \neq 0$$

As described above, a typical choice of M is $M = 0$ to deal with the inversion problem of σ function. There are also many empirical studies only considering top selling products, where M can be large too.

To illustrate the bias, we give two examples. Firstly, without loss of generality we have $\mathbb{E}[\xi_{jt}] = 0$ when covariates x has an intercept column (a column of 1's). However, after selection on $s_{jt} > M$, we will tend to select products with more attractive unobserved characteristics ξ into our estimation subset. As a result, if we focus on (j,t) that satisfies $s_{jt} > M$, the mean of ξ_{jt} will not be zero anymore and shift towards positive territory, i.e., $\mathbb{E}[\xi_{jt}|s_{jt} > M] > 0$.

Secondly, to provide more intuition about the source of the bias, consider the following example. Assume that the true value $\beta_0 > 0$ for the only one covariate, meaning that the higher this attribute, the more attractive the product is to consumers. If in the selected subset we see a product with very low x_{jt} and yet we observe large enough sales $s_{jt} > M$, then we know that ξ_{jt} must be very large and

that it is the very attractive unobserved characteristic that makes some customers choose this product (j,t). As a result, we have a negative correlation between x and ξ in the selected subset where $s_{jt} > M$. That is,

$$Cov(x_{jt}, \xi_{jt} | s_{jt} > M) < 0$$

This will result in our estimator converging to a value that is smaller than β_0 in large sample. If we take these biased estimates into downstream task such as new product design, pricing or assortment optimization, we will likely to arrive at sub-optimal decisions. The bias caused by subset selection in the current settings is very similar to classic selection bias problems in the econometrics literature [90]. In classic selection bias settings, the selection process is often straightforward: for example, it can be individuals selecting themselves into some treatment. However, the subset selection in the current setting results from the particular estimation routine of the BLP framework, which cannot take zero-share products.

To summarize, the selection bias can be severe and common in demand estimation in long-tail markets, which often arise in empirical applications, especially in modern "big data" world. Being able to correct for the bias is crucial, especially if we do not want our biased estimate to affect downstream task and lead to misinformed managerial decisions.

2.3 Description of Proposed Two-Stage Estimators

Here we briefly described our proposed solution and some intuition. We call our propose Two-stage Estimators. As the name suggests, it consists of two stages, where the first stage our task is pure predictive and the second stage we obtain estimates for the casual parameters of interests.

Stage 1: prediction stage

In the first stage, the goal is to train machine learning models to predict market shares s_{jt} using instruments z_{jt}. We store the final predicted shares denoted by $\hat{s}_{jt}$ for future use during second stage.

At this stage, we will be using data of all products including the ones with zero sales, i.e., $s_{jt} = 0$. We also apply cross-fitting procedure similar to [30] and [73] in this stage. We postpone the details of the discussion till the next section. The output of this stage is the predicted market shares $\hat{s}_{jt}$ for all (j, t). One might think that, we could choose a range of machine learning algorithms such as random forest, gradient boosting and deep neuron network to fit to target s_{jt} using features (z_{jt}. However, there is one complication here: the market share of a product j in market t does not only depend on its own characteristics z_{jt} but also depends on its values for other competing products in the same market. It is not possible either to build a regressor to predict s_t or s_{jt} using $z_t := (z_{1t}, z_{2t}, ..., z_{J_t t})'$ because the number of products in different markets can be different and machine learning algorithms require a fixed dimensionality for input features. We propose a new market share predictor based on deep learning models inspired by mixed logit functional form. In our simulation results we test the performance of our new proposal.

Stage 2: estimation stage

In the second stage, we only keep the subset with positive sales, i.e. $s_{jt} > 0$ but we utilize the predicted market shares $\hat{s}_{jt}$ from the first stage to correct for the selection bias and consistently estimate the preference parameters θ.

Before going into the details of the procedure which is provided in the next section, we discuss the intuition of the procedure. We can think of the high level ideas to be assigning (j, t) pair with higher $\hat{s}_{jt}$ more weights and (j, t) pair with lower $\hat{s}_{jt}$ less weights. We will be grouping product market pair (j, t) into different buckets

based on predicted market shares $\hat{s}_{jt}$ and treat them differently. To contrast with this, standard BLP method gives all product with positive sales equal weights. The reason we want to adjust weights this way is that the products with higher predicted shares are less vulnerable to selection bias since they are going to have positive sales and end up in the estimation subset anyway. However, if we focus on the bucket of (product, market) pair with extremely low predicted market shares, within this bucket subset selection based on realized market shares $s_{jt} > M$ will generate the most severe selection bias, because intuitively the products here are very likely to have 0 observed sales and it heavily depends on the draw on unobserved characteristics ξ_{jt}: if ξ_{jt} is very positive then s_{jt} could have some chance being greater than M; otherwise, however, it will almost certainly be $< M$ and be left outside the estimation subset. As a result, if we select subset based on $s_{jt} > M$, the distribution of ξ_{jt} in the selected estimation subset will be severely tilted.

The exact way we carry out this weighting is similar to the bound estimator proposed by [51] and will be described in section 4 in details. To compare our proposal versus bound estimator by [51], one analogy would be in propensity score matching vs. matching by covariates in causal inference literature. The way [51] forms buckets (which is called hypercubes in [51]) and grouping (j,t)'s is like doing matching by covariates where the covariates are instruments z_{jt}. With the help of a first stage, our proposal forms buckets and groups (j,t)'s based on a one-dimensional variable, the predicted market shares $\hat{s}_{jt}$, which is like doing matching by propensity score where $\hat{s}_{jt}$ serves as the propensity score. For propensity score it is shown that matching by (or conditioning on) propensity score is sufficient to ensure unbiased treatment effect estimates while here in our context we assume that predicted share function $\mathbb{E}[s_{jt}|z_{jt}]$ is a sufficient statistic that we need to control for in order to eliminate selection bias. The difference is that in causal inference, we match by covariates or propensity score to group similar units from treatment and control together into one bucket in order to

recover the true treatment effects while here we do not have treatment minus control but here our objective function is a sum of contributions from each bucket as shown in details in Section 2.4.

One main advantage of our two-stage estimator over bound estimator, similar to the advantage of matching by propensity scores over matching by covariates, is that our proposal scales much better as the number of instruments d_z grow. Because of the curse of dimensionality, even with moderate d_z, bound estimator could become inaccurate since it is impossible to find good matches.

2.4 Two-Stage Bound Estimator

We formally describe our algorithm in detailed steps below.

2.4.1 Stage 1: prediction

We formulate the nonparametric estimation problem for market shares. We have T markets and for each market $t = 1, ...T$ we know that there are $J_t + 1$ products being sold there. For any product j in market t, (j, t) where $j = 0, 1, 2, ...J_t$, we observed its market share $s_{jt} \in [0, 1]$ and note that $\sum_{j=0}^{J_t} s_{jt} = 1$ for all markets t. $j = 0$ always denotes outside option. The goal is to be able to predict this market share s_{jt} for each (j, t). Note that we can observe products with 0 market shares and we want our model to do well on those products too.

We may have features about the consumers in particular market t (say average consumer age, income etc..). But just to start off simple, we assume that we only have access to product-level features. Assume for each product market pair, (j, t) we observe a p dimensional vector z_{jt}, which is the value of instruments. Our fitted model will output a number $\hat{s}_{jt}$ for each (j, t) as the predicted market share. It is crucial that the input features to our machine learning model in this stage are instruments

z_{jt} instead of covariates x_{jt}. This is to ensure that we are making prediction based on something exogeneous to unobservable characteristics and the reason for that will be clear later on. Note that in the special cases where all covarites are exogeneous, instruments and covariates are the same, i.e., $z_{jt} = x_{jt}$. Sometimes in BLP we apply nonlinear transformations of instruments to use as additional instruments. This is related to the discussion of optimal instruments. We note that this is not necessary here in our first stage since we will be using highly non-linear machine learning methods. Including self-computed nonlinear transformation might be useful for a different reason. In machine learning tasks one might be able to get better predictive performance if researchers with some prior can precompute some transformations of raw feature variables that are important and correlated to the target variable and this is usually called feature engineering. Nonlinear transformation of original instruments might be useful in this fashion. In our simulation exercise, we just use the original instruments without adding additional nonlinear transformation of them.

The difficulties of directly applying any machine learning algorithms is that we are predicting a number between 0 and 1 and that the target share depends on the characteristics z_{jt} of other products in the same market t. To give a concrete example, let's say z_{jt} is of dimension 1 and represents some attractive feature: the higher the value, the better the products. If for market t_1, there are only two products being offered where $z_{1t_1} = 1$ and $z_{2t_1} = 1$, then we expect these two products to have similar market shares. For some other market t_2, however, we have $z_{1t_2} = 1$ again but $z_{1t_2} = 10000$. Then people in this market will prefer the second product $(2, t_2)$ much more than $(1, t_2)$. We can imagine that if our machine learning algorithm only has knowledge of products' own characteristics, it will do a very poor job predicting market shares for $(1, t_1)$ and $(1, t_2)$ since they have the same $z_{1t_1} = z_{1t_2} = 1$ but they are very likely to have dramatically different market shares or sales. This example might be a bit extreme but is able to illustrate our points.

One might think of an easy workaround to be including all products' z values in the feature input of machine learning algorithms. However that wouldn't work on well either since different markets could have different number of products. For example, if in some market t_1, there are only two products being offered while in another market t_2, we have 10000 products being offered. In this case, if we want to train a machine learning model that includes z_{jt} for all $j = 1, ..., J_t$ in the same market t when predicting market share for (j, t), we will have a problem since for market t_1, the dimension of input features to our machine learning algorithm would be $2d_z$ while for market t_2, the dimension is $10000d_z$. This is not feasible since machine learning algorithms do require a fixed length of input features. We describe our proposal next.

2.4.1.1 Deep market share model

The idea to overcome the difficulties of directly applying machine learning lies in the structure of discrete choice models. We first pose the familiar logit functional form to model choice probabilities but instead of a linear functional form, we use a general functions f on the exponents. That is

$$p_{jt}(w) = \frac{e^{f(z_{jt})}}{1 + \sum_{k=1}^{J_t} e^{f(z_{kt})}} \tag{2.8}$$

, where p_jt denotes the model choice probability for (j, t). Note that logit transformation used above is very similar to the softmax functions that are commonly used in neuron network, we can try to use neuron network to do general predictions for market share. We call the resulting neuron network structure as deep demand net and we describte the structure below. In Equation (2.8), we use a general mapping function $f : R^p \to R$, where $p = d_z$ is the number of features we observe. In our deep market share model, we specify f in (2.8) using Neuron Network and the network structure for f can be simple or complex, shallow or deep. Note that we can recover the standard logit functional form if we set function f to be just a one layer neuron

network without hidden layer and nonlinear activation functions, i.e., $f(x) = x'w$ where w denotes the free parameters of the neuron network structure.

Again, we use w to denote all free parameters in the neuron network structure we specify for f. Some times we call them trainable weights in the neuron networks. We are then able to minimize the following loss L^2 function over w:

$$\min_{w} \sum_{t=1}^{T} \sum_{j=1}^{J_t} \left(p_{jt}(w) - s_{jt}\right)^2 \tag{2.9}$$

Note that products with 0 sales, i.e., $s_{jt} = 0$, are included also in the evaluations of above loss functions.

Note that even though we can specify arbitrarily complex neuron network structure so that f can be very complicated, the above formulation of p_{jt} in Equation (2.8) still has the independence of irrelevant alternatives (IIA) assumption. We can overcome it thanks to the mixed logit idea: we could duplicate the same neuron network structure n times, where n can be thought of as number of "representative customers". That is, we have f_i,$i = 1, 2, ...n$ and all n models have exactly the same network structure but they do not share the same parameter values: denote their weights by $w_i, i = 1, 2..., n$. Then following the random coefficient or mixed logit model, the final model predicted market share for market product pair (j, t) is given by,

$$p_{jt}(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{e^{f_i(x_{jt})}}{1 + \sum_{k=1}^{J_t} e^{f_i(x_{kt})}} \tag{2.10}$$

Note that n is not a trainable parameter and we need to specify n first. In simulation we use $n = 100$. We minimize the same loss function, but over all w_i:

$$\min_{w_i, i=1,2,...n} \sum_{t=1}^{T} \sum_{j=1}^{J_t} \left(p_{jt}(w) - s_{jt}\right)^2 \tag{2.11}$$

Note that this formulation includes the traditional random coefficient model as a special case. To see this, we can specify all f's as a one-layer neuron network: $f_i(x) = x'w_i$, where w_i's are parameters to fit with restrictions: w_i for $i = 1, 2, ..., n$

are normally distributed with certain mean and variance where only the mean and variance are free parameters to fit.

Our formulation in Equation (2.10) is much more flexible than traditional random coefficient model in two ways:

1) f_i can be highly nonlinear and deal with large number of covariates easily, unlike the traditional linear utility framework

2) The heterogeneity in customers are not restricted to normal distribution as in the traditional random coefficient models. For example, it could well be a mixture of two normals and our model should be able to fit it very well.

Because of the L^2 norm loss function used in (2.11), the population minimizer would be the conditional expectation of observed shares s_{jt} conditional on z_t, i.e. $\mathbb{E}[s_{jt}|z_t]$, where $z_t = (z_{1t}, z_{2t}, ..., z_{J_t t})$. Since by assumption observed shares s_{jt} are generated by multinomial sampling with true choice probability π_{jt} and the multinomial sampling is independent with everything else, we have that

$$\mathbb{E}[s_{jt}|z_t] = \mathbb{E}[\pi_{jt}|z_t]$$

Therefore, our first stage procedure is trying to estimate this unknown function $\mathbb{E}[\pi_{jt}|z_t]$ using our deep market share model. We describe the detailed steps in the next section. In fact, if the goal is purely prediction for market shares of certain products in certain markets instead of recovering causal parameters of interests, we could just stop here at the first stage. If we care about consistent estimation of causal parameters of interests, then there is more to be done: we explain later how we could use the estimate of $\mathbb{E}[\pi_{jt}|z_t]$ from first stage to correct for the selection bias and get consistent estimate of causal parameters in stage 2. Intuitively, products (j,t) with very high $\mathbb{E}[\pi_{jt}|z_t]$ would be less affected by selection bias since if we are more certain that this product will have positive sales for extremely high probability. On the contrary, if $\mathbb{E}[\pi_{jt}|z_t]$ is very low for a particular product market pair (j,t), then we know that this data point (j,t) will be problematic and contributing to biases in the final

estimates since whether it will make it to the selected subset () for estimation or not depends a lot on whether the random draw of ξ_{jt} is large or not. That means that the selections on unobservables ξ_{jt} are much more severe for products with very low $\mathbb{E}[\pi_{jt}|z_t]$ compared with products with very high $\mathbb{E}[\pi_{jt}|z_t]$. And those (j,t)'s with very small $\mathbb{E}[\pi_{jt}|z_t]$ are the ones we need to pay attention to if we want to correct for the selection bias in the naive BLP procedure that throws away products with 0 sale or small sales. Note that when evaluating the expectation $\mathbb{E}[\pi_{jt}|z_t]$, the only randomness is ξ conditional on z. We assume instrument variables researchers have found are independent with unobservables ξ and thus the expectation is just with respect to the unknown joint distribution of all ξ's.

2.4.1.2 Detailed procedure

We will be training a deep market share model to predict market shares s_{jt} (including zero shares) using instrument z_t (from all products in the same market t). We model the shares based on Equation (2.10). We set $n = 100$ and for each f_i, we only use 1 layer neuron network with linear activation function. That means that we are modelling all f_i's, $i = 1, 2, ..., 100$ in Equation (2.10) as $f_i(z_{jt}) = w_i' z_{jt}$. We use stochastic gradient descent as optimizer to fit the model to observed market shares of all products (including 0-sale products). Denote our estimated version from deep market share by $\hat{f}(z_t)$ or simply $\hat{s}$. The true function we want to approximate is denoted by $f(z_t) = \mathbb{E}[\pi_t|z_t]$. Note that f or $\hat{f}$ is a mapping from $\mathbb{R}^{J_t}$ to $\mathbb{R}^{J_t}$, where J_t can vary across different t. Both $\hat{f}$ and f are function of z_t since they depend on information from all products in the same market. We use subscript of $\hat{f}$ and f to denote the product market pair (j,t), i.e., $f_{jt}(z_t) = \mathbb{E}[\pi_{jt}|z_t]$ for any t and any $j = 1, 2, ..., J_t$.

We next describe a procedure often called cross-fitting due to its similarity to cross-validation. We randomly partition the entire data set denoted by I into two

folds based on markets so that roughly $\frac{T}{2}$ markets of data are in the first half, denoted as I_1 and same number of markets in the second half, denoted as I_2. Essentially, we want to record out-of-fold predicted market shares. We train one machine learning model on the first half using instruments z_t to predict market shares s_{jt} and denote the model fitted there by $\hat{f}^{I_1}(z_t)$, where the superscript I_1 indicates that the model is trained on data I_1. Note that we use $\setminus$ to denote the set difference operation and in our case $I_1 = I \setminus I_2$. Then we make out-of-fold predictions on I_2 to get $\hat{f}^{I_1}(z_t)$ for all $(j,t) \in I_2$ using $\hat{f}^{I_1}$ trained on $I \setminus I_2 = I_1$. Similarly we train another deep market share model using data from I_2 and make predictions $\hat{f}^{I_2}_{jt}(z_t)$ for all $(j,t) \in I_1$. We use l to denote the fold assignment mapping, i.e., $l(j,t) = I_1$ if and only if $(j,t) \in I_1$ and similarly $l(j,t) = I_2$ if and only if $(j,t) \in I_2$. Then we can unify the notations: for all $(j,t) \in I$, the out-of-fold predicted market shares are denoted by $\hat{f}^{I \setminus l(j,t)}_{jt}(z_t)$. $\hat{f}^{I \setminus l(j,t)}_{jt}(z_t)$ for all (j,t) are computed and stored as our final output from first stage, which will be used in the second stage to correct for selection bias.

Using our notations, a naive first stage will feed into second stage predictions $\hat{f}^{I}_{jt}(z_t)$ for $\forall (j,t) \in I$, where I is the entire data set. The difference between our $\hat{f}^{I \setminus l(j,t)}_{jt}(z_t)$ and the naive version $\hat{f}^{I}_{jt}(z_t)$ for $\forall (j,t) \in I$ is that when we make predictions from trained machine learning models on a point (j,t), we ensure through two-fold cross-fitting that the model has never seen this particular point (j,t) during its training process. The name cross-fitting is partially due to the similarity to cross-validation in machine learning. From [30] and the discussions therein, cross-fitting can help prevent the bias introduced by using machine learning methods to estimate nuisance parameters/functions from affecting our estimation of causal parameters of interests. In our particular case, the causal parameters we want to estimate is $\theta = (\beta', \lambda')'$ while the nuisance function we will estimate in this first stage is f. The intuition of cross-fitting is similar to cross-validation too. In cross-validation, the goal is to evaluate a fitted machine learning model $\hat{f}$ and if we do not perform cross-

validation and directly make predictions on data points our model have seen during training, the model performance will be biased and over-estimated. We thus perform cross-validation and use out-of-fold predictions to evaluate the performance of fitted model $\hat{f}$. In cross-fitting, the goal is not model evaluation. Instead, we want to take the model predictions in a second stage to fit another model. Again, we want to avoid biasing the results of the downstream fitting task in the second stage and thus we use out-of-fold predictions, $\hat{f}_{jt}^{I\setminus l(j,t)}(z_t)$, instead of $\hat{f}_{jt}^{I}(z_t)$. However, in our simulation setup, we do not really find much difference between feeding $\hat{f}_{jt}^{I}(z_t)$ vs. $\hat{f}_{jt}^{I\setminus l(j,t)}(z_t)$ for $\forall(j,t) \in I$ into the second stage. As a result, later in our simulation results session we only show the result without cross-fitting for simplicity.

Note that in the first stage nothing prevents from including 0's in the machine learning prediction task. We should incorporate the information from 0's into our final predictors $\hat{f}$.

We implement the first stage using a deep learning framework in Python called Keras.

2.4.2 Stage 2: estimation

On a high level, we will only focus on first forming the buckets based on $\hat{f}_{jt}^{I\setminus l(j,t)}(z_t)$ from first stage and then minimize our objective function which is the summation of violations of moment inequality over all buckets, subject to constraints that observed market shares equal to model predicted market shares, i.e.,

$$\sigma_j(\delta_t; x_t, \lambda) = s_{jt} \qquad t = 1, 2, ..., T \quad j = 1, 2, ..., J_t$$

We will cast our second stage estimation problem as MPEC formulation proposed by [83]. We will introduce objective function and constraints in different subsections below. The only major difference between our second stage and standard BLP estimator is the loss function we are minimizing.

2.4.2.1 Objective function

Our objective function is very similar to the one used in bound estimator from [51].

We introduce some notations first:

We first apply Laplace transform of the observed market shares similar to [51]. Define $\tilde{s}_{jt} = \frac{ns_{jt}+1}{n+J_t+1}$. This will ensure that actually we can carry out the market share inversion.

Next we introduce a conditional moment inequality from [51]. The following moment inequalities hold at the true value of $\theta_0 = (\beta_0, \lambda_0)$. We state the following Lemma from [51] without proof. For details and the proof see [51].

Lemma 2.1.

$$\mathbb{E}[\delta^u_{jt}(\lambda_0) - x'_{jt}\beta_0 | z_t] \geq 0 \geq \mathbb{E}[\delta^l_{jt}(\lambda_0) - x'_{jt}\beta_0 | z_t] \tag{2.12}$$

, where δ^u_{jt} and δ^l_{jt} are defined as

$$\delta^u_{jt}(\lambda) := \sigma_j^{-1}(\tilde{s}_t, x_t; \lambda) - \log\left(\frac{\tilde{s}_{jt}}{\tilde{s}_{0t}}\right) + \log\left(\frac{\tilde{s}_{jt}+\eta}{\tilde{s}_{0t}-\eta}\right) \tag{2.13}$$

$$\delta^l_{jt}(\lambda) := \sigma_j^{-1}(\tilde{s}_t, x_t; \lambda) - \log\left(\frac{\tilde{s}_{jt}}{\tilde{s}_{0t}}\right) + \log\left(\frac{\tilde{s}_{jt}-\eta}{\tilde{s}_{0t}+\eta}\right) \tag{2.14}$$

Same as bound estimator proposed by [51], our estimator's loss function is baesd on Equation (2.12).

Next we partition I and form buckets based on first stage output $\hat{f}^I_{jt}(z_t)$. We define a collection of intervals $\mathcal{C}$ as

$$\mathcal{C} = \{\mathcal{C}_{a,r} : \mathcal{C}_{a,r} = (\frac{(a-1)}{2r}, \frac{a}{2r}], \quad a \in \{1, 2, ..., 2r\}, r = r_0, r_0+1, ...\} \tag{2.15}$$

In practice we start r at $r_0 = 2$ and cap it at $r_T = 50$. For the value of η in the definition of δ^u_{jt} and δ^l_{jt}, we follow the recommendation by [51] and set $\eta = 10^{-6}$. We can then group points (j,t) into buckets using the collection of indicator functions $\hat{\mathcal{G}}$ defined as

$$\hat{\mathcal{G}} := \left\{ g_{jt}^{a,r} : g_{jt}^{a,r}(z_t) = 1_{\{\hat{f}_{jt}^{I}(z_t) \in \mathcal{C}_{a,r}\}} \text{ for any } \mathcal{C}_{a,r} \in \mathcal{C}, t = 1, 2, ..., T, j = 1, 2, ..., J_t \right\}$$

We utilize those inequalities to form our moment conditions for different buckets formed by indicator functions $g^{a,r}$'s in $\hat{\mathcal{G}}$:

$$\mathbb{E}[\left(\delta_{jt}^{u}(\lambda) - x_{jt}'\beta\right) z_{jt} g_{jt}^{a,r}(z_t)] \geq 0 \tag{2.16}$$

$$\mathbb{E}[\left(\delta_{jt}^{l}(\lambda) - x_{jt}'\beta\right) z_{jt} g_{jt}^{a,r}(z_t)] \leq 0 \tag{2.17}$$

Therefore our objective function would be minimizing over θ the sample analog of Inequality (2.16) and (2.17). Define

$$\bar{\rho}_T^{u}(\theta, g) = \frac{1}{T\bar{J}} \sum_{t=1}^{T} \left(\sum_{j=1}^{J_t} \left(\delta_{jt}^{u}(\lambda) - x_{jt}'\beta\right) z_{jt} g_{jt}(z_t) \right) \tag{2.18}$$

$$\bar{\rho}_T^{l}(\theta, g) = \frac{1}{T\bar{J}} \sum_{t=1}^{T} \left(\sum_{j=1}^{J_t} \left(x_{jt}'\beta - \delta_{jt}^{l}(\lambda)\right) z_{jt} g_{jt}(z_t) \right) \tag{2.19}$$

Note that since z_{jt} has dimension d_z, both $\bar{\rho}_T^{u}(\theta, g)$ and $\bar{\rho}_T^{l}(\theta, g)$ have dimension d_z too. We define two loss functions below; one is feasible while the other is infeasible since it has unknown population true function in its definition.

$$\hat{Q}_T\left(\theta, \hat{\mathcal{G}}\right) = \sum_{g \in \hat{\mathcal{G}}} \left([\bar{\rho}_T^{u}(\theta, g)]_- \right)' [\bar{\rho}_T^{u}(\theta, g)]_- + \sum_{g \in \hat{\mathcal{G}}} \left(\left[\bar{\rho}_T^{l}(\theta, g)\right]_- \right)' \left[\bar{\rho}_T^{l}(\theta, g)\right]_- \tag{2.20}$$

,where $[x]_- := \max(-x, 0)$

Our objective function for second stage is then given below as

$$\min_{\theta} \hat{Q}_T\left(\theta, \hat{\mathcal{G}}\right) \tag{2.21}$$

Note that $\hat{Q}_T$ above depends on $(\theta, \hat{\mathcal{G}})$ where θ is the value of causal parameters of interest, z is the instruments empirical research decides to use and lastly $\hat{\mathcal{G}}$ is

the set of nuisance functions obtained with the help of first stage estimates. By the definition of this objective function, we implicitly assume equal weighting between different moment conditions and between different values of a, r for the g functions. We opt for equal weighting for simplicity but there is nothing preventing us from using the same weighting function as proposed in for example [51] and [5].

The main difference between our second stage and bound estimator in [51] is that we have a first stage estimating $f_{jt}(z_t) = \mathbb{E}[\pi_{jt}|z_t]$ which is then used to form buckets or hypercubes $\hat{\mathcal{G}}$ while in bound estimator, there are no first stage and the hypercubes are formed in a brute force way. This difference gives our estimators advantage as the dimension of instrument z_{jt} grows. Since in the framework of [51], even if we just have a moderate number of instruments, we will encounter the so-called curse of dimensionality problem and all hypercubes will have either 1 or 0 number of data point in its. We basically estimate a 1-dimensional quantity, the predicted market share from the first stage and then use this quantity to form hypercubes or buckets in the second stage. The advantage of two-stage estimator over bound estimator by [51] is very similar to the advantage of propensity score matching over matching by covariates.

Our two-stage bound estimator essentially combines the idea of matching by propensity score and bound estimator by [51] to improve the performance when the dimensionality d_z is large. The basic ideas behind the moment inequalities in (2.12) and the resulting loss function (2.21) are the same as described in [51]. We briefly recall the intuition here. We assume in the assumptions that there exists a subset of product market pair that we consider as "safe set" in the sense that if we focus our estimation on that subset alone, we will be able to recover causal parameters consistently even if we have to throw away products with zero sale. The problem lies in the rest "unsafe" products which are heavily affected by selection bias. The idea of the upper and lower bounds in expression (2.12) is that for safe products, both upper

and lower bounds become tighter and tighter as number of markets T goes to infinity and eventually the upper and lower bounds come together and we have a moment equality in the limit, instead of moment inequality. By assumptions, that moment equality holds true only when θ equals to the true value θ_0. That means that we are able to identify true parameter values only using information from safe set. On the other hand, since the unsafe products will bias the estimation results and thus we want to control their contributions to the loss function that we are minimizing. It turns out that, for unsafe products, both upper and lower bounds are satisfied for any values of θ within a neighborhood.

2.4.2.2 Computation

We use the MPEC formulation from [83]. Instead of actually inverting the function σ^{-1} using contraction mapping as in the original BLP paper [20], we enforce the constraints that model predicted shares are equal to empirical shares for better numerical performance.

$$\sigma(\delta_t; x_t, \lambda) = \tilde{s}_{jt}$$

Note that here we are using the Laplace shares $\tilde{s}_{jt}$. This part is exactly the same as the MPEC implementation of BLP estimator.

To sum up, in the second stage, we are solving the following optimization problem,

$$\min_{\theta} \hat{Q}_T\left(\theta, \hat{\mathcal{G}}\right)$$
$$\text{s.t.} \quad \sigma\left(\delta_t; x_t, \lambda\right) = \tilde{s}_t \text{ for any } t$$

We denote the solution of the above optimization problem as $\hat{\theta}^{TS}$ where the superscript stands for "Two-stage".

2.5 Two-Stage Weighting Estimator

Here, we propose another two-stage estimator that differs from the two-stage bound estimator described in section 2.4 only on the second stage. The weighting estimator relies on a different idea, reweighting, which is different from the moment inequality based arguments used in bound estimator. We want to give higher weights to products that are less vulnerable to selection bias and vice versa. In our two-stage bound estimator, we use first stage estimates $\hat{f}_{jt}(z_t)$ to weight different (j,t)'s. In this section, we propose another two-stage estimator that directly assign weights in the second stage, which is much more intuitive than bound estimator and two-stage bound estimator. Also, it is very easy to implement, compared with two-stage bound estimator and the original bound estimator.

Since the first stage is exactly the same as described in section 2.4.1. We directly jump into the second stage.

2.5.1 Stage 2: reweighted BLP

The intuition of reweighting is that we want to assign lower weights to products subject to higher degrees of selection effects (similar to unsafe products in two-stage bound estimators) and use less information from them in our estimation of causal parameters.

We take $\hat{f}^I_{jt}(z_t)$ from first stage and form buckets based on the values of $\hat{f}^I_{jt}(z_t)$. To be precise, we rank product market pairs (j,t) based on $\hat{f}^I_{jt}(z_t)$ and form K number of buckets. During our simulation we set $K = 80$ and each of the 80 buckets has the same number of (j,t) pairs. We denote buckets by B_k, $k = 1, 2, ..., B$ where B_1 has products with the smallest predicted shares. For $\forall (j,t) \in B_k$, we assign them weights equal to

$$w_{jt} = \frac{\frac{1}{|B_k|}\sum_{(j',t')\in B_k : s_{j't'}>0} \hat{f}^I_{j't'}(z'_t)}{\frac{1}{|B_k|}\sum_{(j',t')\in B_k : s_{j't'}>0} s_{j't'}} \tag{2.22}$$

In English, the weight for a (j,t) pair is the ratio between average predicted market share and the average observed shares over products with positive sales within the same bucket. Another key step is where we apply those weights. Actually, we multiply observed shares s_{jt} by w_{jt} and apply BLP on the weighted shares $\dot{s_{jt}} := s_{jt} \cdot w_{jt}$. The procedure will be exactly the same as traditional BLP except for that we use first stage estimates to reweight the empirically observed market shares before putting it in the BLP estimation. Following MPEC implementation, the optimization problem in the second stage of the proposed two-stage weighting estimator is then given by,

$$\min_{\beta \in \mathbb{R}^p, \lambda \in \mathbb{R}^p, \delta \in \mathbb{R}^J} \sum_{t=1}^{T} \sum_{j=1}^{J_t} \left(z_{jt} \left(\delta_{jt} - x'_{jt}\beta \right) \right)' W \left(z_{jt} \left(\delta_{jt} - x'_{jt}\beta \right) \right) \tag{2.23}$$

$$\text{s.t.} \quad \sigma\left(\delta_t; x_t, \lambda\right) = \dot{s}_t \qquad \forall t = 1, 2, ..., T \tag{2.24}$$

Comparing between the optimization problem above and the MPEC formulation for standard BLP estimation, the objective function (2.23) is exactly the same as (2.6). The only difference is in the constraints. In the right hand side of (2.24) we use the reweighted version of observed market shares $\dot{s}_{jt}$ while in (2.7) the original observed shares s_{jt} are used.

We explain the intuition of the second stage next. For product market pair (j,t) subject to low degrees of selection, we have $w_{jt} \approx 1$ and thus everything is the same as in standard BLP for those products. However, for "unsafe" products we will be having $w_{jt} < 1$ and the reweighting will push s_{jt} down towards $\mathbb{E}[\pi_{jt}|z_t]$, which reduces its contribution to moment conditions in the estimation. We show the promising simulation results of two-stage reweighting estimator in the next section. The inference part is still work in progress.

2.6 Simulation

In this section we presents results from our simulation experiments and showcase our proposed estimators' performance compared with existing ones. Note that the results in this section is preliminary and might change later in the actual paper.

2.6.1 Setup

We assume that there are in total $T = 25$ markets and for any market t, there are $n_t = 10000$ customers. The number of products in any market t is $J_t = 50$. Utility of customer i purchasing j in market t is given below,

$$U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$$

, where ϵ_{ijt} is the standard logit error term, superscript $X_{jt}^{(m)}$ denote the m^{th} covariate for product market pair (j, t). Note that we have in total 5 observed features and the unobserved characteristics are generated as $\xi_{jt} \sim \mathcal{N}(0, 0.5)$. All $\beta^{(m)}$'s are independent normal random variables with mean value 1 and standard deviation 0.5. That is, $\beta_{im} \sim \mathcal{N}(1, 0.5)$, $\forall m$.

Among the five covariates, the first one $X_{jt}^{(1)}$ is called vertical feature where it is generated as $X_{jt}^{(1)} = \frac{j}{10} + \mathcal{N}(0, 1)$, $j = 1, 2, ..., 50$. We call it vertical since there are parts that has consistent values between markets in addition to some noisy making different markets' offering different. The rest four is called non-vertical features and they are generated by $X_{jt}^{(m)} = \mathcal{N}(0, 1)$, $j = 1, 2, ..., 50$, $m = 2, 3, 4, 5$. All features can be endogenous, but for simplicity here we assume all covariates are exogenous, i.e., $X_{jt} \perp \xi_{jt}$. For traditional BLP throwing away products with 0 sales and our proposed two-stage weighting estimator, instruments we will be using are given by $(X_{jt}, X_{jt}^2, X_{jt}^3 - 3X_{jt})$, which follows the choices in [51]. For bound estimator and our two-stage bound estimator, we just use X_{jt} as instruments without applying further non-linear transformation.

We generate the data using the same setup 1000 times and try BLP, bound estimator and our proposed two-stage estimators on all 1000 samples and compute the average and standard deviation of the estimates. We show simulation results for two cases: $\beta_0 = -10$ and $\beta_0 = -12$, which yield different percentage of zero sales products (which is calculated as the average number of zero sale products across all 1000 simulation runs).

2.6.2 Results

We present the simulation results for $\beta_0 = -10$ in Table 2.3 below. Results for $\beta_0 = -12$ are in Table 2.4.

Avg % zeros =26.6%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
$\bar{\beta}_1$	1.000	0.830	0.890	0.982	0.982
		(0.003)	(0.002)	(0.007)	(0.003)
$\bar{\beta}_2$	1.000	0.895	0.885	0.967	0.988
		(0.003)	(0.002)	(0.010)	(0.003)
λ_1	0.500	0.670	0.451	0.482	0.530
		(0.005)	(0.002)	(0.009)	(0.004)
λ_2	0.500	0.492	0.446	0.459	0.496
		(0.005)	(0.005)	(0.017)	(0.004)

Table 2.3: Simulation Results from 1000 Runs for $\beta_0 = -10$ and $\sigma_\xi = 0.5$

We only show the mean and variance estimate for the first two covariates to illustrate the point. The first covariate is the vertical feature. We show the mean estimates $\bar{\beta}$ and standard deviation estimates λ for the first two random coefficients. The column "True" shows the true value of the parameters and "BLP" column corresponds to standard BLP estimator where we throw away 0's. The column "Gandhi et al." corresponds to the bound estimator from [51]. The column "Two-stage Bound Estimator" shows the performance for two-stage bound estimator explained in section

Avg % zeros =41.8%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
$\bar{\beta}_1$	1.000	0.522	0.856	0.954	0.933
		(0.009)	(0.003)	(0.011)	(0.006)
$\bar{\beta}_2$	1.000	0.858	0.847	0.938	0.980
		(0.003)	(0.003)	(0.012)	(0.003)
λ_1	0.500	0.791	0.430	0.463	0.550
		(0.010)	(0.003)	(0.008)	(0.006)
λ_2	0.500	0.472	0.433	0.417	0.494
		(0.007)	(0.006)	(0.019)	(0.005)

Table 2.4: Simulation Results from 1000 Runs for $\beta_0 = -12$ and $\sigma_\xi = 0.5$

2.4 and the last column "Two-stage Reweighting Estimator" presents the result of two-stage reweighting estimator described in section 2.5. We repeat 1000 times for the same simulation setup and the estimates reported in the table are averaged across all 1000 simulation runs. All numbers in parenthesis are the corresponding standard errors, which are calculated from different simulation runs as well.

We can see that consistent with what [51] found, the standard BLP estimator that leaves out the 0's sale products will have very severe bias, especially for β on the vertical feature. The bias is larger when the long-tail pattern is more severe, comparing the "BLP" column from Table 2.3 and 2.4. Interestingly, also consistent with [51], we found that the heterogeneity of vertical attributes (λ) will be biased upwards while the magnitude of mean value of β is biased lower.

Note that the magnitude of the bias on $\bar{\beta}_1$ for $\beta_0 = -12$ case is huge for BLP estimator: it is almost half of the true value 1. Even with just 5 covariates, the bound estimator does not perform very well: the estimate of $\bar{\beta}_1$, which is 0.856, improves upon BLP estimates by a lot but is still 15% lower than the true value 1. More importantly, both $\bar{\beta}_2$ and λ_2 are even a bit worse than BLP estimates. Both our proposals, two-stage bound estimator and two-stage reweighting estimator,

perform well in this simulation experiment in Table 2.3 and 2.4. They have less bias than bound estimator by [51] in almost all 4 parameters presented here. In particular, our two-stage weighting estimator looks very promising and performs the best especially on estimating λ's. It is also much simpler to implement than two-stage bound estimator and the original bound estimator.

Considering how common long-tail pattern is in empirical applications in various fields such as operations management, economics, and marketing, and how severe the bias can be, it is important for researchers to correct for the selection bias in order to recover true customer demand and make optimal managerial decisions.

2.7 Conclusion

We study classic demand estimation problem in long-tail data where direct application of BLP framework leads to substantial bias in the estimates. That biased estimates in turn could lead to sub-optimal decisions. We propose two different two-stage estimators where machine learning algorithms are used in the first stage in order to correct for the selection bias in standard BLP procedure. Our two-stage bound estimator and two stage weighting estimator combines machine learning tools and bound estimators proposed by [51]. It improves the performance of the original bound estimator especially when the number of covaraites or instruments is larger than 5. Our two-stage reweighting estimator tackles the problem in a completely different way: we apply weighting to adjust the observed data so that directly applying BLP on the reweighted data will give us correct estimates. Given how common long-tail pattern is nowadays, our proposals could improve data-driven decision making in a lot of empirical exercises in areas such as operations, economics and marketing.

Part II

Financial Engineering and Machine Learning Models in Asset Pricing

Chapter 3

Buy Rough, Sell Smooth

3.1 Introduction

A recent line of research has found evidence that stock price volatility is *rough*, in the sense that the evolution of volatility is rougher than the paths of ordinary Brownian motion. The evidence for rough volatility comes from two sources: the time series behavior of realized volatility, and an empirical regularity of option-implied volatility at short maturities that turns out to be well explained by roughness. See [53], [13], [50], [16], and [37] for background and further references.

Rough models of stochastic volatility replace an ordinary Brownian motion driving the dynamics of volatility with a *fractional* Brownian motion (fBM). The fBM family, indexed by a single parameter, includes ordinary Brownian motion and also processes with smoother and rougher paths. Empirical estimates here and in [53] and [16] find parameter values smaller than 1/2 (the case of ordinary Brownian motion), corresponding to rougher paths. We refer to these estimates as measures of *realized* roughness.

By *implied* roughness, we mean estimates extracted from option prices. Implied volatilities from equity put options are ordinarily skewed, meaning that they are larger at lower strikes, particularly at short maturities. But the steepness of this skew typically falls quickly as the maturity extends — more quickly than predicted by most stochastic volatility models. Rough volatility models capture this feature. Stocks with greater realized roughness exhibit fast mean-reversion in volatility; stocks

with greater implied roughness exhibit a fast decay in their implied-volatility skew.

The implications of these empirical regularities have received little attention beyond option markets. In this article, we seek to shed light on the possible sources and consequences of rough volatility by studying a trading strategy that trades stocks — not options — based on roughness in volatility. We sort stocks based on measures of realized or implied roughness and analyze a strategy that goes long the roughest quintile and short the smoothest quantile. When sorted on implied roughness, the strategy earns excess returns of 6% or more, after controlling for standard factors. The strategy is profitable in 13 out of the 17 years in our sample, including 2007, 2008, and 2009. The strategy based on realized roughness earns somewhat lower returns and is less robust to standard controls.

These results have several implications. First, they show that roughness matters for stock returns and is not just a feature of option markets. Second, they point to potential differences between implied and realized roughness, though in theory the two should coincide. Third, we will argue that the profitability of our implied rough-minus-smooth strategy reflects compensation for near-term idiosyncratic event risk. The fast decay in the implied volatility skew associated with implied roughness indicates near-term downside uncertainty that will be resolved quickly. We support this interpretation by examining the performance of our strategy near two types of events: our strategy earns higher returns near earnings announcements (which mainly resolve company-specific uncertainty) and lower returns near interest rate announcements by the Federal Reserve (which resolve market-wide uncertainty).

Efforts to date to model an underlying source of roughness have focused on market microstructure and the splitting of large orders, particularly [38], [62]. However, these models do not offer clear predictions on what types of stocks should exhibit greater roughness, which limits their application to our setting. Nevertheless, we investigate possible connections between roughness and market liquidity. We confirm a positive

association between roughness and illiquidity (which may be seen as consistent with [62]); but we also find that controlling for illiquidity reduces but does not eliminate the profitability of our implied strategy. Moreover, this strategy is limited to stocks with significant options trading, and these are generally larger and more liquid stocks. The profitability of our strategy therefore cannot be explained by an illiquidity premium.

Our results present an interesting contrast to the work of [91]. They find that a steep skew (corresponding to expensive puts at low strikes) forecasts negative earnings surprises, a finding we confirm in more recent data. This pattern supports a strategy of buying stocks with lower skews and selling stocks with steeper skews. One might expect stocks with a faster skew decay (greater implied roughness) to start with a steeper skew, in which case the strategy of [91] would lead to selling rough and buying smooth, just the opposite of the strategy we find profitable. Moreover, we find that roughness does not forecast earnings surprises, reinforcing the notion that the profitability of rough-minus-smooth reflects compensation for risk rather than cash flow predictability. Together these patterns indicate that the information in roughness is distinct from the skewness measure in [91].

Section 3.2 provides background on realized and implied roughness, and it explains the procedures we use to estimate both quantities. In Section 3.3, we evaluate the performance of strategies that buy the roughest quintile of stocks and short the smoothest quintile of stocks each month. We evaluate strategies using realized and implied measures of roughness, after controlling for standard factors. In Section 3.4, we control for additional factors through double sorts that hedge out other effects, including several measures of illiquidity and the levels of implied volatility and skewness. We find that returns on the implied strategy are robust to these controls. We also test robustness to these controls using [42] time-series averages of cross-sectional regressions. In Section 3.5, we find that the performance of our strategy is enhanced when restricted to stocks with earnings announcements in the subsequent month and

diminished near Federal Reserve announcements. We interpret these findings as evidence that rougher stocks (particularly as measured by implied roughness) are those facing near-term downside uncertainty.

3.2 Realized and Implied Roughness

3.2.1 Realized Roughness

To discuss roughness, we first recall the definition of fractional Brownian motion; for additional background, see [67] and Section 7.2 of [79]. A fractional Brownian motion with Hurst parameter $H \in (0,1)$ is a mean-zero Gaussian process $\{W_t^H, -\infty < t < \infty\}$ with stationary increments and covariance function given by

$$\mathbb{E}[W^H(t)W^H(s)] = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right). \tag{3.1}$$

The case $H = 1/2$ corresponds to ordinary Brownian motion. With $H \in (1/2, 1)$, fractional Brownian motion exhibits long-range dependence; processes with $H \in (0, 1/2)$ have paths that are rougher than those of ordinary Brownian motion, with small H indicating greater roughness.

As one indication of greater roughness, we have the following property of the moments of the increments of fractional Brownian motion. For any $t \in \mathbb{R}$, and $\Delta \geq 0$, and any $q > 0$,

$$\mathbb{E}[|W_{t+\Delta}^H - W_t^H|^q] = \mathbb{E}[|Z|^q]\Delta^{qH}, \quad Z \sim N(0,1). \tag{3.2}$$

With smaller H, increments over a short interval Δ have larger moments.

As an example of a rough volatility model for an asset price $\{S_t, t \geq 0\}$, we could set

$$d\log S_t = \mu\, dt + \sigma_t\, dW_t \tag{3.3}$$

$$d\log \sigma_t = \nu\, dW_t^H; \tag{3.4}$$

this is a special case of a single-factor version of what [53] call the rough Bergomi model, after [17]. More generally, the model specifies a mean-reverting log volatility process

$$d \log \sigma_t = -\kappa(\log \sigma_t - m) + \nu \, dW_t^H. \tag{3.5}$$

Here, μ, κ, m, and ν are constants, W is an ordinary Brownian motion, W^H is a fractional Brownian motion with $H \in (0, 1/2)$, and W and W^H may be correlated. The parameter H determines the roughness of the volatility process.

Empirical evidence for roughness in the time series of volatility can be found in [53], [16], [64], and later in this chapter. [2] present an approximation method for rough volatility models that suggests a simple interpretation: rough volatility arises from mixing mean-reverting volatility processes with different speeds of mean reversion, driven by an ordinary Brownian motion, including components with arbitrarily fast mean reversion. The connection between roughness and fast mean reversion is also supported by the analysis of option prices in [52].

If we could observe $\log \sigma_t$ at times $t = 0, \Delta, 2\Delta, \ldots$ for some small $\Delta > 0$, we could estimate H by estimating

$$\mathbb{E}[|\log \sigma_{t+\Delta} - \log \sigma_t|^q] \tag{3.6}$$

for various values of $q > 0$, and then applying (3.2) to extract H. This is the method of [53], which they apply more generally to estimate roughness, without necessarily assuming the specific model in (3.3)–(3.4) or (3.5).

In practice, σ_t cannot be observed and must be estimated, so we proceed as follows. Using trades from the Trade and Quote (TAQ) data, we apply the realized kernel method of [12] to estimate the daily integrated variance of returns; taking the square root yields our estimated daily volatility.[1] We obtained similar results using

[1]We use the non-flat Parzen kernel as implemented in Kevin Sheppard's toolbox at https://www.kevinsheppard.com/MFE_Toolbox.

the realized variance of 5-minute returns, but the realized kernel method is designed to be less sensitive to microstructure noise.

The rest of the estimation procedure works with these daily volatilities, which we write as $\hat{\sigma}_d$, with d indexing days. We apply (3.6) with $q = 2$, estimating second moments over intervals of ℓ days, $\ell = 1, 2, \ldots, 10$. In each month, for each stock and each lag ℓ, we calculate

$$z_2(\ell) = \frac{1}{T-\ell} \sum_{d=1}^{T-\ell} (\log \hat{\sigma}_{d+\ell} - \log \hat{\sigma}_d)^2, \tag{3.7}$$

where T is the number of days in the month. Based on (3.2), we expect

$$z_2(\ell) \approx \nu^2 \ell^{2H}.$$

We therefore run a regression

$$\log z_2(\ell) = \beta_1 + \beta_2 \log \ell + \epsilon, \tag{3.8}$$

to estimate H as $\beta_2/2$. We also estimate the volatility of volatility ν by setting $\log \nu = \beta_1/2$. This procedure yields an estimate of H (and ν) for each stock in each month.

[53] estimate (3.7) and (3.8) for moments of several orders q and then run a regression of the slope in (3.8) against q. We find that using several moments rather than just $q = 2$ leads to very similar estimates of H. [2]

3.2.2 Implied Roughness

By implied roughness we mean the value of H obtained by fitting option prices to a rough volatility model.

A conventional approach to evaluating an implied parameter would proceed as follows. Choose a specific model with some free parameters — in this case, a rough

[2] In tests of alternative estimation methods on simulated data, for which we know H, we have found that the main source of error is the estimation of the daily integrated variances $\hat{\sigma}_d^2$ from intraday returns, rather than the estimation of H from the daily volatilities.

volatility model; find the parameters that bring the model's option prices closest to a set of market prices.

Applying this approach to extract H from option prices raises two issues. The first is a practical consideration: pricing options in rough volatility models requires Monte Carlo simulation, so inverting prices to evaluate H for hundreds of stocks and months is computationally daunting. The second issue is more fundamental: a misspecified model may lead to an incorrect value of H, even if the "true" volatility process in rough.

To circumvent these issues, we follow a simpler and more robust approach, based on the term structure of the at-the-money (ATM) skew. Write $\sigma_{BS}(k,\tau)$ for the Black-Scholes implied volatility of an option with time-to-maturity τ and log-moneyness $k = \log(K/S)$, where K is the option's strike price and S is the current level of the underlying. The ATM skew at maturity τ is given by

$$\phi(\tau) = \left| \frac{\partial \sigma_{BS}(k,\tau)}{\partial k} \right|_{k=0} \tag{3.9}$$

An empirical regularity of the ATM skew is that it flattens at longer maturities. This pattern is illustrated in Figure 3.1, which shows fitted implied volatilities for JPMorgan Chase on June 5, 2012, using data from OptionMetrics. (We discuss the details of the fitting procedure below.) The horizontal axis shows the ratio of the strike price to the current stock price, so the ATM skew is the slope at a ratio of 1. The different curves correspond to different maturities. The slope is steepest (most negative) at the shortest maturity of three days and quickly flattens as we move to longer maturities.

The expansions of [49], [13], [37], and [44] characterize the rate of decay of the ATM skew for a very broad range of rough volatility models. These results (in particular as in [49]) show that the ATM skew admits an approximation of the form

$$\phi(\tau) \approx \text{constant} \times \tau^{H-1/2}, \quad \text{as } \tau \downarrow 0. \tag{3.10}$$

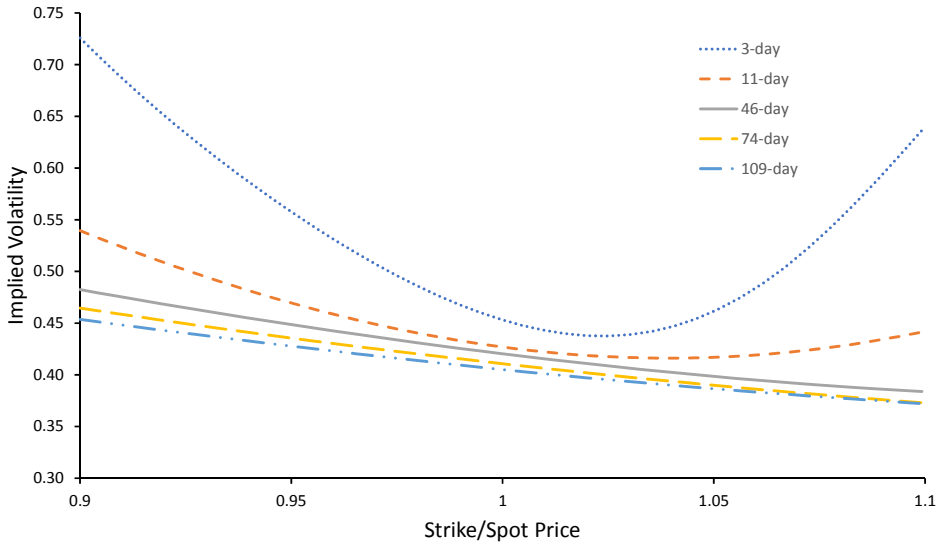

Figure 3.1: JPM implied volatilities on June 5, 2012. The curves show cubic spline fits at various maturities using raw data from OptionMetrics, plotted against the ratio of the put strike K to the spot price S. The ATM skew is the slope at $K/S = 1$. Its absolute value falls quickly as the maturity increases.

In other words, the ATM skew exhibits a power law decay at short maturities, with an exponent determined by H.

This idea is illustrated in Figure 3.2, which replicates similar figures in [53]. The horizontal axis records time-to-maturity τ, and the vertical axis records ATM skew $\phi(\tau)$. Each dot in the figure shows an estimate of $\phi(\tau)$, all calculated on September 15, 2005 (left panel) or June 20, 2013 (right panel), based on OptionMetrics data. The smooth curve in the figure shows a power law fit to the data, from which we estimate the exponent. In this example, the exponents are -0.48 (left) and -0.452 (right) corresponding to $H = .02$ and $H = 0.048$, respectively.

This is the approach we will use to calculate an option-implied value of H, after providing details of the calculation. The method is easy to use and readily lends itself to evaluating an implied H for hundreds of stocks, each day for nearly 20 years. The method is robust because it exploits the general property of rough volatility models in (3.9) rather than the detailed structure of a specific model.

Some may object to using the rate of decay of the ATM skew to extract an implied

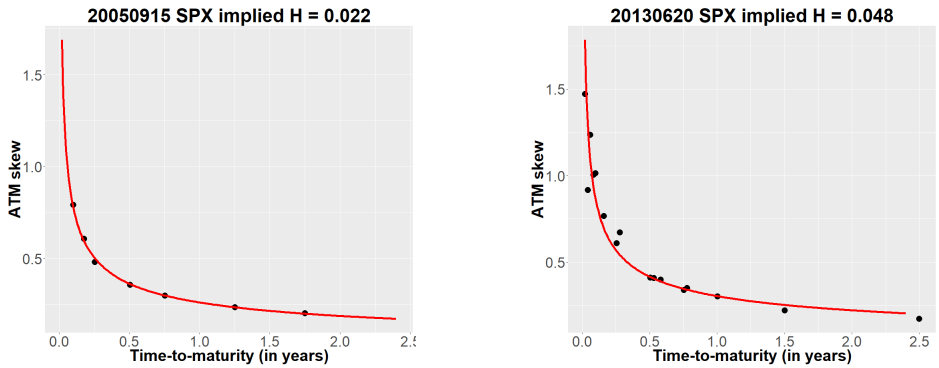

Figure 3.2: Term structure of the ATM skew for the S&P 500 index, as in similar figures in [53]. The charts plot the slope of the ATM skew against option maturity on Sep 15, 2005 (left) and Jun 20, 2013 (right), using OptionMetrics data.

measure of roughness on the grounds that certain stochastic volatility models driven by ordinary Brownian motion may also be able to fit the term structure of ϕ. For example, [18] fit what appears to be a power law decay using a linear combination of two exponentials. Some might prefer to follow the more conventional approach with which we began this section, fitting a specific model to market prices and finding the value of H that fits best. But if the model fits option prices well, *that approach will lead to the same value of* H because if the model fits the data, then the market prices satisfy (3.10). Using (3.10) directly is simply a more efficient and more robust way of arriving at the implied H. Calling it implied roughness is also much simpler than calling it the rate of decay of the ATM skew (plus 1/2).

To carry out this approach, we proceed as follows. First, we merge CRSP and OptionMetrics data to link stock prices and option prices. Next, we filter out options following standard rules in the literature; these are detailed in the appendix. On each day for each stock, using only the filtered data, we use a cubic spline to fit implied volatility as a function of $\log(K/S)$. We take the derivative of the spline at $\log(K/S) = 0$ as the ATM skew $\phi(\tau)$. Then we run a regression

$$\log \phi(\tau) = c + (H - 1/2) \log \tau + \epsilon;$$

that is, we add 1/2 to the estimated slope in this regression to evaluate the implied

H.

In addition to the realized measure discussed in Section 3.2.1 and the implied measure discussed here, we have tested a third measure — realized roughness of implied volatility, as in [64]. In this approach, for each stock we take the ATM implied volatility, and we evaluate the realized roughness (following (3.7)–(3.8)) from the stock's time series of implied volatility. We have found that investment results based on this measure are very similar to those using realized roughness, so we do not discuss them further.

3.2.3 Descriptive Statistics of Realized and Implied Roughness

Our focus is on the cross-sectional relationship between roughness and stock returns, so in Table 3.1 we present summary statistics on the cross-sectional variation of implied and realized roughness. In each month we calculate the mean, standard deviation and several quantiles (25%, 50%, 75%) of implied and realized roughness measures for all stocks; we then take the time-series average of these summary statistics and report them in the table.

As discussed in Section 3.2.2, we have values of implied roughness for only a subset of stock-month pairs. We refer to this subset as the "implied universe." In contrast, by the "full universe" we mean the larger set of stock-month pairs for which we have sufficient data to calculate a realized H and link TAQ, CRSP, and Compustat data. See the appendix for details on the filters applied.

In the last column of Table 3.1 we report summary statistics for realized roughness on the implied universe. The results in the table indicate that implied estimates of H are a bit larger than realized estimates and that this may be partly due to differences in the implied and realized universes, but the differences are small. [64] find that values of realized H estimated from the time series of implied volatility are generally

larger than values estimated from realized volatility, and they attribute the difference to a smoothing effect over an option's time to maturity. This effect may play some role in our estimates of implied H.

	Implied H	Realized H	Realized H on Implied Universe
avg Mean	0.18	0.07	0.09
avg S.D.	0.21	0.10	0.10
avg 25th pctile	0.06	0.00	0.02
avg median	0.18	0.06	0.08
avg 75th pctile	0.30	0.14	0.15

Table 3.1: Monthly averages of cross-sectional summary statistics. The last column shows statistics for realized H estimated from the subset of stocks for which implied estimates are available.

Table 3.2 reports time-series averages of cross-sectional means and standard deviations by industry, using industry classifications from Ken French's website.[3] The estimates are very consistent across different sectors.

[3]http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

	Implied H		Realized H		Realized H on Imp. Univ.	
Industry	Avg Mean	Avg S.D.	Avg Mean	Avg S.D.	Avg Mean	Avg S.D.
Consumer NonDurables	0.18	0.23	0.07	0.10	0.08	0.10
Consumer Durables	0.17	0.19	0.07	0.10	0.09	0.09
Manufacturing	0.17	0.19	0.08	0.10	0.09	0.10
Energy	0.20	0.20	0.08	0.10	0.09	0.09
Chemicals	0.19	0.19	0.08	0.10	0.09	0.09
Business Equipment	0.18	0.21	0.08	0.10	0.09	0.10
Telecom	0.19	0.22	0.08	0.10	0.09	0.10
Utilities	0.17	0.22	0.08	0.10	0.10	0.09
Shops	0.18	0.20	0.07	0.10	0.08	0.10
Health	0.17	0.23	0.07	0.10	0.09	0.10
Finance	0.18	0.19	0.07	0.10	0.10	0.10
Other	0.17	0.21	0.07	0.10	0.09	0.10

Table 3.2: Monthly averages of cross-sectional summary statistics by industry. The last two columns show statistics for realized H estimated from the subset of stocks for which implied estimates are available.

3.3 Sorted Portfolios

In this section, we test the performance of trading strategies that pick stocks based on realized or implied roughness. Each month, we sort stocks based on roughness (realized or implied) and group them into quintile portfolios. We evaluate the performance of a strategy that buys the roughest (smallest H) quintile and shorts the smoothest (largest H) quintile, holding these positions for one month. We calculate value-weighted returns in the month following the month in which portfolios are formed, and then repeat the procedure for the next month.

In addition to calculating average returns, we calculate excess returns (alphas) relative to various factor models: a single-factor (CAPM) model using the overall return of the market, net of the risk-free rate; the three-factor [40] model (with factors for the market, size, and book-to-market) augmented with a momentum factor, as in [25]; the five-factor model of [41] (with factors for the market, size, book-to-market, earnings robustness, and investment conservativeness), again augmented with momentum.

We use stock prices from CRSP, factor returns from Ken French's website, and option implied volatilities from OptionMetrics. The OptionMetrics data starts in 1996, but we start from 2000 because much more data is available after 2000 than in the earler years.

Table 3.3 shows results for stocks sorted on implied roughness. The columns show results for the quintile portfolios, sorted from smoothest (highest H) to roughest (lowest H). The last column shows results for the long-short strategy. The strategy earns an average monthly return of 0.49% (5.9% annually). Its alphas with respect to the various factor models range from 0.47% to 0.52% monthly, or 5.6% to 6.2% annually. The numbers in brackets are [72] t-statistics, and show that these excess returns are all statistically significant. Statistical significance at the 10%, 5% and 1% levels is indicated by *, **, and ***, respectively.

	1 Smooth	2	3	4	5 Rough	5-1
Mean	0.22	0.39	0.37	0.33	0.71	0.49
Std. Dev.	4.82	4.76	4.69	5.08	5.26	2.63
CAPM Alpha	-0.28**	-0.11	-0.13	-0.19	0.19	0.47**
	[-2.51]	[-1.48]	[-1.39]	[-1.35]	[1.37]	[2.43]
FF-3-MOM Alpha	-0.33***	-0.07	-0.07	-0.07	0.16	0.49***
	[-2.88]	[-0.94]	[-0.94]	[-0.64]	[1.22]	[2.63]
FF-5-MOM Alpha	-0.29***	-0.04	-0.04	0.03	0.24*	0.52***
	[-2.74]	[-0.58]	[-0.47]	[0.30]	[1.70]	[2.76]
Implied H	0.46	0.27	0.18	0.09	-0.11	
Size in billion $	14.64	18.87	19.07	15.79	7.84	
Book-to-Market	0.48	0.44	0.42	0.41	0.43	
Number of stocks	153	152	153	152	152	
Portfolio persistence	61%	75%	76%	74%	63%	

Table 3.3: Performance of portfolios sorted on implied roughness. Alphas are monthly values in percent. Numbers in brackets are t-statistics.

The lower half of Table 3.3 shows features of the quintile portfolios. By construction, the average implied H values decrease from left to right. The smoothest quintile has H close to the Brownian value of $1/2$, and the roughest quintile has a negative average H. A negative H is not meaningful as a Hurst parameter, but can certainly arise as an implied parameter through (3.10).

We see from Table 3.3 that the average book-to-market ratio is quite consistent across the quintiles, but size (measured by market cap) seems to be positively correlated with H, a point we will investigate further. The last row shows the percentage of stocks in each quintile that remain in the quintile from one month to the next.

Table 3.4 reports corresponding results using realized roughness. Panel A uses the full universe of CRSP stocks; Panel B limits the set of stocks used each month to the "implied universe," meaning those that pass the filters we use for the implied roughness portfolios in Table 3.4.

Both panels of Table 3.4 show that stocks with rougher volatility (smaller realized H) tend to outperform stocks with smoother volatility (larger realized H). Comparing

		1 Smooth	2	3	4	5 Rough	5-1
PANEL A	Mean	0.22	0.52	0.59	0.71	0.60	0.38
	Std. Dev.	4.99	4.69	4.65	4.33	4.94	2.66
	CAPM Alpha	-0.30***	0.02	0.09	0.24**	0.10	0.40**
		[-3.08]	[0.18]	[1.20]	[2.45]	[0.71]	[2.03]
	FF-3-MOM Alpha	-0.27***	0.00	0.05	0.24***	0.03	0.31
		[-2.97]	[0.00]	[0.61]	[2.71]	[0.22]	[1.56]
	FF-5-MOM Alpha	-0.19**	0.01	0.02	0.13	-0.01	0.17
		[-2.13]	[0.09]	[0.23]	[1.52]	[-0.10]	[0.97]
	Realized H	0.23	0.12	0.06	0.01	-0.05	
	Size in billion $	6.67	5.18	4.62	4.03	3.36	
	Book-to-Market	0.69	0.67	0.64	0.65	0.70	
	Number of stocks	611	613	614	614	614	
	Portfolio persistence	79%	80%	80%	80%	80%	
PANEL B	Mean	0.11	0.27	0.51	0.58	0.59	0.47
	Std. Dev.	5.28	4.89	4.85	4.58	4.89	2.89
	CAPM Alpha	-0.42***	-0.24**	0.00	0.10	0.09	0.51**
		[-3.55]	[-2.16]	[0.01]	[0.94]	[0.71]	[2.51]
	FF-3-MOM Alpha	-0.38***	-0.19*	0.01	0.15	0.13	0.51***
		[-3.49]	[-1.75]	[0.06]	[1.50]	[1.00]	[2.73]
	FF-5-MOM Alpha	-0.25**	-0.13	0.00	0.17	0.08	0.33*
		[-2.31]	[-1.19]	[-0.02]	[1.59]	[0.60]	[1.73]
	Realized H	0.24	0.14	0.08	0.03	-0.04	
	Size in billion $	18.38	16.37	15.31	13.69	12.18	
	Book-to-Market	0.44	0.43	0.43	0.43	0.44	
	Number of stocks	151	151	151	151	152	
	Portfolio persistence	80%	82%	82%	82%	81%	

Table 3.4: Performance of portfolios sorted on realized roughness. Alphas are monthly values in percent. Panel A shows results for all stocks and Panel B is limited to the stocks used in Table 3.3 for comparison.

the last column of Table 3.4 (showing performance of the rough-minus-smooth long-short strategy), with the last column of Table 3.3, indicates that the effect is not quite as strong and not quite as statistically significant sorting on realized as sorting on implied roughness. Portfolio persistence is a bit greater using realized roughness, indicating that this strategy has somewhat lower turnover.

Comparing Panels A and B of Table 3.4, we find that sorting on realized roughness yields higher alphas when we limit the universe of stocks to those for which we can

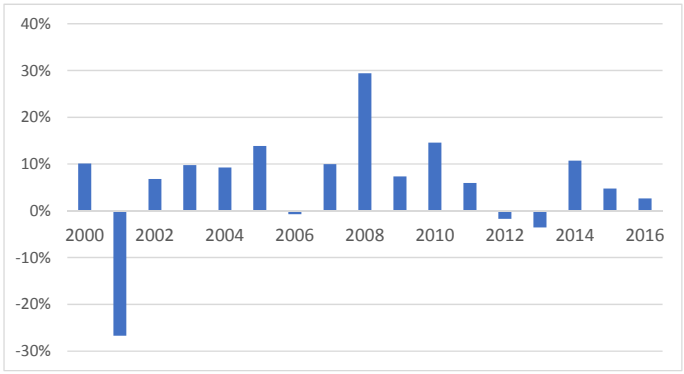

Figure 3.3: Annual performance of rough-minus-smooth strategy based on implied roughness.

also calculate implied roughness. This is surprising because the stocks in the more limited universe are larger on average and have lower book-to-market ratios; smaller stocks and high book-to-market stocks generally have higher expected returns. We see a similar effect in Table 3.3, where controlling for the Fama-French factors improves performance.

It is worth noting that in both panels of Table 3.4 the highest returns are generally associated with the fourth quintile of realized H rather than the fifth quintile. The performance of the realized strategy could be substantially improved by buying the fourth quintile, rather than the fifth, and shorting the first. For consistency and to avoid data snooping, we work exclusively with the original long-fifth, short-first strategy; however, this may underestimate the efficacy of trading on realized roughness.

Tables 3.3 and 3.4 show average performance over the full period 2000–2016. To illustrate how performance varies over time, Figure 3.3 shows annual performance by year for the implied strategy. Remarkably, sorting on implied roughness, the rough-minus-smooth strategy is profitable in 13 out of the 17 years, including 2007–2009; indeed, 2008 was the strategy's best year. The strategy's only large significant loss is in 2001, and the loss that year is almost entirely attributable to September, the month of the 9/11 attacks. We return to this point in Section 3.5.

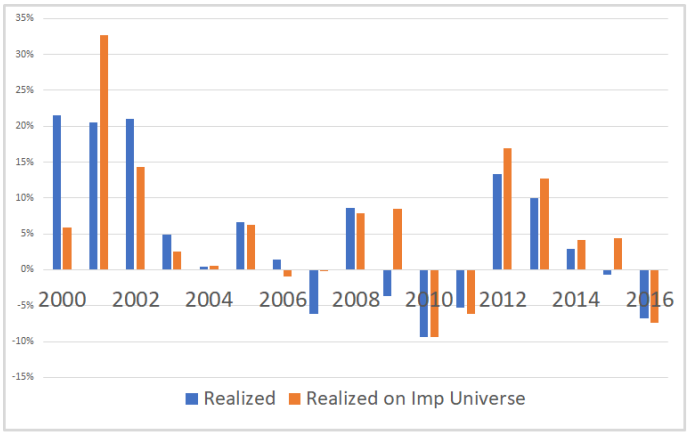

Figure 3.4: Annual performance of rough-minus-smooth strategy based on realized roughness, using all stocks or just the implied universe.

Figure 3.4 shows annual performance of the strategy based on realized roughness. The figure shows performance of the realized strategy on the full universe and on the implied universe. Except in the early part of the sample, where the option data is more limited, the realized strategy generally performs similarly on the full and restricted sets of stocks. This confirms that the performance in Figure 3.3 is not attributable to the set of stocks included in the implied universe. Indeed, comparing Figures 3.3 and 3.4 shows that the realized and implied strategies have done well at different times, suggesting that combining the two signals could lead to even better performance. However from Table 3.4 we see that the realized strategy provides a smaller FF-5-MOM alpha than the implied strategy. We will see in Section 3.4 that the realized strategy is also less robust to controls for other factors.

The performance of the implied strategy in 2008 raises the question of whether sorting on roughness implicitly tilts the long-short portfolio to favor some industries over others. For example, a strategy that shorts bank stocks would have performed well in 2008. However, we saw in Table 3.2 that roughness estimates are similar across industries. Moreover, the average implied and realized H estimates for finance companies in particular are in the middle of the ranges across industries, indicating that a rough-minus-smooth strategy does not tend to favor or disfavor financial stocks.

3.4 Controlling for Other Factors

To better understand the performance of the rough-minus-smooth strategies, in this section we add controls for additional factors. We first discuss factors that might influence performance and then evaluate their impact using two methods — double sorts and [42] regressions.

3.4.1 Liquidity

We observed previously that in Table 3.3 the average market cap across the five quintiles increases with H: rougher stocks tends to be smaller on average. This pattern suggests the possibility that roughness may reflect lower liquidity and therefore that a rough-minus-smooth strategy earns an illiquidity premium. This possibility is tempered by the fact that the stocks that pass the filters for calculating implied roughness are larger, on average, than those that do not. The question therefore requires a more systemic investigation.

A connection between realized roughness and liquidity was noted in an early version of [16], but it was removed from subsequent versions of that paper. [16] compare estimates of realized roughness with daily volume of trading in a stock.

In addition to trading volume, we consider the widely-used [4] illiquidity measure. The Amihud measure for a single stock in a single month sums the absolute values of the daily returns and divides the sum by the dollar volume for the month. Larger values of the Amihud measure are interpreted as indicating lower liquidity, whereas larger values of trading volume are associated with greater liquidity.

Figures 3.5 and 3.6 compare, respectively, realized and implied estimates of H with the log of the Amihud measure and log daily volume. Each dot in the figure corresponds to a single stock in a single month. Consistent with the earlier version of [16], we find a positive correlation (0.55) between realized H and log daily volume.

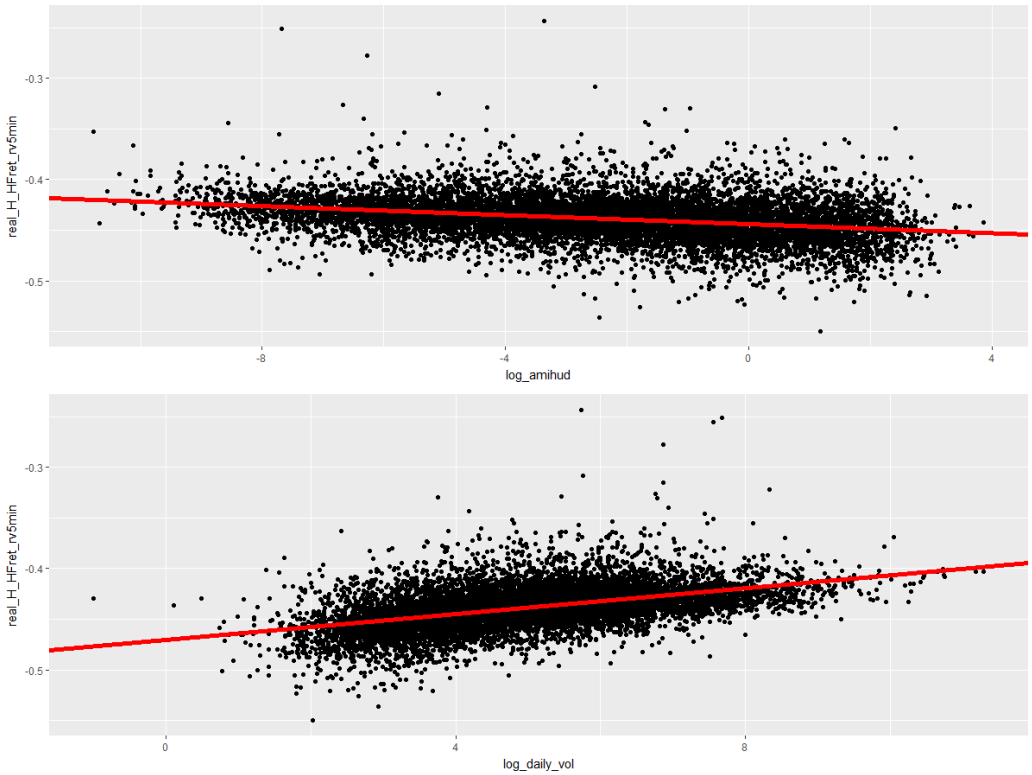

Figure 3.5: Realized roughness and liquidity. The figures plot realized roughness against the log of the Amihud illiquidity measure (top) and log daily volume (bottom). Each point shows a single stock in a single month.

Consistent with this pattern, we find a negative correlation (-0.46) between realized H and the log Amihud measure.

The results using implied roughness in Figure 3.6 are qualitatively similar but not as strong. The correlation between implied H and log daily volume is 0.28, and the correlation with the log Amihud measure is -0.40.

Beyond these empirical patterns, a potential link between roughness and liquidity is interesting because of efforts to explain realized roughness through market microstructure; see [38] and [62]. However, the explanations developed to date are highly stylized, and they do not make clear predictions about whether greater roughness should be associated the more or less liquidity.[4]

[4]According to Mathieu Rosenbaum (personal communication), [62] implies a longer transient price impact when H is smaller, which would be consistent with the correlations we find.

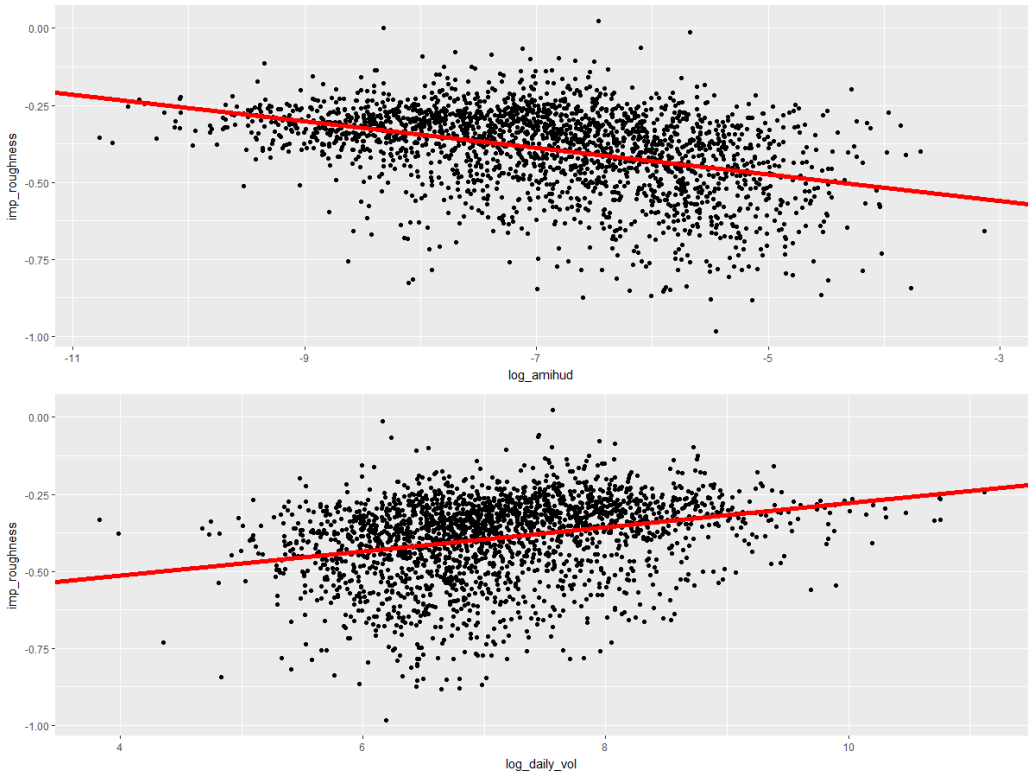

Figure 3.6: Implied roughness and liquidity. The figures plot implied roughness against the log of the Amihud illiquidity measure (top) and log daily volume (bottom). Each point shows a single stock in a single month.

3.4.2 Implied Volatility and Skewness

Implied roughness is a relatively complex feature of a stock's implied volatility surface, involving differences in implied volatilities across both strikes and maturities. To try to isolate the source of alpha in the implied rough-minus-smooth strategy, we will therefore control for more basic features — the level of the ATM implied volatility and the shape of implied volatility skew.

Several authors (particularly [34] and [91]) have documented predictability in stock returns using measures of implied volatility and skewness. A fast decay in the ATM skew (low implied H) is potentially associated with high degree of near-term skewness or implied volatility. We therefore control for these factors.

As our measure of ATM implied volatility, we use the implied volatility for a one-month call with strike closest to the spot price as reported in implied volatility

surface data set from OptionMetrics. We denote this by $\sigma_{1m}^{Call}(\frac{K}{S} = 1)$. Similar to [91], we use as our measure of implied volatility skew

$$\text{XZZ-skew} = \sigma_{1m}^{Put}(\frac{K}{S} = 0.9) - \sigma_{1m}^{Call}(\frac{K}{S} = 1), \tag{3.11}$$

the difference between the one-month implied volatility for a put with moneyness closest to 0.9 and the one-month implied volatility for a call with strike closest to the spot price.

[91] find that larger values of their skew measure predict lower stock returns in the cross section, a pattern that we find holds up as well using more recent data and a slightly different skew measure. Interestingly, this effect appears to run in the opposite direction of what we find using implied roughness. A smaller implied H indicates a faster decay of the ATM skew. If this indicates a higher initial value of the ATM skew, then the finding of [91] would suggest that stocks with smaller implied H have lower stock returns, yet we find exactly the opposite. This suggests that the performance of the rough-minus-smooth strategy is not explained by the XZZ-skew, a hypothesis we will check in the next sections.

3.4.3 Double Sorts

To control for factors like liquidity or skewness that might influence the returns on our roughness quintile portfolios, in this section, we apply a standard double-sorting procedure.

Suppose, for example, that we want to control for illiquidity, using the Amihud measure. For each month, we proceed as follows. We sort stocks into deciles according to the Amihud measure. Within each of these illiquidity deciles, we sort stocks by roughness (realized or implied). We then take the roughest quintile from each of the illiquidity deciles — this is our rough portfolio. Similarly, we form our smooth

portfolio by grouping all stocks that are in the smoothest quintile of any of the illiquidity deciles.

Under this construction, all levels of illiquidity are represented in the rough and smooth portfolios, so the performance of the rough-minus-smooth strategy should be unaffected by illiquidity: we have hedged out illiquidity. We sort into ten portfolios based on illiquidity in the first step in order to achieve a better balance of the conditioning factor between our controlled rough portfolio and smooth portfolio. The same procedure allows us to hedge out the effect of any other factor by first sorting on that factor.

We apply double sorts that condition on the following variables, one at a time:

- Average daily volume for each stock;
- The Amihud illiquidity measure;
- Turnover, measured as a stock's monthly trading volume divided by the average shares outstanding of that stock during the month;
- ATM implied volatility, as measured by the implied volatility for a 30-day option with strike closest to spot price;
- XZZ-skew, as defined in (3.11);
- Size (as measured by log market cap), book-to-market, and trailing 12-month return.

Table 3.5 shows the performance of the rough-minus-smooth strategy based on implied roughness after controlling for each of these factors through double sorts. The table shows average returns and alphas using either FF3-Mom or FF5-Mom factor models.

The first three rows of the table consider liquidity measures. Sorting first on average daily volume or the Amihud illiquidity measure reduces but does not eliminate

the profitability of the strategy. Some reduction in performance is to be expected, given the correlation we documented in Section 3.4.1 between implied roughness and these measures. But the profitability of the strategy remains significant, particularly as measured by alpha relative to the Fama-French 5-factor with momentum, ranging from 3.1% to 5.4% per year, depending on the measure used, with t-statistics ranging from 2.0 to 3.0. Controlling for turnover actually increases the mean return of the strategy, with average monthly returns of 0.54%, and increases the t-statistics to around 4.0. In short, liquidity by itself cannot account for the performance of the rough-minus-smooth strategy.

The next two rows of the table control for implied volatility and the ATM skew. Controlling for ATM implied volatility improves the average return and alphas to 7%, except for the FF5Mom alpha, which decreases a bit to 4.9% annually. Controlling for the XZZ-skew measure of [91] has only a small effect on the average return, alphas and t-statistics, and all alphas remain statistically significant. Thus, these well-known features of the implied volatility surface — the level of ATM volatility and skewness in implied volatility — cannot account for the performance of the rough-minus-smooth strategy.

The last three factors in the table serve as robustness checks. Sorting on size, book-to-market, and trailing returns may slightly reduce the performance of the strategy but does not eliminate — and may even strengthen — statistical significance.

Table 3.6 shows corresponding results based on realized roughness, using the full universe of stocks (Panel A) or the implied universe (Panel B). Here we find that controlling for liquidity (through average daily volume or the Amihud measure) removes the significance of returns and alphas of the rough-minus-smooth strategy. Controlling for size does as well in Panel A. These results suggest a strong association between realized roughness and illiquidity. In contrast, controlling for implied volatility and the ATM skew actually enhances the performance of the strategy. This further indi-

Conditioning Variable	Mean Return	CAPM Alpha	FF3Mom Alpha	FF5Mom Alpha
Average Daily Volume	0.23*	0.21	0.23*	0.26**
	[1.84]	[1.60]	[1.81]	[2.01]
Average Daily Amihud	0.45***	0.46***	0.46***	0.40***
	[3.23]	[3.31]	[3.21]	[2.65]
Turnover	0.54***	0.53***	0.52***	0.45***
	[3.96]	[3.90]	[3.53]	[2.98]
XZZ Skew	0.46***	0.44**	0.49***	0.45***
	[2.79]	[2.54]	[2.96]	[2.61]
ATM Implied Volatility	0.59***	0.59***	0.59***	0.41**
	[3.54]	[3.51]	[3.32]	[2.28]
Size	0.41***	0.42***	0.46***	0.42***
	[3.24]	[3.31]	[3.53]	[3.16]
Book-to-Market	0.34**	0.32**	0.36**	0.35**
	[2.44]	[2.17]	[2.57]	[2.34]
12-Month Return	0.45***	0.43***	0.43***	0.48***
	[3.29]	[3.06]	[3.17]	[3.38]

Table 3.5: Performance of rough-minus-smooth portfolios using implied roughness, constructed through double sorts on various factors, for the period Jan 2000 through Jun 2016. Mean return and alphas are monthly values in percent. Numbers in brackets are t-statistics based on Newey-West standard errors.

cates that the effect of roughness, whether realized or implied, is not already reflected in the ATM volatility or the ATM skew.

3.4.4 Fama-MacBeth Regressions

To further investigate whether the performance of the rough-minus-smooth strategy is explained by other factors, we run regressions based on the specification

$$Ret_{i,t} = b_{0t} + b_{1t}H_{i,t} + b'_{2t}CONTROLS_{i,t-1} + e_{i,t}, \tag{3.12}$$

where $Ret_{i,t}$ is the return of stock i in month t; $H_{i,t}$ is either realized or implied roughness of stock i in month t; $CONTROLS_{i,t-1}$ is a vector of controls; and the $e_{i,t}$ are error terms. We estimate coefficients and their standard errors through [42] regressions: in each month t, we run cross-sectional regressions to estimate b_{0t}, b_{1t}, and b_{2t}; we then take the time-series averages of these regression coefficients and use their time-series variation to estimate standard errors. Compared to the double sorts tested previously, these regressions have the advantage of allowing the simultaneous

Conditioning Variable	Mean Return	CAPM Alpha	FF3Mom Alpha	FF5Mom Alpha
PANEL A: Full Universe				
Average Daily Volume	0.14	0.15	0.14	0.05
	[1.47]	[1.63]	[1.51]	[0.60]
Average Daily Amihud	0.12	0.17	0.12	-0.05
	[0.94]	[1.37]	[1.03]	[-0.45]
Turnover	0.38***	0.38**	0.33**	0.20
	[2.60]	[2.49]	[2.34]	[1.46]
XZZ Skew	0.54***	0.57***	0.53***	0.30*
	[3.10]	[3.22]	[3.36]	[1.95]
ATM Implied Volatility	0.54***	0.56***	0.59***	0.41**
	[2.81]	[2.92]	[3.17]	[2.15]
Size	0.12	0.18	0.13	-0.04
	[0.90]	[1.47]	[1.27]	[-0.37]
Book-to-Market	0.35***	0.38***	0.35***	0.20
	[2.63]	[2.90]	[2.73]	[1.55]
12-Month Return	0.51***	0.54***	0.50***	0.37**
	[3.06]	[3.12]	[3.19]	[2.34]
PANEL B: Implied Universe				
Average Daily Volume	0.22	0.25	0.21	0.08
	[1.43]	[1.64]	[1.48]	[0.56]
Average Daily Amihud	0.40*	0.47**	0.41**	0.22
	[1.94]	[2.31]	[2.40]	[1.35]
Turnover	0.60***	0.61***	0.63***	0.53***
	[3.13]	[3.09]	[3.38]	[2.79]
XZZ Skew	0.49**	0.53**	0.51***	0.27
	[2.41]	[2.51]	[2.88]	[1.54]
ATM Implied Volatility	0.62***	0.63***	0.62***	0.43*
	[2.81]	[2.77]	[2.88]	[1.95]
Size	0.49**	0.55***	0.53***	0.35**
	[2.41]	[2.70]	[3.13]	[2.10]
Book-to-Market	0.31*	0.34**	0.33**	0.14
	[1.83]	[2.05]	[2.05]	[0.88]
12-Month Return	0.55***	0.60***	0.57***	0.41**
	[2.96]	[3.08]	[3.47]	[2.47]

Table 3.6: Performance of rough-minus-smooth portfolios using realized roughness, constructed through double sorts on various factors, for the period Jan 2000 through Jun 2016. Mean return and alphas are monthly values in percent. Numbers in brackets are t-statistics based on Newey-West standard errors.

inclusion of multiple controls, but they have the disadvantage of imposing linearity on the relationship between returns and controls.

An alternative approach would be to run a panel regression to estimate (3.12) with no dependence on t in the coefficients. Since we are mainly interested in the cross-sectional relationship between roughness and returns, we would include month fixed-effects; and since monthly returns have very low autocorrelation, we would estimate standard errors clustered by month, following [75]. However, as also discussed in [75], Section 3, Fama-MacBeth standard errors are more accurate than panel regressions with clustered standard errors under two conditions that are appropriate to our setting: (1) the main source of dependence in error terms comes from time effects (correlations in returns of different stocks in the same month); and (2) the number of time periods (201 months) is not very large compared with the number of stocks per month (up to 1108 stocks per month in the implied universe and 3577 per month for the full universe). The dependence in (1) is dealt with effectively by Fama-MacBeth regressions. The values in (2) would require the estimation of a very large covariance matrix between different stocks based on limited data in order to cluster by time. In light of these considerations, we use Fama-MacBeth regressions.

Table 3.7 shows the results. Panel A tests implied H; Panel B test realized H on the implied universe; and Panel C tests the realized H on the full universe of stocks. Each panel shows two regressions, one including only the corresponding roughness measure, and one including multiple controls. All explanatory variables have been standardized (cross-sectionally in each month) to make the coefficients comparable. Returns are in decimals, so a return of 5% is recorded as 0.05.

Panel A confirms the negative relationship between returns and implied H; including controls increases the magnitude and significance of the coefficient. Panel B shows that realized H has a significant relationship with returns when restricted to the implied universe, but this relationship is eliminated by the controls. In Panel C we

find no significant relationship between realized H and returns on the full universe of stocks, with or without controls. Interestingly, our results confirm a strong negative relationship between returns and the skewness measure of [91], while also showing in Panel A that this control does not explain the effectiveness of implied roughness.

Our controls include return volatility and implied volatility, so the regressions in Table 3.7 also control for the volatility risk premium ([27]) measured as the difference between implied and realized volatility. In particular, Panel A shows that the profitability of the implied strategy cannot be attributed to the volatility risk premium.

	PANEL A		PANEL B		PANEL C	
Variable	Reg 1	Reg 2	Reg 3	Reg 4	Reg 5	Reg 6
Intercept	0.0043	0.0046	0.0043	0.0046	0.0088*	0.0078
	[0.83]	[0.91]	[0.83]	[0.90]	[1.66]	[1.49]
Implied H	-0.0010**	-0.0014***				
	[-2.04]	[-3.43]				
Realized H			-0.0015**	-0.0003	-0.0003	-0.0002
			[-2.10]	[-0.68]	[-0.51]	[-0.56]
XZZ Skew		-0.0034***		-0.0034***		-0.0036***
		[-5.30]		[-5.20]		[-6.56]
ATM volatilities		-0.0063***		-0.0062**		-0.0047**
		[-2.62]		[-2.56]		[-2.25]
Log Option Volume		-0.0033*		-0.0034*		-0.0015
		[-1.85]		[-1.91]		[-1.44]
Log Option Open Interest		0.0025		0.0024		-0.0012
		[1.58]		[1.53]		[-1.22]
Log Stock $ Volume		0.0044		0.0046		0.0006
		[1.38]		[1.45]		[0.25]
Log Stock Volume		0.0019		0.0018		0.0061***
		[1.06]		[1.03]		[3.58]
Turnover		-0.0019		-0.0020		-0.0027**
		[-1.47]		[-1.54]		[-2.51]
Book-to-Market		-0.0003		-0.0002		-0.0010
		[-0.29]		[-0.21]		[-0.36]
Log Size		-0.0095***		-0.0094***		-0.0079***
		[-2.89]		[-2.84]		[-3.01]
Past 6M Return		-0.0006		-0.0007		-0.0007
		[-0.49]		[-0.56]		[-0.59]
Past 12M Return		0.0010		0.0011		0.0010
		[0.93]		[0.98]		[1.01]
Past Return Volatility		-0.0024*		-0.0024		-0.0043***
		[-1.65]		[-1.63]		[-2.76]
Past Return Skew		-0.0005		-0.0004		-0.0002
		[-0.95]		[-0.88]		[-0.60]
Adj. R^2	0.29%	13.15%	0.46%	13.18%	0.14%	9.21%

Table 3.7: Fama-MacBeth regression results. Panel A, B, C each have two regression results, one with only one regressor (either implied or realized H) and the other including a complete set of controls. Panel A shows results for implied H. Panel B presents results for realized H on the implied universe. Panel C uses realized H and the unrestricted universe. Numbers in brackets are t-statistics based on Newey-West standard errors.

3.5 Event Risk: Earnings Announcements and FOMC Meetings

In this section, we argue that cross-sectional differences in implied roughness of individual stocks reflect differences in near-term downside risk; we interpret the profitability of the rough-minus-smooth strategy as compensation for bearing this risk. We support this interpretation by considering the performance of the strategy around two types of events: company-specific earnings announcements, and interest rate announcements by the Federal Reserve's Open Markets Committee (FOMC). We present three pieces of evidence to support our argument. The strategy's profitability is greatest when restricted to stocks with earnings announcements in the subsequent month, when the potential for near-term idiosyncratic risk is high; roughness does not forecast earnings, suggesting that the strategy's profitability reflects compensation for risk rather superior selection of profitable companies; the strategy is not profitable in the lead-up to FOMC announcements — a period of elevated aggregate near-term risk rather than idiosyncratic near-term risk.

3.5.1 Earnings Announcements

3.5.1.1 Testing for Earnings Surprise Predictability

We begin by testing whether roughness predicts earnings surprises, as a possible explanation for the profitability of our strategy. Positive earnings surprises tend to be followed by stock price appreciation, so a signal that forecasts earnings surprises can serve as the basis for a profitable trading strategy. We will see, however, that this does not explain the profitability of the roughness signal.

We focus on the subset of data defined by

$$I^{ea} = \{(i,t)\text{: stock } i \text{ has an earnings announcement in month } t\},$$

using earnings announcement data from IBES. Letting I denote the full universe of stock-month pairs for which we have an implied roughness measure, $I \setminus I^{ea}$ denotes the subset that do not have an earnings announcement.

To measure earnings surprises, we use the standardized unexpected earnings (SUE) score from IBES. SUE measures the difference between a company's actual earnings and the mean forecast by analysts, normalized by the standard deviation of analyst forecasts in the previous quarter. To test for a relation between SUE and roughness, we use the Fama-MacBeth regression approach, meaning that we first run the following regression for every month t,

$$SUE_{i,t} = b_{0t} + b_{1t} H_{i,t-1} + e_{i,t}, \quad (i,t) \in I^{ea},$$

where $H_{i,t-1}$ denotes the implied roughness calculated for stock i in month $t-1$. We then average the b_{1t} over all months t and calculate standard errors adjusted for autocorrelation.

For comparison, we run the same analysis replacing implied roughness with the ATM skew in (3.11). Using data through 2005, [91] show that a greater ATM skew forecasts negative earnings surprises. In other words, before companies report disappointing earnings, low-strike puts become more expensive. [91] interpret this as evidence that investors with inside information trade on that information through options and that the stock market is slow to incorporate the information in option prices.

The left panel of Table 3.8 reports estimated coefficients and t-statistics for the two regressions. The bottom row confirms the finding of [91], with the benefit of more than ten years of additional data. The coefficient on the ATM skew is large, negative, and statistically significant. In contrast, the coefficient on implied roughness is indistinguishable from zero. Implied roughness does not forecast earnings surprises, and the implied roughness signal is distinct from the information in the ATM skew.

FM Regression		Portfolio Sorting
Variable	Coef	Difference in SUE
Implied H	0.035	-0.087
	[0.155]	[-0.324]
ATM skew	-2.744***	-0.300***
	[-2.827]	[-3.277]

Table 3.8: Predicting earning surprises using roughness by Fama-MacBeth regressions and portfolio sorting. Left panel shows coefficients and t-statistics in Fama-MacBeth regressions of standardized unexpected earnings (SUE) on implied roughness and ATM skew. Right panel shows the difference in average SUE in the top and bottom quintiles of stocks sorted by implied roughness or ATM skew. Numbers in brackets are t-statistics based on Newey-West standard errors.

The right panel of Table 3.8 further supports these conclusions. In this analysis, in each month t we limit ourselves to stocks with earnings announcements in month $t+1$. We sort these stocks into quintile portfolios based on roughness in month t. The table shows the difference in average SUE (in month $t+1$) between the highest and lowest roughness quintiles. The table shows the same comparison for stocks sorted on ATM skew in month t. We again see that a higher ATM skew forecasts negative earnings surprises whereas there is no relation between roughness and SUE. The profitability of the rough-minus-smooth strategy is not grounded in forecasting earnings.

3.5.1.2 Strategy Performance Near Earnings Announcements

Next we compare the performance of the rough-minus-smooth strategy when restricted to subsets of stocks based on the timing of earnings announcements. Specifically, we evaluate performance in three cases:

- I^{ea}: sort stocks with announcements in month t based on roughness in month $t-1$;
- $I \setminus I^{ea}$: sort stocks without announcements in month t based on roughness in

month $t-1$;

- $I^{ea,100}$: same as I^{ea} but only if at least 100 stocks in I have announcements in month t.

In all cases, portfolios are formed in month $t-1$ and returns are evaluated in month t.

Performance results under these restrictions are shown in the top panel of Table 3.9. Compared with the right-most column of Table 3.3, restricting attention to earnings-announcement stocks I^{ea} improves monthly alphas by roughly 40%, from around 0.50 to around 0.70. The estimated alphas are now only marginally significant, but this may be because the sample size (the number of stocks available each month) is now smaller. The results for $I^{ea,100}$ support this hypothesis: in months with at least 100 stocks available, the estimated monthly alpha goes above 1.0 (an annual alpha of more than 12%) and is highly significant. (These results are not sensitive to the choice of 100 as threshold.) In contrast, when we exclude stocks with earnings announcements, the $I \setminus I^{ea}$ alphas are smaller than the alphas in Table 3.3 and not statistically significant.

Taken together, the results in the top panel of the table show that sorting on roughness is most effective when applied to stocks facing a near-term idiosyncratic risk in the form of an earnings surprise. We interpret this to mean that greater roughness signals greater near-term downside risk, and that this risk is compensated with a price discount and a subsequent higher average return.

The analysis in the top panel is necessarily restricted to the universe I of stock-month pairs for which implied roughness is available. As a benchmark, the second panel shows market returns and alphas for the restricted sets of stocks used in the top panel. The second panel treats each restricted set as a long-only portfolio. The bottom row shows that stocks without earnings announcements earn lower returns;

	Mean Return	CAPM Alpha	FF3Mom Alpha	FF5Mom Alpha
Rough Minus Smooth (implied roughness universe)				
Earnings Announcement Stocks	0.71*	0.71*	0.70*	0.74*
(I^{ea})	[1.71]	[1.68]	[1.73]	[1.72]
EA Stocks – Threshold 100	1.00***	1.03***	1.07***	1.11***
($I^{ea,100}$)	[2.60]	[2.67]	[2.91]	[2.95]
No Earnings Announcement	0.30	0.28	0.29	0.29
($I \setminus I^{ea}$)	[1.41]	[1.23]	[1.25]	[1.22]
Long Only (implied roughness universe)				
Earnings Announcement Stocks	0.63*	0.12	0.14	0.20
(I^{ea})	[1.70]	[0.88]	[1.08]	[1.49]
EA Stocks – Threshold 100	0.47	-0.03	-0.01	0.09
($I^{ea,100}$)	[1.13]	[-0.28]	[-0.06]	[0.76]
No Earnings Announcement	0.29	-0.22***	-0.17***	-0.15**
($I \setminus I^{ea}$)	[0.81]	[-3.23]	[-2.69]	[-2.53]
Long Only (full universe)				
Earnings Announcement Stocks	0.77**	0.24**	0.21**	0.20**
(F^{ea})	[2.23]	[2.33]	[2.20]	[2.12]
No Earnings Announcement	0.38	-0.16***	-0.17***	-0.18***
($F \setminus F^{ea}$)	[1.12]	[-2.91]	[-2.79]	[-3.07]

Table 3.9: Strategy performance around earnings announcements. Top panel: Implied roughness strategy performance on stocks with earnings announcements in the next month (I^{ea}), in months with at least 100 candidate stocks ($I^{ea,100}$), and on stocks without earnings announcements $I \setminus I^{ea}$. Middle panel: Long-only performance on the same sets of stocks. Bottom panel: Long-only comparison of stocks with and without earnings announcements in the full universe of stock-month pairs. Numbers in brackets are t-statistics based on Newey-West standard errors.

but the main implication of the second panel is that the results in the top panel cannot be attributed to the restrictions in the definitions of I^{ea}, $I^{ea,100}$, and $I \setminus I^{ea}$. Moreover, the average implied H values in these three sets are nearly identical and all in 0.17–0.18.

This point is reinforced by the bottom panel. Here we drop the restriction to I and compare performance on the full universe of stocks with earnings announcements F^{ea} and without $F \setminus F^{ea}$. Stocks with earnings announcements earn higher returns than stocks without. Put differently, investors are compensated for bearing earnings announcement risk. Sorting on roughness identifies the stocks where this risk compensation is greatest.

These observations invite speculation on the implied strategy's losses in September 2001, which we mentioned in our discussion of Figure 3.3. Based on quintiles formed in August, the strategy would be long stocks facing near-term downside uncertainty. These stocks may have proved to be the most vulnerable to the disruptions and shock of the 9/11 attacks, leading the strategy to incur large losses.

3.5.2 Strategy Performance Near FOMC Announcements

We now turn from considering individual corporate events to FOMC announcements, which are among the most important scheduled events for the aggregate market. Indeed, [65] find that the excess return of the stock market is mainly earned during the 24-hour window before the earnings announcement; in other periods the average excess return is not statistically different from zero. If, as we have suggested, implied roughness ranks stocks on near-term idiosyncratic risk, then our strategy should not be expected to enhance returns in the lead-up to FOMC announcements.

Following [65], we consider announcements for the eight scheduled FOMC meetings each year. (Public announcements began in 1994, and our sample starts in 2000.) We define the pre-announcement period as the interval from the close of trading on

day $d-2$ to the close on day d, where d denotes the FOMC announcement date. We compare the performance of our strategy when it is restricted to invest in (or outside of) the pre-announcement period.

Our strategy is based on monthly data, so these timing restrictions require some explanation. When we limit ourselves to investing in pre-announcement periods, we evaluate performance only in the eight months of the year with scheduled announcements. In each such month, we take the return for the month to be the return over the two days that make up the pre-announcement period. We can apply this restriction to stock-month pairs in the implied roughness universe, in which case we label it $I^{preFOMC}$, and we can apply the restriction to the full universe of stock-month pairs and label it $F^{preFOMC}$.

We label the opposite restrictions $I^{nonFOMC}$ and $F^{nonFOMC}$. For the four months of each year without an FOMC announcement, the "nonFOMC" return is the just the ordinary monthly return. For the other eight months, the "nonFOMC" return is the return for the month excluding the two-day pre-announcement window.

The results are shown in Table 3.10, which has the same format as Table 3.9. The top panel compares the rough-minus-smooth strategy with the "preFOMC" and "nonFOMC" restrictions; the second panel shows long-only results with the same restrictions and limited to the universe of stock-month pairs for which we have implied roughness; the bottom panel shows long-only results when the restrictions are applied to the full universe of stock-month pairs.

The bottom panel is closest to the work of [65] and consistent with their conclusions: stocks earn higher returns during the pre-announcement period than at other times. The pattern is nearly identical in the middle panel, indicating that the I universe is representative of the full universe in its response to FOMC announcements.

In the top panel, the results flip. Sorting on implied roughness is not profitable during the pre-announcement period, when all stocks are facing a high degree of near-

	Mean Return	CAPM Alpha	FF3Mom Alpha	FF5Mom Alpha
Rough Minus Smooth (implied roughness universe)				
pre FOMC ann	0.09	0.11*	0.08	0.11*
($I^{preFOMC}$)	[1.53]	[1.72]	[1.40]	[1.78]
non pre-FOMC ann	0.43**	0.40**	0.42**	0.43**
($I^{nonFOMC}$)	[2.42]	[2.16]	[2.36]	[2.38]
Long Only (implied roughness universe)				
pre FOMC ann	0.41***	0.27*	0.33**	0.32**
($I^{preFOMC}$)	[3.20]	[1.92]	[2.24]	[2.12]
non pre-FOMC ann	0.12	-0.36***	-0.35***	-0.30***
($I^{nonFOMC}$)	[0.34]	[-3.58]	[-3.25]	[-2.70]
Long Only (full universe)				
pre FOMC ann	0.41***	0.26	0.37*	0.36*
($F^{preFOMC}$)	[2.59]	[1.38]	[1.81]	[1.76]
non pre-FOMC ann	0.26	-0.25**	-0.29***	-0.29***
($F^{preFOMC}$)	[0.78]	[-2.49]	[-2.80]	[-2.66]

Table 3.10: Strategy performance around FOMC announcements. Top panel: Implied roughness strategy performance in the pre-announcement period ($I^{preFOMC}$) and outside the pre-announcement period ($I^{nonFOMC}$). Middle panel: Long-only performance of the implied universe I during the same time periods. Bottom panel: Long-only performance of the full universe during the same time periods.

term systematic risk. The rough-minus-smooth strategy earns its returns the rest of the year, away from the pre-announcement period.

Recall that implied roughness measures the rate of decay of the ATM skew. A larger ATM skew indicates greater concern for downside risk, so a projected rapid decay in the ATM skew suggests concerns for downside risk that will be resolved quickly. Taking the results of this section together with those of Section 3.5.1.2, we see that proximity to a company-specific event enhances the performance of our strategy whereas proximity to an aggregate event has the opposite effect. This pattern suggests that the near-term downside risk captured by implied roughness is idiosyncratic. Moreover, the profitability of the rough-minus-smooth strategy suggests that investors are compensated for bearing this particular type of risk.

Our investigation does not explain why this near-term idiosyncratic risk should earn a risk premium. But the puzzle is not specific to our setting. Leaving aside roughness, the bottom panel of Table 3.9 records a well-known phenomenon of stocks

earning higher returns around earnings announcements. Sorting on implied roughness pushes this effect further.

3.6 Conclusions

We have investigated strategies for trading stocks based on measures of roughness in their volatility. We have compared long-short strategies based on realized roughness (calculated from high-frequency stock returns) and implied roughness (calculated from option prices). Both measures support a strategy of buying stocks with rougher volatilities and selling stocks with smoother volatilities; but sorting on implied roughness yields higher returns and is more robust to controlling for other factors. In particular, it is robust to controlling for illiquidity and the level of the ATM skew.

We have argued that implied roughness provides a measure of near-term idiosyncratic risk: a stock with greater implied roughness is one that the market perceives to have downside uncertainty that will be resolved quickly. On this interpretation, the profitability of our rough-minus-smooth strategy reflects compensation for bearing this risk. The performance of our strategy is enhanced near earnings announcements, when stocks face elevated idiosyncratic risk, and it is suppressed near FOMC announcements, when the dominant near-term risk is systematic.

Our work raises interesting questions for the rough volatility framework. Part of the appeal of this framework is that it simultaneously explains key features of realized volatility and the implied volatility surface extracted from option prices. Yet we find important differences in working with realized and implied measures of roughness. Estimating either measure of roughness from limited data presents significant difficulties, so it is unclear if the differences we observe present a challenge to the theoretical framework or simply call for better estimation methods.

Chapter 4

Heterogeneous Treatment Effects in Asset Pricing

4.1 Introduction

Flexible estimation of heterogeneous treatment effects (HTE) are key to many application areas, such as personalized medicine and optimal resource allocation, and have received lot of attentions in research areas of causal inference and machine learning. Various work has proposed causal variants of powerful machine learning methods that are tailored to HTE estimation. Some recent work includes methods based on random forests [87], boosting [76], and neural networks [80]. Also, [73] proposes an algorithm called R-learner that is more flexible than formal proposals in the sense that empirical researchers can choose any machine learning algorithms or black-box predictors in the estimation of HTE, and the resulting estimator still enjoys certain good properties under additional assumptions. The intended application domain for those estimation methods is usually in health care. We apply the technique to financial asset pricing.

Since the seminal work by [40], numerous papers have found new factors that can explain cross-sectional stock returns ([32] and [57]). Throughout this chapter, by factor models or factors, we mean the firm-level characteristics associated with the factor as potential predictors for future stock returns in the cross section By factor investment strategy, we mean the usual long-short strategies formed by single variable sorting based on the firm-level characteristic considered, as in [40], [41] and this large literature on empirical asset pricing. For example, by HML (high-minus-low), we mean the trading strategy where we periodically sort all stocks in a certain universe

based on book-to-market ratio, and buy the stocks with the highest book-to-market ratio and short the ones with the lowest.

In addition to finding new factors that contribute to explaining cross-sectional returns, understanding a factor's limitation and scope is also important: for what kind of stocks this factor works particularly well and for what kind of stocks it does not work at all in terms of investment returns of long-short strategies formed by the characteristic. Take the well-known value factor as an example: practitioners often refer to cases in which buying cheap stocks according certain valuation metric does not generate excess returns in future as a "value trap"[1]. Identifying value traps and characterizing those stocks a priori would be of great interests so that people could avoid them when performing value investing, considering the popularity of the value factor in the industry. Understanding what kind of stocks responds the strongest to higher book-to-market ratio is also important since it might shed some light on what drives the profitability of the value factor or book-to-market ratio. Similarly people use "momentum trap"[2] to refer to stocks or situations where past winners suffer in the coming periods.

Inspired by the terms "value traps" and "momentum traps", we use the term "characteristic traps" to refer a subset of stocks for which a naive factor investment strategy based on some firm characteristic does not tend to yield good returns in the future. On the positive side, we use the phrase "characteristic responders" to denote a subset of stocks with the following property: the factor investing strategy based on the focal characteristic performs the best when restricted to the particular subset, compared with cases where we apply the same trading strategy to the entire universe. It is very intuitive that future returns of different stocks at different times

[1]For example, see https://www.bloomberg.com/opinion/articles/2017-11-30/the-12-signs-a-cheap-stock-is-a-value-trap.

[2]https://www.bloomberg.com/news/articles/2018-03-26/morgan-stanley-warns-of-momentum-trap-for-rebounding-tech-stocks

might respond to certain firm characteristic differently. In this study we interpret the focal characteristic as the "treatment", and the goal of this chapter is to utilize HTE estimation techniques to identify and characterize characteristic responders and traps in a systematic and data-driven fashion.

The connections between HTE and factor models are as follows. We consider one focal factor/characteristic at a time. Use again value factor as an example. We view the book-to-market ratio as the continuous treatment variable Z and treat the next month's stock return as outcome variable Y. The treatment effect τ we want to study is thus the effect of book-to-market ratio on future stock returns. The treatment effect is heterogeneous in the sense that for different stocks with different characteristics, book-to-market ratios might affect future returns differently. This heterogeneity is denoted by $\tau(x_{jt})$ as a function of controls or features denoted by x_{jt} for stock j and month t. We collected around 40 firm characteristics other than the treatment characteristic and use them as controls x_{jt}. Based on findings from [40], we expect that value factor has positive average returns and that stocks with higher book-to-market ratio have higher future returns. Then "characteristic responders" for value, or "value traps", should be the stocks with the largest $\tau(x_{jt})$ (most positive). On the other hand, "value traps" would be the stocks with smallest $\tau(x_{jt})$ and their $\tau(x_{jt})$ could be close to 0 or even negative.

Our paper focuses only on the most standard factors–size, value, and momentum–as treatment characteristics for two reasons. First, these standard factors are probably the most well-known and are also widely used in practice, leading to a greater interest in understanding how the factor investment strategies work and where they work best/worst. Second, the fact that our approach works well for the most basic factors instead of having to cherry-picking factors from the literature is reassuring. Of course our procedure can be applied to any other factors/characteristics of interests.

To illustrate the difference between our HTE approach and the traditional ho-

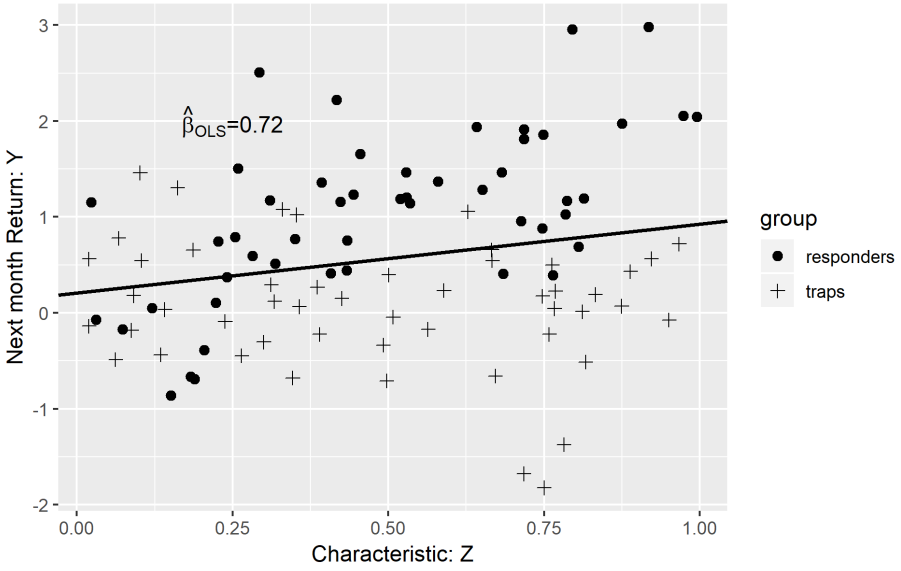

Figure 4.1: Homogeneous treatment effect estimation when the true effect is heterogeneous

mogeneous approaches in asset pricing, we conduct a simple simulation in Figures 4.1 and 4.2. In the simulation, we randomly generate 100 points from uniform distribution as x-axis variables Z. We set $Y = 0 * Z + \mathcal{N}(0, 0.8)$ for 50 points and $Y = 2 * Z + \mathcal{N}(0, 0.8)$ for the rest. Suppose that Z is some firm level characteristic whose effect on return is of interests and Y is future stock returns. Figure 4.1 represents the traditional approach: we treat all stocks the same way and run a linear regression to estimate the slope. Figure 4.2 corresponds to our new proposal, where we still keep the linear assumption but the effect of Z on Y can be different for the two groups of stocks. In the simulation, we know the data generating process, but in reality, we need to differentiate the two groups in a data-driven fashion. We also simplify $\tau(x_{jt})$ in the simulation to have only two levels of different treatment effect, 0 and 2, corresponding to our characteristic trap and characteristic responder set respectively. The line in Figure 4.1 is an OLS regression line while the two lines in Figure 4.2 is OLS applied to each group. We also mark the value of the slopes on the plots.

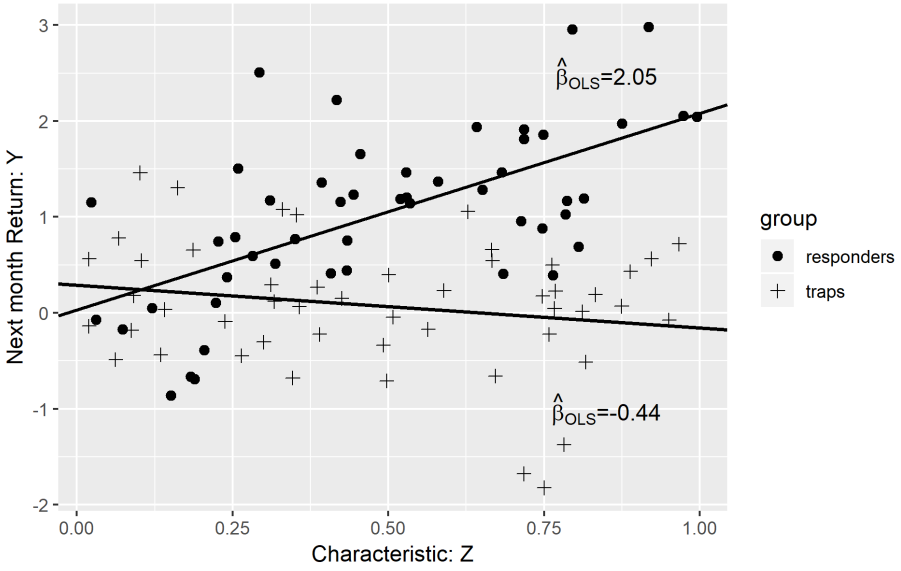

Figure 4.2: Heterogeneous treatment effect estimation when the two groups with different treatment effects are known

For HTE estimation, we focus on the R-learner proposed by [73] for its flexibility and nice theoretical properties. The resulting estimates $\hat{\tau}(x_{jt})$ can be used to identify characteristic traps and opportunities. For our test period 2006-2018, at the end of month $t-1$, we set the quintile with the highest $\hat{\tau}(x_{jt})$ to be the characteristic-responder subset and the 20% of stocks with the lowest $\hat{\tau}(x_{jt})$ to be the characteristic-trap subset. We form two long-short portfolios restricted to both subsets of stocks and hold the two long-short portfolios for month t. We repeat the same steps at the end of month t. For the value factor, the long-short strategy restricted to the characteristic-responder subset has an average monthly return of 0.71%, whereas the same HML strategy applied to the characteristic-trap set has an average monthly return -0.06%. As a benchmark, standard HML strategy applied to the entire universe of stocks has a monthly average return of 0.25% for the same time period. All long-short portfolios are equal weighted. We've found similar results for size and momentum as treatment variables. The difference in HML average returns of the characteristic responders and traps is statistically significant.

These results have the following implications. First, the treatment effects for size, value, and momentum are indeed heterogeneous and our procedure successfully identified the characteristic responders and characteristic traps. Similar to the applications in personalized medicine, if we successfully identify a sub-population with good treatment effects, we could only prescribe drugs to that specific population. Here, we could apply factor the investing strategy only to the characteristic-responder subset where the strategy is particularly effective. Secondly, using fitted treatment effects $\hat{\tau}(X_{jt})$, we can characterize the two subsets: for example, for the value factor, $\hat{\tau}(X_{jt})$ tends to be very small when stock j's trailing return is very bad compared with other stocks. We conduct this type of analysis in section 4.5. Lastly, we could form a long-short of long-short trading strategy for each factor. Taking value as an example, since HML strategy applied to the quintile with the highest $\hat{\tau}(X_{jt})$ outperforms the HML strategy restricted to the quintile with the smallest $\hat{\tau}(X_{jt})$, we could long the HML portfolio restricted to the value responders and short the HML portfolio formed within the value traps. This long-short of long-short strategy has an average monthly return of 0.77%. The same strategy for size and momentum factor have average monthly returns of 1.54% and 1.18% respectively and both are very significant statistically. This long-short of long-short strategy takes advantage that the fact that our $\hat{\tau}(X_{jt})$ predicts actual $\tau(X_{jt})$ well enough such that the quintile with the highest $\hat{\tau}(X_{jt})$ has higher true $\tau(X_{jt})$'s than the quintile with the lowest $\hat{\tau}(X_{jt})$.

Our work is related to several streams of literature. First, our work falls into the active research area applying machine learning to empirical asset pricing. [55] apply different kinds of machine learning algorithms to predicting returns using features mainly including firm characteristics[3]. [28] is similar but focuses on deep learning. The objective of their papers and many other studies are purely to predict cross-sectional returns using characteristics in a non-linear way. To some extent the ul-

[3]and their interactions with other aggregate features

timate goals of those papers and ours are the same: predicting returns in the cross section. However, we view our detailed analysis of different factors/characteristics as a middle ground between linear factor models and simple applications of off-the-shelf black-box machine learning algorithms to return prediction: we study how the effects and predictability of treatment characteristics on future returns vary with other characteristics. [47] apply adaptive group LASSO to study the predictive relationship between characteristics and future returns in a nonparameteric way. Their work assumes an additive structure between different characteristics; therefore interaction effects between any features are excluded from the model and can only be added in an ad-hoc fashion by empirical researchers. Our work complements their approach in the sense that we focus purely on how all other characteristics interact with one treatment characteristic in a non-linear fashion.

Second, our work is related to causal inference and machine learning, and HTE estimation in particular. A recent study by [43] in the finance context proposes a procedure to conduct statistical inference on new factors' contribution to explaining cross-sectional returns. Their focus is to get the inference correct after the model-selection step that chooses a subset of features from a high dimensional set of features or control variables. Their proposal is related to post-selection inference literature such as [14] and [31]. All of those papers focus on unbiased estimation and correct inference on homogeneous effects when machine learning is involved. For theory of HTE estimation, recent studies include [73], [76] and [28]. Our paper applies R-learner in [73] to an interesting empirical problem in asset pricing.

Lastly, our work is related to existing studies in finance on factors, characteristics and expected return in the cross section. [35] find that the data actually support characteristics, rather than loading on undiversifiable risk factors, to explain cross-sectional stock returns. As mentioned previously, in this chapter we do not focus on factor or covariance structure but rather on characteristics, which is the same as most

machine learning papers in finance such as [55] and [47]. From the efficient market point of view, we expect that characteristics have predictive power for future stock returns because they are proxies for loadings on certain fundamental risk factors that demands risk premium. Along this line of thinking, our HTE estimation for one treatment characteristic could potentially capture how well the treatment characteristic approximate the loadings to the fundamental risk factor: for "characteristic responders", the treatment characteristic is very good proxy for the actual loading on the underlying risk factor and as a result the treatment effect is strong; for "characteristic traps", the treatment characteristic might be very poor in approximating the loading and thus the treatment effect is weak or even has the opposite sign. On the other hand, there might be alternative explanations for the predictability of firm characteristics. For example, many existing studies in finance focus on source and driver of the profitability of momentum strategies. The causal-inference and machine-learning hybrid approach we take allows us to interpret our estimation results and may shed some lights on why certain characteristics or factors are profitable. In our empirical study, we use more recent data to confirm some of the findings in [59] regarding how momentum strategy's profitability varies depending on firm size . Our approach has two distinct advantages over the traditional approach in papers like [59]. First, we could control for many more confounders at the same time in a nonlinear fashion compared with portfolio sorting, for which controlling for one additional characteristic by double sorting is easy but becomes very hard for more characteristics; second, our procedure is fully automated. If we have no prior knowledge on what characteristic would affect profitability of momentum strategy, our procedure systematically finds out what features matter most and in what way. Most importantly, we are able to provide new insights on the problem by showing that other features such as return volatilities are also important for the profitability of momentum strategies.

The rest of the chapter is organized as follows: section 4.2 explains the background

of HTE estimation with a brief introduction to causal inference and machine learning and in particular the estimator we use in our empirical study, R-learner by [73]. Section 4.3 details our algorithmic procedure of applying R-learner combined with gradient boosting to asset pricing, and section 4.4 presents our main results. In section 4.5, we go deeper in understanding the HTE estimates and what our fitted models have captured. Section 4.7 concludes the paper.

4.2 Heterogeneous Treatment Effects Estimation

The main model we consider is given as below:

$$Y_i = \mu^* (X_i) + Z_i \tau^* (X_i) + \epsilon_i, \qquad \mathbb{E}[\epsilon_i | X_i, Z_i] = 0 \tag{4.1}$$

, where Y denotes the outcome we care about and Z denotes a continuous treatment variable whose effect on Y is denoted by $\tau^*(X)$. X's are the features other than the treatment. $\mu^*(X)$ is the base-case outcome as a function of X when $Z = 0$. The condition $\mathbb{E}[\epsilon_i | X_i, Z_i] = 0$ is often called exogeneous assumption or no endogeneity assumptions in linear regression models. We will arrive at equation 4.1 later in this section. First, we recall the framework of causal inference and treatment effect estimation in the binary treatment case, which is dominant in the causal inference literature. We then discuss generalizations to continuous treatment cases and assumptions made therein. Note that binary and continuous treatment variables are conceptually very similar, but to be clearer, we use different notations: Z denotes continuous treatment variable while we use W if the treatment is binary. Lastly, we discuss how we could estimate HTE, and cast the traditional regression approach in asset pricing as a simple special case to build the connections. Basically, we treat future stock returns as Y, one firm characteristic as treatment W or Z, and the rest of firm characteristics as feature X's.

4.2.1 Causal inference framework: Binary treatment W

We first briefly recall the causal inference framework before diving into HTE estimation. The framework we take is often called potential outcome framework. We assume we have n i.i.d. samples of (X_i, Y_i, W_i) drawn from a joint distribution $f(X, Y, W)$, where

- $Y_i \in \mathbb{R}$: observed outcome for individual i.
- $X_i \in \mathbb{R}^p$: features for individual i of dimension p
- $W_i \in \{0, 1\}$: whether i receives treatment or not
- "Potential Outcome": $\{Y_i(0), Y_i(1)\}$ depending on whether i gets treatment: $Y_i = Y_i(W_i)$

The focus is on estimating HTE defined as:

$$\tau^* (x) = \mathbb{E} [Y (1) - Y (0) | X = x] \tag{4.2}$$

, which are also called conditional average treatment effects (CATE). We use superscript * to emphasize that it is an unknown population quantity. Note that, as hinted in the introduction section, in this chapter's empirical study we mainly use continuous treatment variables like book-to-market ratio but we could also think of interesting binary treatment variables. For example, W could represent some events of firms, such as dividend cancellation announcements, where the effects of the event could potentially be heterogeneous.

The *fundamental problem of causal inference* is that we only observe one of $\{Y_i(0), Y_i(1)\}$ for each i. The other one in $\{Y_i(0), Y_i(1)\}$ is always missing. In the medical trial setting this problem can be stated as how long will the patient in the control group live had him taken the drug. In our case, the problem can be stated as what will the next month return be had the stock announced a dividend cancellation

in the current month. This is what makes causal inference problem special: if we can observe both $Y(0)$ and $Y(1)$, we get to observe treatment effect $\tau_i = Y_i(1) - Y_i(0)$ for each observation i, too, and HTE estimation reduces to a supervised learning task using any ML methods to fit to the data set $(X_i, Y_i(1) - Y_i(0))$.

To tackle the fundamental problem of causal inference, we need some structure and assumption in order to recover τ^*. There are usually two types of studies that try to infer treatment effects: randomized experiments and observational studies. Randomized experiments are settings where researchers are able to conduct experiments in a controlled environment by randomly assigning treatment to individuals, which are common in drug trials and A/B testing conducted by internet companies such as Facebook and Google. Experiments remain the golden standard of causal inference, but, unfortunately, in most economic applications, experiments are not feasible. In this case, we need to work with observed data only and that is called observational studies. One standard assumption to assume in observational studies is no unmeasured confounders: conditioning on the observed features X_i, treatment assignment W_i is independent of potential outcome $(Y_i(0), Y_i(1))$, i.e.

- No unmeasured confounders

$$(Y_i(0), Y_i(1)) \text{ independent of } W_i \text{ conditional on } X_i \tag{4.3}$$

Note that in a controlled randomized experiment, assumption (4.3) is automatically satisfied. However, in observational studies, we usually need to assume equation (4.3) is true. If we have reasons to believe that assumption (4.3) is violated, we have an endogeneity problem and need to find instrumental variables (IV) for identification of treatment effects. See corresponding chapters in textbook such as [89] for more details along that direction. We explain the meaning of this assumption in our empirical context in the next subsection, where we describe the continuous treatment case.

Under assumption 4.3, we can decompose observed Y_i into the following three terms:

$$Y_i = \mathbb{E}\left[Y\left(0\right)|X = X_i\right] + W_i\tau\left(X_i\right) + \epsilon_i \tag{4.4}$$

,where the last term ϵ_i satisfies

$$\begin{aligned}\mathbb{E}\left[\epsilon_i|W_i, X_i\right] &= \mathbb{E}\left[Y_i - \mathbb{E}\left[Y_i\left(0\right)|X = X_i\right] - W_i\tau\left(X_i\right)|W_i, X_i\right] \\ &= 0\end{aligned}$$

Our goal is then to flexibly estimate HTE, that is, the CATE function $\tau\left(X_i\right)$.

4.2.2 Causal inference framework: Continuous treatment Z

We generalize the model to continuous treatment variable case where the treatment is denoted by Z. Continuous treatment is not as widely studied in the causal inference literature compared with binary treatment. Again, outcome variable is denoted by Y ,and the observed outcome becomes $Y_i = Y_i(Z_i), Z_i \in \mathcal{Z}$, where $\mathcal{Z}$ can be a continuum, and for each individual i, we only get to observe its value for one Z_i, $Y_i = Y_i(Z_i)$. The problem here is more complicated than the binary case, and thus we need to assume more than just equation (4.3). In particular, we assume that conditional on X, the effect of increasing one unit of Z is constant regardless of current levels of Z, in addition to no unmeasured confounders. The assumption in this case can be described as below:

- Linear treatment and no unmeasured confounders

$$\mathbb{E}\left[Y_i|X_i = x, Z_i = z\right] = \mu^*\left(x\right) + z\tau^*\left(x\right) \tag{4.5}$$

The assumption above implies equation (4.1) in the starting paragraph of this section, which we repeat below:

$$Y_i = \mu^*\left(X_i\right) + Z_i\tau^*\left(X_i\right) + \epsilon_i, \qquad \mathbb{E}\left[\epsilon_i|X_i, Z_i\right] = 0$$

The specific model of equation (4.5) has also been studied recently by, for example [9] and [1]. We will assume equation (4.5) throughout the paper, and therefore equation (4.1) always holds in our setting. The zero expectation of residual ϵ conditional on X, Z in equation (4.1) is often called exogenous assumption. It may seem at first to be a strong assumption; however, we explain what this assumption means in our empirical setting in the following section, and this assumption turns out to be no more restrictive than what is assumed in usual factor model estimation. In fact, equation (4.1) is much more flexible than standard assumptions of regression approach used in empirical asset pricing.

4.2.3 Linear factor models as a special case

Before going to HTE estimation, we cast the traditional regression approach to factor models as a special case of our main specification in equation (4.5). We simplify the HTE estimation problem into a linear, homogeneoues setting and build the connection between this setting and the estimation of traditional linear factor models.

Because we are generally interested in continuous variables in asset pricing models, we cast regression models under our causal framework using equation (4.1), which is implied by assumption (4.5). We assume the following two additional assumptions which simplifies the problem significantly.

- The treatment effect is homogeneous, that is.

$$\tau^*(x) = \tau \tag{4.6}$$

- Linear in features X for base case response $\mu^*(x)$:

$$\mu^*(x) = x\beta \tag{4.7}$$

Then under assumptions (4.5), (4.6) and (4.7), equation (4.1) becomes

$$Y_i = X_i\beta + Z_i\tau + \epsilon_i, \qquad \mathbb{E}\left[\epsilon_i | Z_i, X_i\right] = 0 \tag{4.8}$$

OLS theory says that $\hat{\tau}_{OLS}$ is consistent for true τ. In linear factor models, we want to estimate whether or not a particular factor, say, book-to-market ratio for stock j in month t, can predict stock j's return for month $t+1$, after controlling for other firm characteristics represented in X_{jt}. What we have is essentially panel data regression, but we care only about the cross-sectional relationship between book-to-market ratio and next month stock returns. Therefore one way to do the analysis is to run a panel data regression with time fixed effects, which focuses on variations in the cross section and averages out along the time axis. In terms of the point estimates, adding time fixed effect is equivalent to demeaning both the target variable and all regressors from their monthly average. Denote stock return by r, stock features by X, the book-to-market ratio by BM. Following [40], we use the log of book-to-market ratio, $\log BM_{jt}$, in the regression for it has distribution closer to normal distribution. We have the panel regression equation with cross-sectionally demeaned variables:

$$r_{jt} - \bar{r}_t = \left(X_{jt} - \bar{X}_t\right)\beta + \left(\log BM_{jt} - \log \bar{BM}_t\right)\tau + \epsilon_{jt} \tag{4.9}$$

,where $\bar{r}_t = \frac{1}{J_t}\sum_{l=1}^{J_t} r_{lt}$ and $\log \bar{BM}_t$ is similarly defined.

Regression equation (4.9) is related to the possibly more popular called Fama-Macbeth regression proposed by [42]. See [75] for more about that connection.

Comparing equations (4.8) and (4.9), we can then map quantities in factor models to ones in the causal framework: $r_{jt} - \bar{r}_t$ is our outcome variable Y, $\left(X_{jt} - \bar{X}_t\right)$ are features/other controls X of dimension p, and, most importantly, $\left(\log BM_{jt} - \log \bar{BM}_t\right)$ is the treatment Z whose effects on next-month returns relative to the cross-sectional averages,$r_{jt} - \bar{r}_t$, are of primary interests. On the other hand, this comparison also shows what assumptions are behind usual OLS estimation for traditional linear regression in factor models (equation (4.9). For OLS estimator in equation (4.9) to be consistent, equation 4.8 must hold, which is often called exogenous condition[4]. We've

[4]Actually the weakest condition to ensure OLS is consistent is $\mathbb{E}\left[\begin{bmatrix} X_i\epsilon_i \\ Z_i\epsilon_i \end{bmatrix}\right] = 0$

seen that this is equivalent to assumptions (4.5), (4.6) and (4.7). We keep assumption 4.5 throughout the paper. However, in section 4.2.4 we introduce HTE estimation techniques developed by the machine learning and causal inference researchers, which help us relax assumptions 4.6 and 4.7 in our empirical studies.

We next discuss the assumption of no unmeasured confounders or no endogeneity assumptions of equation (4.3) and (4.5). Those assumptions imply exogeneous condition like $\mathbb{E}[\epsilon|X, Z] = 0$, which is the type of assumptions that justifies OLS in linear models. We assume equation (4.5) and thus (4.1) throughout this chapter. We argue that this assumption is reasonable for the following two reasons. First, it is no stronger than the assumptions made in the literature when estimating traditional linear factor models using regressions. If we run the type of panel regression in equation (4.9) or similar Fama-Macbeth regressions and start to interpret the sign and magnitude of τ or conduct statistical inference on τ, we are implicitly assuming equation (4.5) or (4.1) holds. Secondly, compared with prior regression approach that assumes exogeneity in the finance literature, our empirical study based on HTE estimation is in a better position to assume no unmeasured confounders and thus mitigate endogeneity problem for the following two reasons: (1) In the existing literature, only a few characteristics are collected as controls X for regression equations like (4.9), whereas we include around 40 features as controls X. Endogenity problem caused by omitted-variable bias is much less of an issue in our study. (2) We also utilize the features X in a much stronger way in that we estimate a nonlinear function of X, $\mu^*(X)$ as in equation (4.1) instead of using a linear form $X\beta$. It is possible that X^2 should have been included as a control, but in the linear setup researchers fail to add them. The mis-specification causes omitted-variable bias and the endogeneity problem. In our procedure described in the following sections, we could estimate a flexible functional form for the base model $\mu^*(X)$ to alleviate this concern compared with the traditional approach.

4.2.4 R-learner for HTE estimation

In this subsection, we explain how we are going to estimate the model in (4.1). There are many alternatives but we focus on one particular estimator, R-learner, proposed by [73]. The problem [73] try to tackle is how to turn a good generic black-box predictor into a good treatment effect estimator that has some nice theoretical properties. In terms of what machine learning tools we can use, R-learner is more flexible than most HTE estimation methods where "causal" variants of machine learning methods still require efforts from specialized researchers. Our empirical studies in section 4.4 focus on applications of R-learner.

In this section, we only assume assumption (4.5), and by taking expectation for both sides of the equation 4.5 conditioning on only X_i, we have

$$\begin{aligned} \mathbb{E}\left[Y_i|X_i, Z_i\right] &= \mu^*\left(X_i\right) + Z_i\tau^*\left(X_i\right) \\ \mathbb{E}\left[Y_i|X_i\right] &= \mu^*\left(X_i\right) + \mathbb{E}\left[Z_i|X_i\right]\tau^*\left(X_i\right) \end{aligned} \tag{4.10}$$

Plugging in (4.10) back into regression equation (4.1), we have the following:

$$Y_i - m^*(X_i) = (Z_i - e^*(X_i))\tau^*(X_i) + \epsilon_i, \qquad \mathbb{E}[\epsilon_i|X_i, Z_i] = 0, \tag{4.11}$$

where

$$\begin{aligned} m^*(x) :=& \mathbb{E}[Y_i|X_i = x] \\ e^*(x) :=& \mathbb{E}[Z_i|X_i = x] \end{aligned}$$

In the binary treatment setting, $e^*(x) = \mathbb{E}[Z_i|X_i = x]$ described above gives the conditional probability of getting treatment and is often called propensity score in the causal inference literature. The exposition here follows from [73] and the only minor difference is that we have a continuous treatment variable Z_i as opposed to binary. Equation (4.11) is our main estimation equation in R-learner. In equation (4.11), we essentially subtract from Y and Z their conditional expectation conditioning on X.

This step is often called residualization and is very intuitive: we want to take out the effects of other controls X_i and isolate the treatment effects τ^*. [78] first Utilized The residualization form of equation (4.11) for the binary case.

We next recall the R-learner HTE estimator and a few of its variations below. First, an oracle with knowledge of true population quantities $m^*(x)$ and $e^*(x)$ can estimate function τ^* by

- Oracle estimator $\tilde{\tau}(x)$:

$$\tilde{\tau} := \arg\min_{\tau(x)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - m^*\left(X_i\right) - \left(Z_i - e^*\left(X_i\right)\right) \tau\left(X_i\right) \right)^2 + \Lambda_n \left(\tau\left(x\right) \right) \right\}, \tag{4.12}$$

where Λ_n is the regularization term that should be tuned by cross-validation (CV). We could think of it as similar to the L^1 penalty term in LASSO regression, or the maximum tree depth in tree related methods. This term is crucial since without this regularization, we could minimize the training error to a point overfitting the data and failing to generalize.

In reality, we cannot implement the oracle estimator, but we can estimate first m^* and e^* and plug our estimates $\hat{m}$ and $\hat{e}$ into the minimization problem in (4.12), which is the R-learner proposed by [73]:

- R-learner $\hat{\tau}(x)$ is estimated as below, where $\hat{m}^{(-i)}(X_i)$ and $\hat{e}^{(-i)}(X_i)$ mean hold-out predictions made by models $\hat{m}$ and $\hat{e}$ fitted to data without the i^{th} data point:

$$\hat{\tau} := \arg\min_{\tau(x)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{m}^{(-i)}\left(X_i\right) - \left(Z_i - \hat{e}^{(-i)}\left(X_i\right)\right) \tau\left(X_i\right) \right)^2 + \Lambda_n \left(\tau\left(x\right) \right) \right\} \tag{4.13}$$

In summary, we could implement the R-learner in the following two steps:

- Step 1: Fit $m(x)$ and $e(x)$ via any black-box predictive methods tuned for optimal predictive accuracy using CV.

- Step 2: Minimize the causal loss function plus regularization term $\Lambda_n(\tau(x))$, again via any black-box methods. Use CV to tune hyperparameter to combat overfitting:

$$\hat{\tau} := \arg\min_{\tau(x)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{m}^{(-i)}(X_i) - \left(Z_i - \hat{e}^{(-i)}(X_i) \right) \tau(X_i) \right)^2 + \Lambda_n(\tau(x)) \right\}$$

Note that we use hold-out predictions for nuisance components $m^*(x)$ and $e^*(x)$ when estimating what we care about in step 2. This usage of hold-out predictions is also known as cross-fitting for its similarity to cross-validation. The difference is that here we want to use hold-out predictions in fitting our main parameters of interests as opposed to evaluating performance of predictive models. It is a widely used trick in making correct statistical inference when machine learning methods are involved ([10] and [30]). Usually, k-fold cross-fitting is used, and in our empirical study, we set $k = 5$.

[73] show R-learner has nice theoretical guarantees. For example, with additional assumptions on the true form of τ^* and the machine learning algorithms used, [73] prove that $\mathbb{E}\left[(\hat{\tau}_n(X_i) - \tau^*(X_i))^2\right]$ converging to 0 as fast as $\mathbb{E}\left[(\tilde{\tau}_n(X_i) - \tau^*(X_i))^2\right]$. Also, as mentioned earlier, although the nice properties from theory require certain conditions, we are free to use any black box predictors in R-learner when implementing it in practice, which makes the approach very flexible.

[73] named their procedure R-learner, based on Robinson's transformation, partly to recognize [78]'s work and partly to emphasize the importance of residualization.

In the next section, we describe how we apply R-learner to our setting of factor models.

4.3 Application of the R-learner to Factor Models

4.3.1 Algorithmic procedure

We first need to choose what feature to use as treatment Z and the remaining features collected will be control features X. We collect and calculate around 40 features, including items from companies' balance sheets, past returns, past volatilities, trading volumes, and so on. The full feature list is given in Table D.1 in the appendix. Note that we have panel data, and therefore the subscript index is $_{j,t}$ instead of just $_i$. We focus our study on characteristics associated with Fama-French 3 factors plus momentum as possible treatments. We pick one of size, value and momentum at a time as the treatment variable, and use the rest as controls. We use log book-to-market ratio $\log BM_{jt}$ as the treatment to describe the procedure, and the steps for the other two treatment variables are the same except for changing the treatment. We use D_{train} to denote the set of (j,t) pairs belonging to training set.

Since we are mainly interested in the cross-sectional relationship between treatment variable and future stock return, inspired by panel regression equation (4.9), we demean the response variable, the treatment variable and all control variables by their cross-sectional averages except for ratio based features[5]. Note that as pointed out before, this demeaning procedure corresponds to adding time-fixed effects in panel regression and it is very intuitive: the outcome we care about is not really the exact level of stock returns but each stock's return compared with the cross-sectional average. In addition, to make values across different time periods more comparable, we further standardize all non-ratio variables by their cross-sectional standard deviations, to ensure that all non-ratio variables have mean 0 and standard deviation 1 for every month. Sometimes this pre-processing step is called feature normalization or standardization in machine learning. We only perform this standardization on

[5]

non-ratio variables since ratio-based variables such as log of book-to-market ratio are already comparable across stocks and months. We list which variables are ratio-based in Table D.1. We use $\dot{}$ to denote our operations of pre-processing on target variable, treatment variable, and control variables. That is, for any variable x_{jt} of stock j at month t,

$$\dot{x}_{jt} := \begin{cases} x_{jt} & \text{for ratio-based feature } x \\ \frac{x_{jt} - \frac{1}{J_t}\sum_{k=1}^{J_t} x_{kt}}{\sqrt{\frac{1}{J_t - 1}\sum_{v=1}^{J_t}\left(x_{vt} - \frac{1}{J_t}\sum_{k=1}^{J_t} x_{kt}\right)^2}} & \text{otherwise} \end{cases} \tag{4.14}$$

Thus, for our application, using Robinson's transformation in the R-learner, under assumption (4.5), we have that

$$\dot{r}_{j,t+1} - m^*(\dot{X}_{j,t}) = \tau(\dot{X}_{j,t})\left(\dot{Z}_{j,t} - e^*(\dot{X}_{j,t})\right) + \epsilon_{j,t}, \tag{4.15}$$

where

$$m^*(x) := \mathbb{E}[\dot{r}_{j,t+1} | \dot{X}_{j,t} = x]$$
$$e^*(x) := \mathbb{E}[\dot{Z}_{j,t} | \dot{X}_{j,t} = x]$$

For example, when $\log BM_{j,t}$ is the treatment variable Z, since $\log BM_{j,t}$ is a ratio-based feature, we have

$$\dot{Z}_{j,t} = \log \dot{BM}_{j,t} = \log BM_{j,t}$$

Note that all variables involved in equation (4.15) have a $\dot{}$ above them.

[73] try several different machine learning methods such as boosting and LASSO, in step 1 and 2 of the R-learner procedure, to fit m, e, and τ models. We decide to focus on gradient boosting ([48]) and the particular implementation we choose is xgboost by [29]. Gradient boosting is sometimes considered as the best off-the-shelf regression method in machine learning for its impressive performance and relative ease to use. The particular implementation by [29] is very efficient, powerful, and has

been used by many winning teams in predictive data science competitions[6]. It also has the ability to deal with missing values easily, which comes handy in our situations because a lot of balance sheet items are missing for some stocks from time to time. When training m, e, and τ models in the R-learner procedure described in Table 4.2, we always use xgboost.

We next describe how cross-validation is used to do some basic hyper-parameter tuning for xgboost when fitting m, e, and τ models. We decide to use five-fold CV and the folds are split randomly but based on months. To be precise, we randomly split all months in the training data into five folds, denoted by $l = 1, 2, 3, 4, 5$. We use L to denote the mapping from months t to ID of folds resulting from the random fold splitting; that is, $L(t) = l$ if and only if month t is assigned to fold l. All stock month pairs (j, t) s.t. $L(t) = l$ belong to fold l, $l = 1, 2, 3, 4, 5$. We define then $L(j, t) := L(t)$. For the m model, we train five different versions based on which fold is left out in the training data. For example, we train one model using data from folds $l = 2, 3, 4, 5$ and leaving out fold $l = 1$. Denote this fitted model as $\hat{m}^{(-1)}$ where the superscript minus 1 means that data points in fold $l = 1$ are excluded during training. $\hat{m}^{(-l)}$, $\hat{e}^{(-l)}$ and $\hat{\tau}^{(-l)}$ are similarly defined for $l = 1, 2, 3, 4, 5$. We choose a few hyper-parameters (listed in Table 4.1) that are particular important to gradient boosting to tune via CV. We use randomized grid search: first we define a reasonable search range for all hyper-parameters to tune and we randomly select five sets of hyper-parameter values within the ranges specified in Table 4.1 in a uniform fashion. We record the five sets of hyper-parameter values and for each set of values, we train the five m models to get $\hat{m}^{(-l)}$ for $l = 1, 2, 3, 4, 5$ and use out-of-fold predictions to form the evaluation metric based on squared loss. The CV evaluation metric for the e model is similarly

[6]See www.kaggle.com for details on such data science competitions

defined. To be precise,

$$CV_m := \sum_{j,t \in D_{train}} \left(\dot{r}_{j,t+1} - \hat{m}^{(-L(j,t))} \left(\dot{X}_{j,t} \right) \right)^2 \tag{4.16}$$

$$CV_e := \sum_{j,t \in D_{train}} \left(\dot{Z}_{j,t} - \hat{e}^{(-L(j,t))} \left(\dot{X}_{j,t} \right) \right)^2 \tag{4.17}$$

We record the hyper-parameter value yielding the best (smallest) evaluation metrics CV_m and CV_e for the m and e models respectively. $\hat{m}^{(-l)}$, $l = 1, 2, 3, 4, 5$ denote models that are trained using the best hyper-parameter values in terms of CV_m, and so do $\hat{e}^{(-l)}$. Now, we are ready to list the first two steps of the R-learner procedure in our setting with full details.

- Step 0: Conduct random splitting into five folds based on months t. Use $L(j, t)$ to denote the resulting mapping from (j, t) to fold ID. We keep the folds the same throughout the following steps.

- Step 1: Use xgboost to fit outcome $\dot{r}_{j,t+1}$ using features $\dot{X}_{j,t}$ only. Randomly generate five sets of hyper-parameter values based on ranges specified in Table 4.1. For each set of hyperparameter values, fit five different versions, $\hat{m}^{(-l)}$ for folds $l = 1, 2, 3, 4, 5$ and record the CV loss CV_m. Pick the set of hyper-parameter values that yields the smallest CV_m. Record the best hyper-parameter values and save $\hat{m}^{(-l)}$, $l = 1, 2, 3, 4, 5$ that are trained with the best hyper-parameters for future use.

- Step 2: Repeat step 1 but for e models. use xgboost to fit outcome $\dot{Z}_{j,t}$ using features $\dot{X}_{j,t}$ only. Randomly generate five set of hyper-parameter values based on ranges specified in Table 4.1. For each set of hyperparameter values, fit five different versions, $\hat{m}^{(-l)}$ for folds $l = 1, 2, 3, 4, 5$ and record the CV loss CV_e. Pick the set of hyper-parameter values that yields the smallest CV_m. Record

the best hyper-parameter values and save $\hat{m}^{(-l)}$, $l = 1, 2, 3, 4, 5$ that are trained with the best hyper-parameters for future use.

Next, we describe the final step of HTE estimation: estimating the τ^* function. There is another detail of R-learner in this step, namely, the cross-fitting (CF). We record out-of-fold predictions from our fitted m and e models in preparation for estimating $\tau(x)$. That is, for each (j,t) pair in the training set, $\hat{m}^{(-L(j,t))}(\dot{X}_{j,t})$ and $\hat{e}^{(-L(j,t))}(\dot{X}_{j,t})$ are computed and recorded. Note that we use the same fold assignment as in the cross-validation stage of steps 1 and 2. In the final step, we use xgboost to minimize the following causal loss function over $\tau(x)$.

$$\hat{L}_n\left(\tau; D_{train}\right) = \sum_{j,t \in D_{train}} \left(\dot{r}_{j,t+1} - \hat{m}^{(-L(j,t))}\left(\dot{X}_{j,t}\right) - \left(\dot{Z}_i - \hat{e}^{(-L(j,t))}\left(\dot{X}_{j,t}\right)\right)\tau\left(\dot{X}_{j,t}\right)\right)^2 \tag{4.18}$$

The fact that out-of-fold predictions $\hat{m}^{(-L(j,t))}$ and $\hat{e}^{(-L(j,t))}$ are used in fitting is to follow CF, which is key to avoid biasing the results. We again use CV to prevent overfitting. We randomly generate five sets of hyper-parameter values, and for each set of hyper-parameter values, we minimize $\min_{\tau(x)} \hat{L}_n\left(\tau; D_{train}\right)$ and determine the number of boosting iterations by early stopping when the following CV error stops decreasing:

$$CV_\tau := \sum_{j,t \in D_{train}} \left(\dot{r}_{j,t+1} - \hat{m}^{(-L(j,t))}\left(\dot{X}_{j,t}\right) - \left(\dot{Z}_i - \hat{e}^{(-L(j,t))}\left(\dot{X}_{j,t}\right)\right)\hat{\tau}^{(-L(j,t))}\left(\dot{X}_{j,t}\right)\right)^2 \tag{4.19}$$

We record the set of hyper-parameter values that yield the smallest CV_τ. Then, we use the best hyper-parameter values to train τ model using all training data by minimizing (4.18), to get the final HTE estimator $\hat{\tau}$. We repeat step 3 below:

- Step 3: Use output from the previous two steps to generate $\hat{m}^{(-L(j,t))}(\dot{X}_{j,t})$ and $\hat{e}^{(-L(j,t))}(\dot{X}_{j,t})$ for all (j,t) in D_{train}. Use xgboost to minimize loss function $\hat{L}_n\left(\tau; D_{train}\right)$ over $\tau(x)$. Again, use CV to control overfitting: use xgboost

Hyper-parameters	Range for random search
max tree depth	$\{3, 4, ..., 20\}$
learning rate	$\{0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.5\}$
bootstrap fraction	$\{0.5, 0.75, 1\}$
feature sampling fraction	$\{0.6, 0.8, 1\}$
γ	$[0, 0.2]$
min leaf weights	$\{1, 2, 3, ..., 20\}$
max delta step	$\{1, 2, 3, ..., 10\}$
max number of trees (iterations)	300
number of trees	Determined by early stopping based on CV loss with a patience of 10 rounds, which means that we terminate training if CV loss stops decreasing for 10 rounds. The number of iterations/trees chosen is then the one minimizing CV loss
early stopping rounds for number of trees	10

Table 4.1: Random search ranges for hyper-parameters to tune

to fit outcome $\dot{r}_{j,t+1}$ using features $\dot{X}_{j,t}$ only. Randomly generate five sets of hyper-parameter values based on ranges specified in Table 4.1. For each set of hyperparameter values, fit five different versions, $\hat{\tau}^{(-l)}$ for folds $l = 1, 2, 3, 4, 5$ and record the CV loss CV_τ. Pick the set of hyper-parameter values that yields the smallest CV_τ. Record the best hyper-parameter values and retrain one single τ model by minimizing $\hat{L}_n(\tau; D_{train})$ with the best set of hyperparameter values over all training data. The resulting model from this last step is our final HTE estimator $\hat{\tau}(x)$.

We summarize the entire R-learner procedure used in our setting in Table 4.2.

4.4 Empirical Results

4.4.1 Data description

Our data period is from Jan 1996 to Dec 2018. All features used in $X_{j,t}$ are listed in Table D.1. We use all stocks listed on NYSE, NASDAQ and AMEX. All price and volume related features in Table D.1 are calculated using CRSP data while all bal-

Steps	Details
Step 0: preparation	1. Conduct random splitting into 5 folds based on months t. 2. Use $L(j,t)$ to denote the resulting mapping from (j,t) to fold ID. 3. $L(j,t)$ is kept the same for use in steps 1, 2 and 3.
Step 1: m model	1. Use xgboost to fit $\dot{r}_{j,t+1}$ using $\dot{X}_{j,t}$ only. 2. Using folds $L(j,t)$ from Step 0 to conduct 5-fold CV, pick the best set of hyper-parameter values that minimizes CV_m from the five randomly generated sets of hyperparameter values based on ranges specified in Table 4.1. 3. Save the final five models with the best hyper-parameter values $\hat{m}^{(-l)}$, $l = 1, 2, 3, 4, 5$.
Step 2: e model	1. Use xgboost to fit continuous treatment $\dot{Z}_{j,t}$ using $\dot{X}_{j,t}$ only. 2. Using folds $L(j,t)$ from Step 0 to conduct 5-fold CV, pick the best hyper-parameter values that minimizes CV_e from five randomly generated sets of hyperparameter values based on ranges specified in Table 4.1. 3. Save the final five models with the best hyper-parameter values $\hat{e}^{(-l)}$, $l = 1, 2, 3, 4, 5$.
Step 3: τ model	1. Take the trained models from previous two steps to generate $\hat{m}^{(-L(j,t))}(\dot{X}_{j,t})$ and $\hat{e}^{(-L(j,t))}(\dot{X}_{j,t})$ for all (j,t) in training set following cross-fitting. Save those out-of-fold predictions for τ model estimation. 2. Use xgboost to minimize $\hat{L}_n(\tau; D_{train})$ over $\tau(x)$. Use the same folds $L(j,t)$ to conduct 5-fold CV in order to control overfitting. Pick the best hyper-parameter values that minimizes CV_τ from five randomly generated sets of hyperparameter values based on ranges specified in Table 4.1. 3. Train a single τ model by minimizing $\hat{L}_n(\tau; D_{train})$ over all data points in training set D_{train} with the best hyper-parameter values obtained in Step 3.2 above. This is our final HTE estimator and we denote it by $\hat{\tau}(x)$.

Table 4.2: R-learner procedure in our empirical exercise

ance sheet related features are calculated using annual report data from Compustat. We merge the monthly data from CRSP and Compustat. Since it is reasonable to believe that market situation evolves over time, we use a 10-year rolling window when applying estimation procedure described in Table 4.2. We do not refit the model every month since the estimates are unlikely to change a lot by shifting the training window by just one month. Instead, we repeat the entire procedure in Table 4.2 at each year end, using the past 10 years' data and we keep the estimated models the same without refitting for 12 months until the next year end. Therefore we obtain an estimated function, $\hat{\tau}$, for each year using only past data (the previous 10 years preceding the current year) to avoid look-ahead bias in our backtest. For each stock month pair (j, t) in our backtest, we make stock-month specific prediction, $\hat{\tau}(X_{j,t})$, using model $\hat{\tau}$ obtained from data in the most recent 10-year rolling window. Since our entire data period is from Jan 1996 to Dec 2018 and we use a 10-year rolling window for training, the first month for our out-of-sample test period is Jan 2006 using feature values from Dec 2015. The test set ends in Dec 2018. Therefore, all results presented in this section are from Jan 2006 to Dec 2018.

4.4.2 Long-short of long-short test

4.4.2.1 Description of the test

Note first that we have no way to observe true $\tau^*(X_{j,t})$ to evaluate the performance of $\hat{\tau}(X_{j,t})$ in terms of predicting $\tau^*(X_{j,t})$. Thus, we choose a very intuitive portfolio sorting test that is similar to what is usually performed in asset pricing papers. If again we use the value factor as the treatment for illustration purposes, the high level idea is to test whether we are able to utilize our $\hat{\tau}(X_{j,t})$ to identify subsets of stocks in which an HML strategy restricted to those subsets have different average returns. If $\hat{\tau}(X_{j,t})$ successfully capture the heterogeneity of the treatment effects, then we expect to see that an HML strategy applied to the subset of stocks with highest

$\hat{\tau}(X_{j,t})$ would have the greatest average returns since highest $\hat{\tau}(X_{j,t})$ suggests that those stocks' future returns respond the strongest to the higher value measures. They are the characteristic responders. On the contrary, if we focus on the subset of stocks that have the smallest $\hat{\tau}(X_{j,t})$, these stocks' cross-sectional returns should be affected the least by value measures or even respond negatively to higher valuation measures. These stocks would be the value traps where an HML strategy won't generate a good expected return. We design a long-short of long-short portfolio test to check whether this is the case in our data.

We describe our test procedure here. At the end of each month t, we sort all stocks into five quintiles based on $\hat{\tau}(X_{j,t})$. We call the five quintiles $\hat{\tau}$ quintiles[7]. Next, within each τ quintile, we perform the usual long-short strategy based on treatment variable considered: sorting based on treatment into five quintiles. Using value as an example, within each tau quintile, we further sort stocks into five quiniles. We call these five quintiles treatment quintiles. We long the treatment quintile with the highest treatment and short the lowest to form the long-short trading strategy within each τ quintile. We also call this trading trategy P5 - P1 since we sort stocks on treatment characteristic into five portfolios, and we long the fifth portfolio while shorting the first portfolio. We hold all five P5-P1 long-short strategies for each tau quintile during month $t+1$ and repeat everything at the end of month $t+1$. We calculat average returns and conduct alpha analysis for the long-short strategies restricted to each tau quintile. As a benchmark, we also repeat our analysis on naive long-short strategies based on the treatment variable for the entire universe of stocks during our test period, ignoring information from $\hat{\tau}(X_{j,t})$'s. For the average returns,

[7]Note that, for some months we might see lots of stocks having exactly the same $\hat{\tau}$. We make sure that the two extreme $\hat{\tau}$ quintiles have stocks in them through the following way. We calculate $\min\hat{\tau}$, $\max\hat{\tau}$ and $20th$, $40th$, $60th$, $80th$ percentiles of $\hat{\tau}$. Then $\hat{\tau}$ 1 quintile is the stocks with $\hat{\tau}(X_{j,t}) \in [\min, quantile_{20\%}]$. l^th quintile, where $l = 2, 3, 4, 5$, is defined similarly. $\hat{\tau}$ 2 quintile: stocks with $\hat{\tau}(X_{j,t}) \in (q_{20\%}, q_{40\%})$; $\hat{\tau}$ 3 quintile: $\hat{\tau}(X_{j,t}) \in [q_{40\%}, q_{60\%})$; $\hat{\tau}$ 4 quintile: $\hat{\tau}(X_{j,t}) \in [q_{60\%}, q_{80\%})$;$\hat{\tau}$ 5 quintile: $\hat{\tau}(X_{j,t}) \in [q_{80\%}, \max]$. This way number of stocks might be different for the 5 $\hat{\tau}$ quintiles and the two extreme quintiles $\hat{tau}$ 1 and $\hat{tau}$ 5 quintiles tend to be larger

we would expect that as we move from the first τ quintile (20% stocks with the smallest $\hat{\tau}(X_{j,t})$) to the fifth τ quintile (20% of stocks with the largest $\hat{\tau}(X_{j,t})$), the long-short strategies' average returns should be roughly increasing. Instead of testing monotonicity, we opt for a simpler approach focusing on the two extreme τ quintiles: we form a strategy by "longing" the long-short strategy restricted to the fifth τ quintile and "shorting" the long-short strategy applied to the first τ quintile. We statistically test whether or not this long-short of long-short strategy will generate positive returns or not and also conduct alpha analysis with respect to standard factors. We call this test "long-short of long-short" test since we take a long position on one long-short strategy while shorting another long-short strategy. Connecting to the simulation plots we show in Figure 4.2, we think that $\hat{\tau}$ 5 quintile consists of the value responders while the $\hat{\tau}$ 1 quintile has the value traps in this case. To some extent, our long-short of long-short test is testing the difference in the slopes or treatment effects between the two groups[8]. We repeat the same test for different treatment variables: value, size, and momentum.

Our long-short of long-short test is designed for two purposes. First, the goal is to test whether or not our $\hat{\tau}(X_{j,t})$ is able to differentiate subsets of stocks for which long-short strategy returns differ. Because observing true treatment effects function $\tau^*(x)$ is impossible, we have to come up with some way to test whether our fitted model $\hat{\tau}$ is useful in that the high $\hat{\tau}(X_{j,t})$ stocks are indeed having larger treatment effects than low $\hat{\tau}(X_{j,t})$ stocks. We choose long-short of long-short test since it is intuitive and has an interpretation of monthly return of a particular trading strategy. Since we always long the long-short strategy within the $\hat{\tau}$ 5 quintile and short the long-short strategy within the $\hat{\tau}$ 1 quintile, we expect long-short of long-short to have positive

[8]Note that a P5-P1 long-short strategy return does not correspond exactly to slope or treatment effects in the sense that it just reflects how much the response has changed, without taking into account how much the treatment variable changes. And as we see in the result tables later, the range of treatment characteristics within each $\hat{\tau}$ quintile can be quite different.

average returns. If it turns out that our test statistic is negative or insignificantly positive, there can be two reasons: 1) the true effects are homogeneous i.e., $\tau^*(x) = \tau$ (assumption (4.6) holds), or 2) the true treatment effects are heterogeneous but our fitted $\hat{\tau}$ are not accurate enough to the point where stocks in $\hat{\tau}$ 5 quintile have higher τ^* than stocks in the $\hat{\tau}$ 1 quintile.

Second, our long-short of long-short test can potentially be implemented by investors to harvest its returns and alphas. We think of two strategies to utilize our HTE estimates if our long-short of long-short test is significantly positive: (1) We could just apply the long-short strategy within the characteristic responders. Use value again as an example. If we have priors that a P5-P1 HML strategy applied to the entire universe is going to have positive average returns, we could just apply P5-P1 strategy to the characteristic responders: the $\hat{\tau}$ 5 quintile with the largest $\hat{\tau}(X_{j,t})$. And we could raise the alarm whenever we consider buying any stocks due to their attractive valuation metric if they are from the characteristic trap subset. (2) We could directly trade the long-short of long-short strategy. However, we note that the long-short of long-short strategy is different than long-short strategy restricted to some subset. We expect the long-short of long-short strategy to not considerably load on the treatment characteristic. For example, if the treatment variable is book-to-market ratio, the long-short strategy formed on BM ratio is betting on value but our long-short of long-short is expected to be roughly neutral to value: the long-short of long-short strategy is purely betting on differences in P5-P1 strategies' returns between the two subsets, which is an approximation to differences in true treatment effects τ^* between the two subsets. As pointed out, P5-P1 strategies are equal weighted in our study and we acknowledge that due to equal weighting, the strategies we discuss here might not be implementable at large scale if the strategies tend to assign too large weights for tiny cap stocks.

When calculating long-short strategy returns, returns of all long portfolios and

short portfolios are equal weighted, and likewise for the naive long-short strategy applied to the entire universe as a benchmark. We next make a few comments about the choice of equal weighting as opposed to value-weighting, which is more common in finance literature. First, during our training stage described in Table 4.2, all data points are equal-weighted. For a highly flexible machine learning procedure like ours, it is important to make sure the test distribution is the same training distribution. Using equal-weighted portfolios in our long-short of long-short test corresponds to applying equal weights to each data point during training, which is obviously what our procedure does based on, for example, the form of the loss function specified in equation (4.18). We surely can apply different weights during training, say, weighting all data points by market cap. However, value weighting during training also has an issue during the test stage: when calculating long-short portfolio returns, only within the long and short portfolio can we ensure stocks are weighted by their size. If we combine the long and short positions together to form a long-short strategy, again, it is not value-weighted anymore. To see this, we give an extreme example with only four stocks in the universe, whose market caps are $1\$, 1\$, 10\$$ and $10\$$. If, according to some signal, we need to long the two 1$ stocks and short the two 10$ stocks, then in the final long-short portfolio formed, we are giving equal weights to all four stocks, regardless of whether we are using equal weighting or value weighting within the long portfolio or short portfolio. This mismatch or inconsistency is problematic since if we decide to use value weight during training, in the training process, the two 10$ stocks get a weighting that are 10 times of the two 1$ stocks; however, during the sorting portfolio tests, all four stocks get equal weights. In fact, there are no way we could predetermine weights for all stocks during training phase such that it is consistent with the usual value-weighted portfolio-sorting tests because we cannot know the composition of the long-short portfolio beforehand. Second, we argue that equal weighting is not a big concern in our case since the purpose of the long-short

of long-short test is not entirely on finding a profitable trading strategy that can be realized at large scale by investors. We are comparing different equal-weighted long-short strategies to check whether the effect of treatment variables on stock returns is heterogeneous, and, in case the treatment effects are heterogeneous, whether our HTE estimation successfully differentiates characteristic responders and traps from the entire universe. Also, we use the equal-weighted version of the unrestricted long-short strategy applied to the entire universe as a benchmark to make sure any comparison is fair. Our main results are equal-weighted and for completeness, we leave the value-weighted results in the appendix.

We discuss the results for different treatments in the following subsections.

4.4.2.2 Results for value as the treatment variable

We present results of the long-short of long-short test in Table 4.3 for $\log BM_{j,t}$ as the treatment variable. To be precise, $\dot{Z} = \log \dot{BM}$ in this case. Table 4.3 has four panels. The top panel shows long-short strategy returns and alphas for the entire universe (column full), each of the five $\hat{\tau}$ quintile (columns τ 1-5). Essentially we apply the P5-P1 HML strategy to different universes: the full universe and each of the five $\hat{\tau}$ quintiles. To be precise, we form long-short strategy by sorting stocks in full universe or each $\hat{\tau}$ quintile into five quintiles (called treatment quintiles) based on the treatment variable, $\log BM_{j,t}$. We buy the stocks in the highest treatment quintile and short the stocks in the lowest treatment quintile as our long-short strategy. The long-short of long-short strategy, as described before, is just that we long the long-short strategy in the fifth $\hat{\tau}$ quintile and short the long-short strategy in the first τ quintile. The results of the long-short of long-short strategy is presented in the last column, "$5-1$". The four rows in the top panel are, respectively, average returns, alpha to CAPM, alpha with respect to Fama-French 3 factor model plus momentum, and alpha to Fama-French 5 factor model plus momentum. The numbers in brackets

are t-stats calculated using Newey-West robust standard errors ([72]). All return and alpha numbers are monthly in percentage. $*$, $**$ and $***$ are used to indicate significance at levels 10%, 5% and 1% respectively. All portfolios involved are equal-weighted.

We can see that in our test period, the HML strategy applied to the full universe actually has an average monthly return of only 0.25% and is not statistically significant. Although after controlling FF3 and FF5 factors it has significant alphas, the magnitudes, 0.49% for FF-3-MOM and 0.39% for FF-5-MOM, are not large[9]. Average returns of the HML strategy for the five $\hat{\tau}$ quintiles roughly follow an increasing pattern when we move from left to right columns. For the fifth τ quintile, average return of the HML strategy is 0.71% which is almost 3 times of the average return if we apply HML to the full universe. And it is statistically significant, whereas for the first τ quintile, the average return is negative and not significant. The observations from alpha analysis are similar to average returns. From the test results we can see our HTE estimation indeed successfully separates out stocks whose returns respond differently to book-to-market ratios in the cross section, especially by looking at the two extreme τ quintiles. Note that the average return of the HML strategy applied to the full universe sits nicely between the HML average returns for $\hat{\tau}$ 1 quintile and $\hat{\tau}$ 5 quintile. If we look at our long-short of long-short test, the difference between returns of HML strategies restricted to the $\hat{\tau}$ 5 quintile and $\hat{\tau}$ 1 quintile is statistically significant and it also has very significant alphas after controlling for standard Fama-French and momentum factors. The magnitude of alphas are between 0.78% to 0.98% depending on the right-hand-side factors. We have a prior that value factor has positive average returns, i.e., the average effects of value on future return should be positive. Thus, we interpret the fifth τ quintile, which consists of 20% of stocks with

[9]The reason that we have statistically significant alphas with respect to value factor on the right hand side as one control could be that we are using equal-weighted portfolios.

the largest $\hat{\tau}$, as value responders in the value case, whereas we call the first τ quintile the value traps that we identified. Interestingly, by looking at $\hat{\tau}$ quintile 2, we note that FF-5-MOM alphas are more monotonic from $\hat{\tau}$ quintile 1 to 5, compared with average returns of the P5-P1 long-short strategy, which might not be a coincidence. As discussed later in section 4.4.2.6, long-short strategies based on single sorts do not control for anything and are susceptible to confounding. If we view FF-5-MOM alpha as a way to control for exposures to standard factors in the long-short strategy return, we should expect FF-5-MOM alphas to be more aligned with our $\hat{\tau}$ and more monotonic from quintile 1 to 5.

In the second panel, we present cross-sectional summary statistics of $\hat{\tau}$ for the full universe and five $\hat{\tau}$ quintiles. We calculate the $25th$ percentile, median, $75th$ percentile, and the average and standard deviation of $\hat{\tau}$ in the cross section for each month, and we average those summary statistics across all months for the full universe and the five $\hat{\tau}$ quintiles. Recall that $\log BM_{j,t}$ is a ratio-based treatment and we do not standardize it but we do standardize the response: future stock returns. We can then interpret $\hat{\tau}$ as how many standard deviations of a change in Y will be induced if the book-to-market ratio increase by 1%. Note first that by construction, the numbers are increasing from τ 1 quintile to τ 5 quintile. Second, there are still quite some variations in $\hat{\tau}$ left within each τ quintile, especially in the two extreme τ quintiles, 1 and 5. This finding suggests that, according to our model, although we categorize stocks into 5 quintiles, it is likely that stocks within each $\hat{\tau}$ quintile still have heterogeneous treatment effects. If we focus solely on average $\hat{\tau}$ for each $\hat{\tau}$ quintile and compare it with the signs of long-short strategy returns, we find they are largely consistent except for $\hat{\tau}$ quintiles 3 and 4. We note that some discrepancies like this are expected: even if our $\hat{\tau}$ are the true τ^*, due to reasons discussed later in section 4.4.2.6, the relationship between $\hat{\tau}$ and long-short returns could be complicated.

In the third panel, we present the spread of treatment characteristics. Because

we sort stocks into five treatment quintiles within each $\hat{\tau}$ quintile and within the full universe, checking the range of the treatment variable within each $\hat{\tau}$ quintile is useful. Take the first column "full" as an example. For each month, we calculate the average of $\dot{Z}$ across all stocks and average them across all months, which yields the results in the first row of the third panel. For the second and third row, we sort all stocks in the full universe into five treatment quintiles based on the treatment $\log BM_{j,t}$, and calculate, for each month, the average of $\dot{Z}$ in the first treatment quintile with the smallest $\dot{Z}$, and the average of $\dot{Z}$ in the fifth quintile with the largest $\dot{Z}$ and then again we average them across all months. For the last row of the third panel, we calculate the spread of treatment $\dot{Z}$ between the fifth treatment quintile and the first treatment quintile: we simply subtract the second row from the third row. We repeat the same calculations for each of the five $\hat{\tau}$ quintiles in columns $\hat{\tau}1$ to $\hat{\tau}5$. We see from Table 4.3 that the spread of $\dot{Z}$ is similar across different $\hat{\tau}$ quintiles. The distributions of treatment variable $\dot{Z}$ are stable across different $\hat{\tau}$ quintiles. Also, recall that the treatment $\log BM_{j,t}$ is not standardized and therefore in the full universe, average $\dot{Z}$ is not zero (-0.69).

In the bottom panel, we present some additional information about the full universe and the five $\hat{\tau}$ quintiles. Again, we calculate each summary statistic in the cross section for the five $\hat{\tau}$ quintiles and the full universe during each month and then average across all months. We include the number of stocks, book-to-market ratios, market equity in billions of dollars, and average monthly volume in million shares. In addition, in order to get some ideas of how stable each $\hat{\tau}$ quintile, we calculate a persistence statistic for each $\hat{\tau}$ quintile. Fix one $\hat{\tau}$ quintile. For each month t, we calculate the percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month $t+1$. Again, we average this percentage across all months. We can see that the average firm sizes across all five τ quintiles are similar from Table 4.3. Average book-to-market ratios are similar for the two extreme τ quintiles.

4.4.2.3 Results for size as the treatment

We next look at the results for size as the treatment variable in Table 4.4. To be precise, $\dot{Z} = \log \dot{M}E$, where ME stands for market equity. To be consistent across different treatment variables, we present in the top panel of Table 4.4 the P5-P1 long-short strategy's returns and alphas. That means instead of the more familiar small-minus-big (SMB), we are actually presenting results of P5-P1 big-minus-small (BMS) strategies restricted to different subsets of the full universe (and the unrestricted full universe). This ensures that signs of average returns of long-short strategies are the same with the signs of treatment effect τ: if we expect smaller firms to earn higher average returns, we should expect P5-P1 BMS strategy applied to the full universe to have negative average returns and $\hat{\tau}$ to be mostly negative too. Keeping using P5-P1 across all treatment variables also ensures that the long-short of long-short test statistic is expected to be positive for all treatment variables.

From the top panel, we can see that the $\hat{\tau}$ 1 quintile has very negative and significant returns and alphas, -1.43% average returns and -1.86% FF-5-MOM alpha. Since we expect that, on average, size negatively affects returns, $\hat{\tau}$ 1 quintile consisting of stocks with the most negative treatment effect is where we expect the treatment characteristic to work the best. Therefore, $\hat{\tau}$ 1 quintile is the characteristic-responder subset and $\hat{\tau}$ 1 quintile has the characteristic traps identified by our procedure. Our procedure works very well especially on identifying size responders: the $\hat{\tau}$ 1 quntile has the most negative average returns and alphas for BMS strategy compared with the other 4 quintiles and the full universe. The difference in magnitudes are huge too: the average return in $\hat{\tau}$ 1 quintile is -1.44%, whereas the next closest number is -0.03% for $\hat{\tau}$ 2 quintile. Note that average BMS returns increase as we move from quintile 1 to 5 except for $\hat{\tau}$ 3 ($\hat{\tau}$ 4 and 5 quintile have very similar average returns). Interestingly, FF-5-MOM alphas are more monotonic compared with average returns. In particular, $\hat{\tau}$ 1 quintile has the most negative FF-5-MOM alpha,

-1.87% while $\hat{\tau}$ 5 quintile has the least negative FF-5-MOM alpha, -0.05%. For our test period, big-minus-small strategy actually has a slightly positive average return for the full universe, though the magnitude, 0.15%, is really small. The FF-5-MOM alpha for BMS strategy on the full universe is negative, -0.24%, and not significant as expected. The FF-5-MOM alpha and average return for the long-short strategy restricted to the $\hat{\tau}$ 1 quintile are more than 7 times larger in magnitude than those of the long-short strategy for the full universe. Our long-short of long-short test is significantly positive and large in magnitude too: the strategy has an average return of 1.54% and 1.80% FF-5-MOM alpha.

For the second panel of Table 4.4, recall that ME is not ratio-based and thus we cross-sectionally standardize the treatment. As a result, we could interpret the values of $\hat{\tau}$ as how many standard deviations of a change in average returns will be caused by a one-standard-deviation increase in the treatment variable. As expected, most $\hat{\tau}$ are negative except for the $\hat{\tau}$ 5 quintile since size factor works well in the past. If we compare the sign of average $\hat{\tau}$ with the sign of average BMS returns, for the $\hat{\tau}$ 3 and 4 quintiles the signs between long-short average returns and average $\hat{\tau}$ are the opposite. As will be discussed in section 4.4.2.6, this discrepancy is somewhat expected. Similar to the observation on the monotonicity of the FF-5-MOM alpha in the value treatment case, signs of the average $\hat{\tau}$ are more consistent with the signs of FF-5-MOM alphas compared with the signs of BMS average returns: except for the $\hat{\tau}$ 5 quintile where the average $\hat{\tau}$ is positive and the FF-5-MOM alpha is negative, the rest of $\hat{\tau}$ quintiles are consistent.

For the last panel, it is striking that $\hat{\tau}$ is so correlated with firm size. The average of market equity is increasing from $\hat{\tau}$ 1 to 5 quintile and the $\hat{\tau}$ 1 quintile's average firm size is only 0.6 billion\$, which suggests that the SMB strategy works the best[10] among the smallest stocks. This discrepancy in firm sizes across different $\hat{\tau}$ quintiles

[10]Equivalently we could say the BMS strategy performs the worst among the smallest stocks

definitely has an impact on how much of the returns and alphas presented here can be realized by investors trading the long-short of long-short strategy strategy when size is the treatment variable. Because during all model training we exclude the treatment variable from our features X, this must result from the fact that we have some features that are correlated with size, such as total asset, which play an important role in the τ model.

4.4.2.4 Results for momentum as the treatment

Lastly, we explain the results for momentum as the treatment variable in Table 4.5. To be precise, $\dot{Z} = \dot{r}_{2m:12m}$, where the subscript denotes trailing 11-month returns from the beginning of month $t-12$ to the end of month $t-2$, following the definition of momentum characteristic. Note that $r_{2m:12m}$ is not ratio-based and thus we cross-sectionally standardize the treatment variable. Again, in the top panel, we present returns and alphas of long-short trading strategies where we always long 20% of stocks with the highest value of treatment and short the 20% of stocks with the lowest treatment. Thus, the top panel shows returns and alphas for winner-minus-loser (WML) long-short strategy restricted to different subsets of the full universe. If we take the prior that the momentum factor has positive average returns in general, we should expect the WML strategy to have positive average returns and the fitted $\hat{\tau}$ to be mostly positive. However, the WML strategy applied to full universe in our test period has a very small positive return of 0.16% and is far from being statistically significant with a t-stats of 0.29. This finding shows that at least the equal-weighted momentum strategy has not been working well for our test period. As expected, the WML strategy for the full universe has a FF-5-MOM alpha of -0.03%, which is very close to 0 and not significant. For all five $\hat{\tau}$ quintiles, the average returns of the WML strategy increase monotonically from -0.61% in quintile 1 to 0.57% in quintile 5. Similarly, the FF-5-MOM alpha increases from $\hat{\tau}$ quintile 1 to 5 monoton-

ically. A WML strategy restricted to momentum responders identified in $\hat{\tau}$ quintile 5 generates an average return that is more than 3 times higher than WML on the full universe (0.57% vs. 0.16%), though neither is statistically significant. Interestingly, the WML strategy restricted to the $\hat{\tau}$ 1 quintile (the "momentum traps" quintile) has significantly negative FF-5-MOM alpha, which means that shorting the WML long-short strategy or a loser-minus-winner (LMW) strategy restricted to the "momentum traps" identified by our procedure would generate a significantly positive alpha with respect to FF-5-MOM. Consistent with our previous results for value and size, we find the long-short of long-short test remains positively significant and large in magnitude: the average return is 1.18%, and the FF-5-MOM alpha is 1.65%.

For the second panel, we find that the signs of avg $\hat{\tau}$ and median $\hat{\tau}$ are mostly consistent with the signs of FF-5-MOM alpha across different $\hat{\tau}$ quintiles except for $\hat{\tau}$ quintile 4, where the FF-5-MOM alpha is positive, avg $\hat{\tau}$ is barely negative, and median $\hat{\tau}$ is essentially 0. This finding is consistent with the cases in which value or size is used as the treatment variable. For the third panel, we see that interestingly the spread of treatment is smaller within the momentum traps (the $\hat{\tau}$ 1 quintile) compared with the momentum responders (the $\hat{\tau}$ 5 quintile).

For the last panel, we found that average market size decreases monotonically from momentum traps ($\hat{\tau}$ 1 quintile) to momentum responders ($\hat{\tau}$ 5 quintile). Although for the momentum responders, momentum strategy does not perform well for our test period in the sense that neither the average return nor the FF-5-MOM alpha is significantly positive, it is the best performing momentum strategy relative to other quintiles, and it also has the smallest average firm size.

4.4.2.5 Summary of results

We briefly summarize the results of our long-short of long-short test for the treatment variables we tried: value, size, and momentum characteristics. Before reading any

results, the priors on the sign of the three characteristics' average effects are different: for value and momentum, we expect the treatment effects on average to be positive, and for size, we expect the effects to be negative on average. Therefore, for value and momentum, the $\hat{\tau}$ quintile 5 is the characteristic-responder subset, whereas for size $\hat{\tau}$ as the treatment, the $\hat{\tau}$ 1 quintile consists of characteristic responders identified by our HTE estimation procedure. Our test results in Tables 4.3, 4.4 and 4.5 show the average returns of the long-short strategy restricted to characteristic responders are larger than the average returns of the naive long-short strategy on the full universe in all three cases of different treatment variables. For the long-short strategy restricted on characteristic responders, except for momentum treatment case, we always have statistically significant average returns and FF-5-MOM alphas. Those findings above shows that when investing in P5-P1 strategies based on those factors/characteristics, investors could focus on characteristic responders identified by our procedure to boost their performance. P5-P1 long-short strategies applied to the $\hat{\tau}$ 5 quintile always outperform long-short strategies restricted to the $\hat{\tau}$ 1 quintile, and the difference in average returns are statistically significant in all three treatment cases, based on our long-short of long-short test results. Moreover, the long-short of long-short strategies always have significantly positive FF-5-MOM alphas, which suggests that the out-performance of P5-P1 strategy between $\hat{\tau}$ 5 and $\hat{\tau}$ 1 quintile cannot be explained by standard Fama French and momentum factors. The magnitudes are economically significant too: the average returns of our long-short of long-short strategies range from 0.77% to 1.54%, and the FF-5-MOM alphas of the long-short of long-short strategies range from 0.98% to 1.80%, depending on the treatment variables.

Note that as mentioned before, in the long-short of long-short tests, we only test our procedure's ability to predict heterogeneity in treatment effects. Even if the average of treatment effects is zero, as long as the effects are heterogeneous enough[11],

[11]Recall that responders and traps identified by our procedure are defined as the two extreme $\hat{\tau}$

we expect our R-learner procedure to produce significantly positive returns in the long-short of long-short test. In particular, in our out-of-sample test period, the simple P5-P1 long-short strategies applied to the entire universe all have average returns that are statistically insignificant, according to the "full" column in Tables 4.3, 4.4, and 4.5, which suggests that the average treatment effects are non-existent (0) or very small in magnitude. However, our long-short of long-short strategy has positive average returns in all three cases, which suggests for the three treatment variables we've tried: (1) treatment effects are heterogeneous in all three cases and (2) our procedure successfully captures the heterogeneity and is accurate enough in predicting treatment effects. On the other hand, it is possible that for some treatment variable, the treatment effects are both homogeneous and very strong. In this case, we would see that the naive long-short strategy applied on the full universe has significantly positive average return, whereas our long-short of long-short test does not have significantly positive average return. However, we didn't find it to be true in the data when size, value or momentum is used as the treatment variable.

From the second panel in Tables 4.3, 4.4, and 4.5, the signs of average $\hat{\tau}$ or median $\hat{\tau}$ are largely consistent with the signs of long-short strategies' average returns across different $\hat{\tau}$ quintiles. In particular, the signs of average $\hat{\tau}$ are more consistent with the signs of FF-5-MOM alphas of long-short strategies for different $\hat{\tau}$ quintiles. We also find that, in general, within extreme quintiles ($\hat{\tau}$ 1 and $\hat{\tau}$ 5 quintiles), estimated treatment effects $\hat{\tau}$ tend to be more heterogeneous than within the middle three $\hat{\tau}$ quintiles.

Average firm sizes across different $\hat{\tau}$ quintiles are quite consistent, using value as the treatment variable, whereas with momentum as the treatment variable, firm

quintiles, $\hat{\tau}$ 1 and $\hat{\tau}$ 5 quintiles. Which one consists of characteristic traps depends on the prior on the overall treatment effects. For example, for size as the treatment, we expect overall treatment effect to be negative due to SMB factor documented by [40]. Then $\hat{\tau}$ 1 quintile consists of the characteristic responders or "size responders", because it has the smallest or most negative $\hat{\tau}$'s, whereas $\hat{\tau}$ 5 quintile consists of size trap stocks

size negatively correlates with $\hat{\tau}$. Using size as the treatment, firm size positively correlates with $\hat{\tau}$. This correlation between size and $\hat{\tau}$ might result in characteristic reponders or traps mainly consisting of tiny stocks, which could potentially affect the practical implementation of long-short of long-short as an investment strategy for size and momentum cases.

Lastly, we make one additional comment about our long-short of long-short tests. Although we focus on the long-short of long-short strategies in this section, the average returns or alphas of our long-short of long-short strategy are never the loss function or evaluation metric during our model training and cross-validation stage. Our loss function used during the training of τ models is given by equation (4.13) and cross-validation criterion (evaluation metric) is given by equation (4.19). The fact that all of our long-short of long-short tests are significant even though it is never directly used as the training target or evaluation metric shows the robustness of our procedure and suggests that to some extent we capture the true heterogeneity of treatment effects. On the other hand, machine learning calls for directly minimizing the loss function we care about. Incorporating average returns or alphas of our long-short of long-short strategies into loss function is difficult in our case, because our long-short of long-short test relies on the $\hat{\tau}$ predicted by the fitted τ model, which is unknown until the training finishes. However, if we are really interested in investing in the long-short of long-short strategies and want to improve their performance as much as possible, we could at least do cross validation based on the average returns of long-short of long-short strategies in the validation set, which could potentially boost the performance of the strategies. To be more conservative, we choose to not to tune the models that way and stay with the most standard loss functions and cross-validation evaluation metrics when implementing the R-learner.

4.4.2.6 Subtleties on $\hat{\tau}$ vs. long-short strategy returns

To some extent, our long-short of long-short test checks whether $\hat{\tau}$ has successfully differentiated stocks with different true τ^*. We use average returns of long-short strategy restricted to different $\hat{\tau}$ quintiles as approximations to some aggregate measure of true τ^* for the different $\hat{\tau}$ quintiles. If the long-short of long-short strategy has significantly positive returns, then we could conclude that stocks in the $\hat{\tau}$ 5 quintile generally have higher τ^* than stocks in the $\hat{\tau}$ 1 quintile. As we discuss below, however, fixing one $\hat{\tau}$ quintile, several gaps exist between τ^* of stocks in a $\hat{\tau}$ quintile and the returns of the long-short strategy restricted to that $\hat{\tau}$ quintile. Firstly, our measure of τ controls for a lot of firm characteristics listed in Table D.1 while long-short strategy are just single-sort trading strategies based on treatment variable, which is susceptible to confounders. Unfortunately there is no way we could control for that many potential confounders at one time in the sorting portfolio approach.

Second, although our HTE estimation relaxes some standard assumptions, we still assume that treatment affects returns linearly (assumption (4.5)). Similar to the comparison between the Fama-Macbeth regression approach and the sorting portfolio approach in asset pricing literature, long-short strategy returns are more non-parametric for the relationship between treatment and returns. If the true relationship is non-linear for some feature values of X, it is possible that τ^* for one quintile is quite different from the long-short strategy returns restricted to that quintile.

Third, long-short returns could be viewed as differences in response variables, whereas treatment effect τ^* is the change in response per unit change of treatment. As we will see later in the results, it is possible that the ranges of treatment variables are different for different $\hat{\tau}$ quintiles. For example, two different quintiles could have exactly the same τ^* but very different long-short strategy average returns only because one quintile has larger spread of treatment variable than the other.

Lastly, the fact that stocks inside one $\hat{\tau}$ quintile could have different τ^* further

complicates the matter. As we see later in the empirical results, within one $\hat{\tau}$ quintile, in general $\hat{\tau}$ is still quite heterogeneous, which suggests the true τ^*'s are not homogeneous between stocks in that $\hat{\tau}$ quintile. This further obscures the connections between τ^* and long-short strategies' average returns. For example, if a P5-P1 long-short strategy restricted to a quintile has close to 0 return, what should τ^* look like for stocks in the quintile is not clear: they could have some stocks with very positive τ^*, some with very negative τ^* and the average τ^* for stocks with the quintile could be positive, negative or close to 0. In our empirical results, we will be presenting summary statistics of the cross-sectional distribution of $\hat{\tau}$ within each $\hat{\tau}$ quintile.

For the reasons above, we should not expect that some aggregate measure of $\hat{\tau}$ such as mean or median for different $\hat{\tau}$ quintiles to be good predictors for long-short strategy returns restricted to different $\hat{\tau}$ quintiles, because even if $\hat{\tau}(x)$ equals to $\tau^*(x)$ exactly for $\forall x$, the relationship between average or median τ^* for stocks in one quintile and the long-short strategy returns within that quintile could be complicated. Therefore, we do not test whether mean or median $\hat{\tau}$ within each $\hat{\tau}$ quintile has predictive power for the P5-P1 long-short strategy restricted to each $\hat{\tau}$ quintile. Our long-short of long-short tests are less ambitious: we expect the $\hat{\tau}$ 5 quintile to have significantly higher long-short strategy returns than the $\hat{\tau}$ 1 quintile and we try to test it in the data. In addition to the difference between two extreme $\hat{\tau}$ quintiles, we also look at whether long-short average returns are roughly monotonically increasing from $\hat{\tau}$ quintile 1 to 5. However, we do not conduct a rigorous statistical test on the monotonicity.

4.5 Interpreting Heterogeneous Treatment Effects

4.5.1 Important Features for HTE

One advantage of our HTE approach together with the choice of gradient boosting is that we could interpret our results in many different ways. We present some analyses in this section. First, we show the most important features for our τ model. Tree-based models, including gradient boosted trees, often have very intuitive ways to calculate feature importance. The idea is that when the gradient boosting tree algorithm decides on which variable to split, it chooses the variable with the largest reduction in training loss. For each variable, we could sum up the training loss reduction whenever the algorithm splits on this variable, and this total loss reduction by each feature could be thought of as its feature importance. Usually we normalize the sum of all feature importance values to be 1 such that the feature importance is in relative terms. Calculating feature importance is built into most of gradient boosting packages. We directly use feature importance function from xgboost in our analysis. Because we have 13 rolling training windows of 10 years each[12], we first calculate the feature importance value described above for each feature and during each training window. Then, for each feature, we average the feature importance values across all 13 training windows to get a final "average relative gain" for the feature. The average relative gains for all features add up to 1 by construction. Average relative gains shows–in an average training window–the percentage of total loss reduction achieved could be attributed to each feature. We list the top 10 features based on this average feature importance measure for $\hat{\tau}$ models when the treatment variable is value, size, and momentum, in Tables 4.6, 4.7, and 4.8 respectively. Note that we could apply

[12]The first training window is from 199601 to 199512, and the last training window is from 200801 to 201712

this analysis to $\hat{m}$ and $\hat{e}$ models as well, but since HTE are captured by the $\hat{\tau}$ model, we focus on that here. We explain the results in the following subsections.

4.5.1.1 Value

We can see the sum of the top 10 features' average relative gains add up to around 52%, which means the other features also contribute almost half to determining the heterogeneity in treatment effects. To further understand how $\hat{\tau}$ depends on the top three features on the list, we conduct the following analysis. Inspired by portfolio sorting, at the end of month t, we calculate break points that sort stocks into five quintiles based on feature $X^{(1)}$, break points of $X^{(2)}$ to sort stocks into five quintiles, and break points of $X^{(3)}$ to sort stocks into two halves. Those break points will group all stocks into $5 \times 5 \times 2 = 50$ portfolios. We calculate the average of $\hat{\tau}$ for each of the 50 portfolios at each month end, and then average again across all months for the 50 portfolios. Essentially, we form unconditional sorts using $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$ and calculate the equal-weighted $\hat{\tau}$ of all portfolios. This way we could directly see how τ model depends on the three most important features to a level that is not too detailed but meaningful. We plot the results for case in which value is the treatment in Figure 4.3.

We next interpret some of the findings in Figure 4.3. By looking at the most important three covariates, it seems that the value responders are mainly the smaller stocks (first two quintiles) with trailing 2-12 months returns in the middle range (middle three quitiles) and past month's return in the top half. Also note that this plot only reflects contributions from the three most important features and that even the top 10 features together only explain 50% of the heterogeneity in treatment effects. Non-top-10 features help a great deal in defining the value traps and responders.

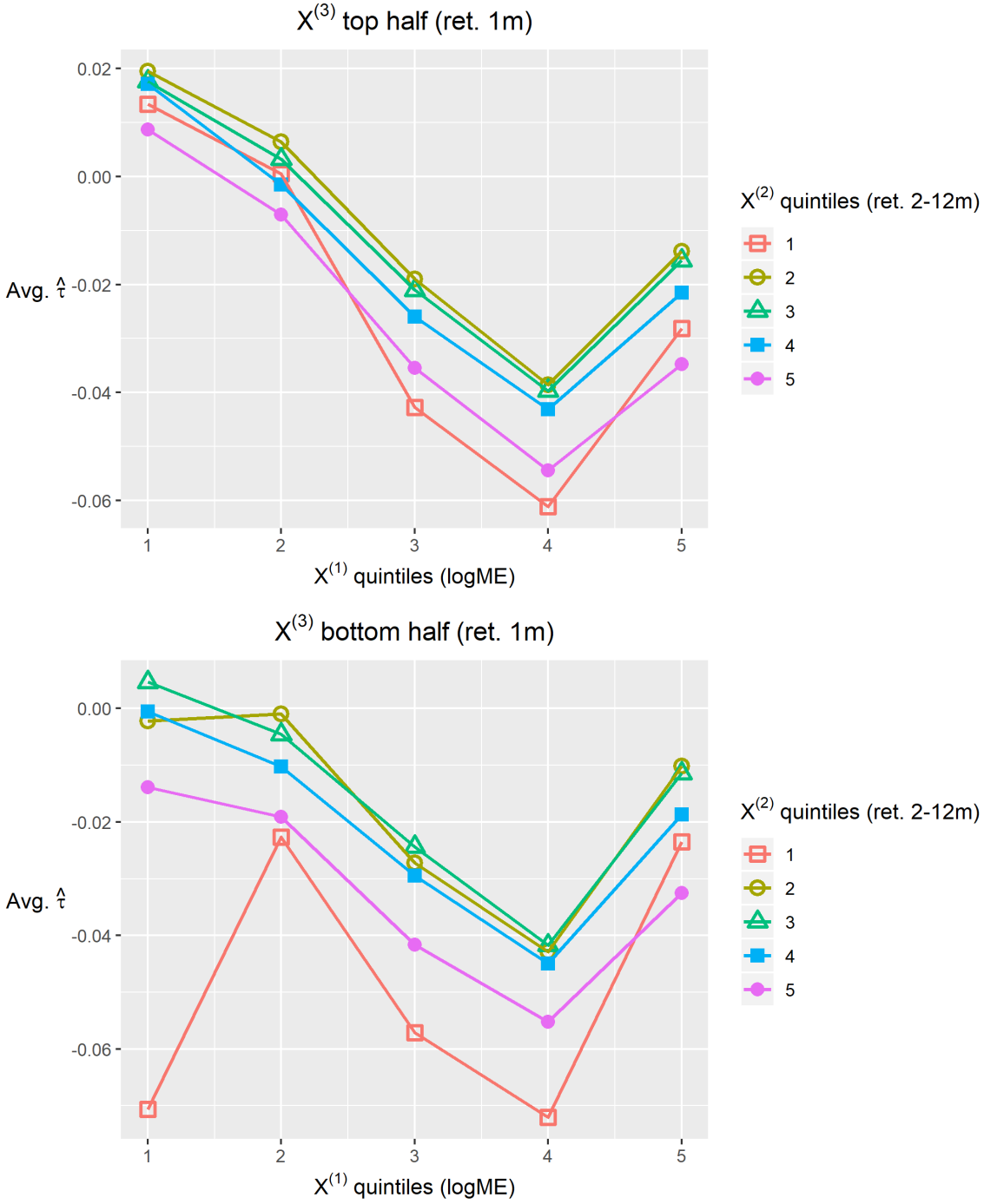

Figure 4.3: $\hat{\tau}$ as a function for the most important 3 features based on feature importance of the τ models fitted in the rolling training windows (value as the treatment).

4.5.1.2 Size

We can see that the sum of the top 10 features' average relative gains add up to around 69%, which is more top heavy compared with $\hat{\tau}$ with value as the treatment. The fact that trailing returns, past month return and past 2-12 month return, are the two most important features suggests that momentum and size have interesting interactions.

We repeat the analysis of plotting $\hat{\tau}$ against quintiles of the most two important features, separately for stocks whose third most important feature value is in the top 50% and bottom 50%. We show the results for size as the treatment in Figure 4.4.

4.5.1.3 Momentum

First, note that the sum of top 10 features' average relative gain add up to around 68%, which is similar to the case where size is the treatment. Also note that size is the most important feature for the heterogeneity effect when momentum is the treatment, which is not surprising at all considering trailing returns are the two most important features for cases in which size is the treatment. The total contribution from the top three features is 46%, which is close to half and the highest among the three treatment variable cases we study. We repeat the same plotting exercise as in the previous section.

From Figure 4.5, we note a striking similarity of the findings with [59] on how the momentum treatment effect varies with firm size. [59] study, in particular, how the momentum long-short strategy's profitability varies with firm size and the number of analysts covering the stocks, in order to test the idea of gradual information spreading among investors. The data period in their study is 1980-1996 and for that period momentum long-short strategy on the full universe works much better than our period (0.53% in their paper vs. 0.16% in Table 4.5). By sorting all firms into 10

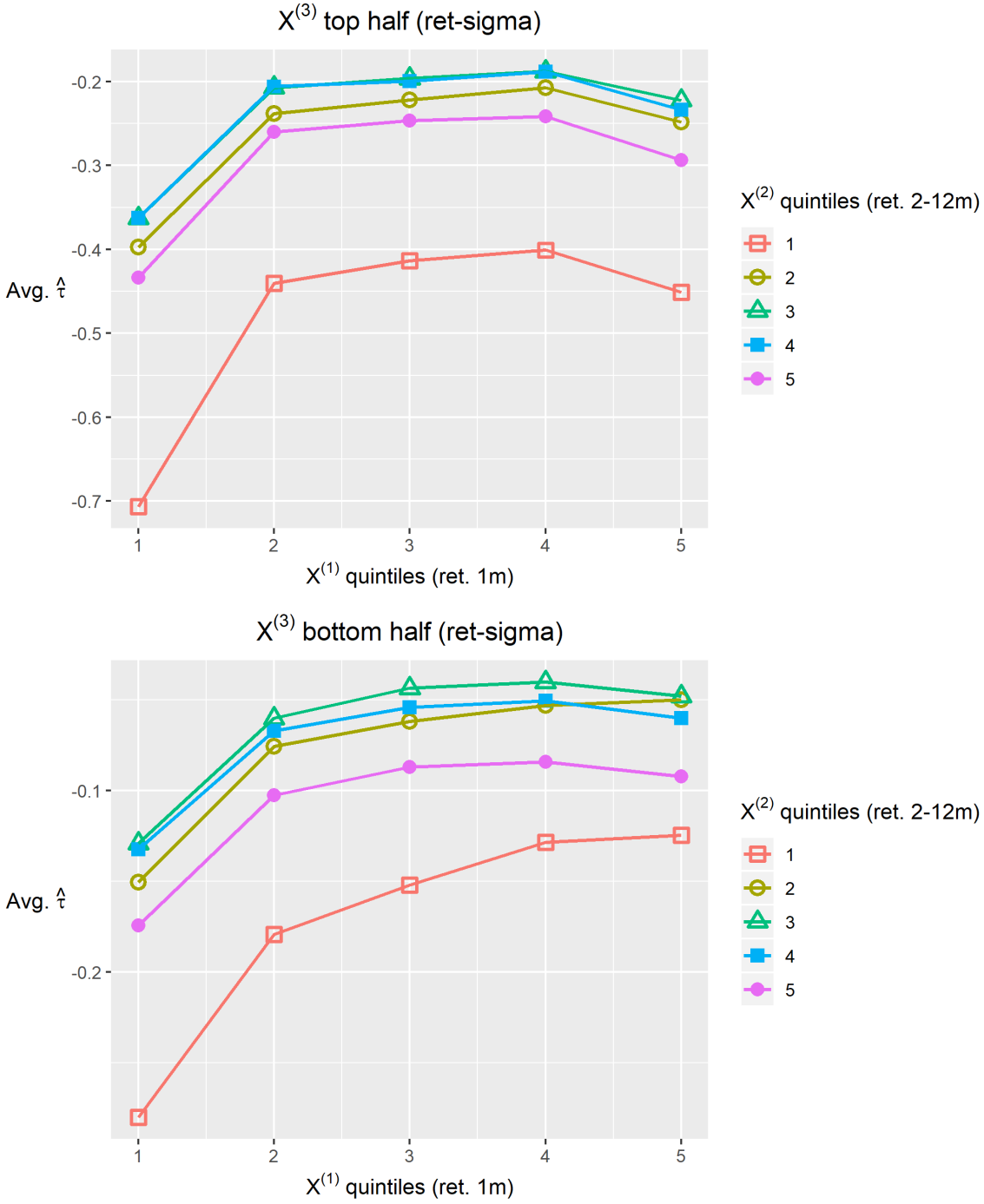

Figure 4.4: $\hat{\tau}$ as a function for the most important 3 features based on feature importance of the τ models fitted in the rolling training windows (size as the treatment).

deciles based on firm sizes and testing P3-P1[13] momentum long-short strategy within each decile, they find the momentum long-short strategy performs the worst for the smallest two size deciles and performs the best for the third smallest decile. The strategy's performance quickly decreases from the third to the largest decile. The general patterns in Figure 4.5 is surprisingly similar to what [59] found: the first two size deciles correspond to the first quintile in our case. The trends are mostly the same though for some lines in the plot, the peak of profitability shows up in the third instead of the second size quintile. This observation is a bit surprising because our data periods overlap only slightly and the ways we form our momentum strategies are different: [59] follows [60] and uses P3-P1 while our P5-P1 trading strategy is based on momentum characteristic ret.2-12m, which follows Carhart's definition in [26].

Similar to our discussions over the relationship between $\hat{\tau}$ and long-short strategy returns, we note here that [59]'s findings are based only on long-short strategy returns, whereas our predicted $\hat{\tau}$ plotted here are the treatment effect estimates controlling for many other features in Table D.1 in a non-linear fashion (recall the e, m models in equation (4.11 and our assumption from equation (4.1)). The portfolio sorting approach cannot control for so many covariates at the same time while linear regression can include more controls as additional regressors but it is susceptible to misspecification of the linear functional forms. Our approach has an advantage in that aspect.

Essentially, we confirm [59]'s empirical findings with more recent data and more controls for potential confounders. Our procedure to arrive at this finding is completely automated without any priors regarding which feature is going to be important and in what fashion. More importantly, in addition to confirming their findings, we are able to provide new insights: the second most important feature is past return

[13]Different than our P5-P1, they sort into three portfolios based on trailing returns and long the 33% of stocks with the largest trailing returns and short the 33% of stocks with the worst trailing returns.

volatility "ret-sigma" and once we move past the smallest stocks, the momentum strategy should work the best (or least poorly in our period) when ret-sigma are among the highest quintile. Coming up with a theory for this observation is not part of the main goals of this chapter, but we think that this could support the idea of gradual information diffusion in [59] in that more volatile stocks probably have more restricted information flow. [59] also find that controlling for size, momentum strategy works better among stocks with low analyst coverage, and it is reasonable to believe that low analyst coverage[14] is often associated with high volatility. Lastly, we note that, despite the similarity in trend, our predicted $\hat{\tau}$ are shifting into negative regions, and most average $\hat{\tau}$ are negative here because momentum strategy applied to the full universe performs much worse than the previous data period used in [59].

4.5.2 Impact of Top 3 Ex-post Important Features on HTE

In this subsection, we investigate the contributions of the three most important (ex-post) features. We check whether our long-short of long-short strategies can still provide alphas after controlling for simple interaction terms between treatment and the top three ex-post features. Given that most of the time the patterns shown in Figures like 4.3 are non-linear, we do not expect simple interactions to absorb our results.

To some extent, our HTE model focuses on finding interactions between functions of other features X and the treatment Z, as described in the term $\tau^*(X)Z$ in equation (4.1). It is natural to control for some simple interaction terms between some features and the treatment variable to see whether our results of long-short of long-short tests still hold. The way we conduct this test is to rerun the same tests, but instead of calculating alphas with respect to standard Fama-French and momentum factors on the right hand side, we also control for six more terms explained below. Denote

[14]We do not collect number of analyst coverage as a control in X.

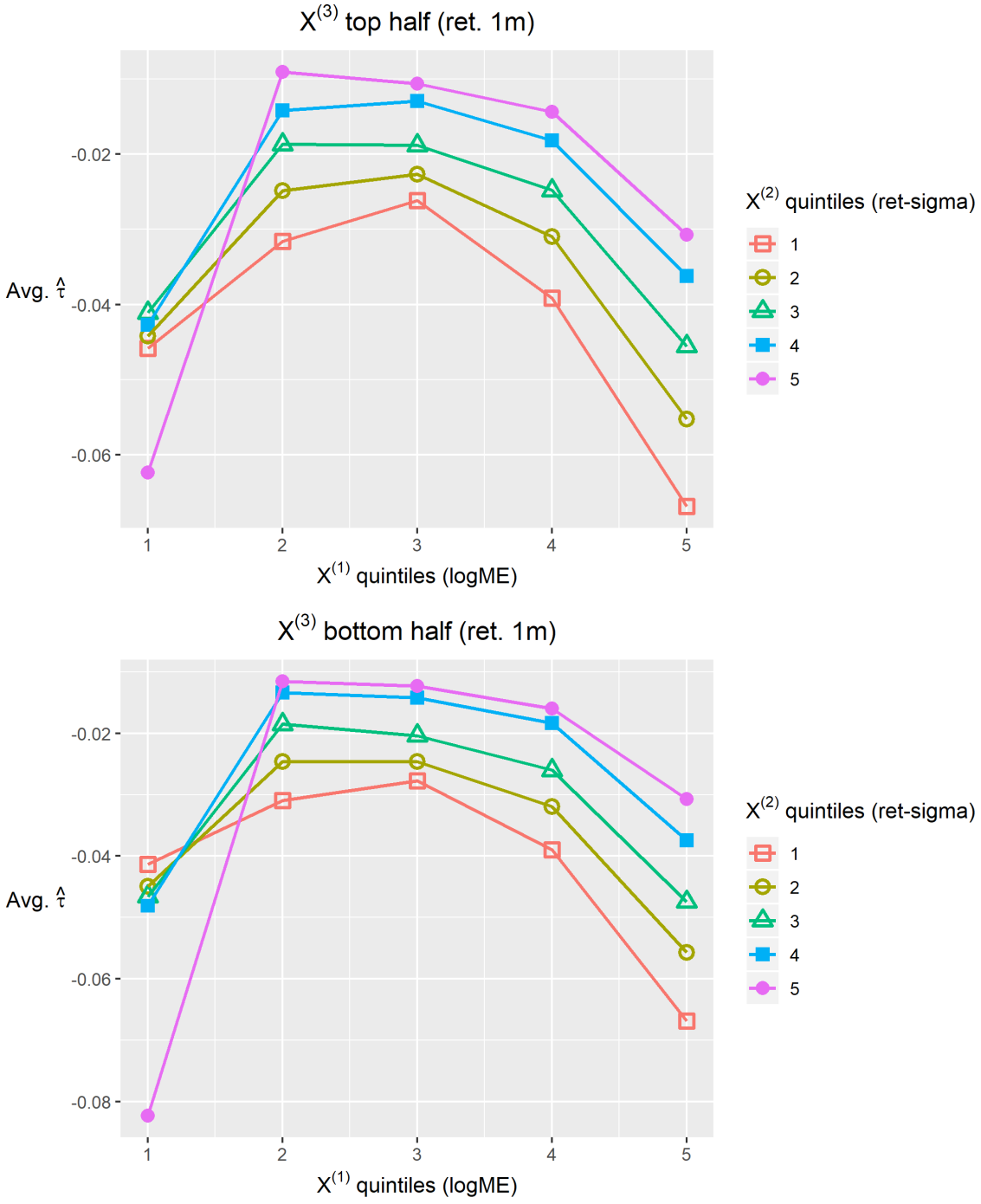

Figure 4.5: $\hat{\tau}$ as a function for the most important 3 features based on feature importance of the τ models fitted in the rolling training windows (momentum as the treatment).

the most important feature according to average relative gain by $X^{(1)}$. At each month end, we sort into five quintiles based on this feature $X^{(1)}$ instead of $\hat{\tau}$. We focus on the two $X^{(1)}$ quintiles with the highest and lowest $X^{(1)}$ and form a P5-P1 long-short trading strategy (based on treatment) within each extreme $X^{(1)}$ quintile as we did before for each $\hat{\tau}$ quintile. We repeat this process at each month end and therefore we have gotten two time series of long-short strategy restricted to two extreme $X^{(1)}$ quintiles. Similarly, we calculate the same two time series for the second and third most important features $X^{(2)}$ and $X^{(3)}$ to get four more time series of long-short strategy returns. We include those six time series on the right hand side to capture interaction effects between the three most important (ex-post) features and our treatment characteristics. We rerun the alpha analyses for the long-short of long-short tests by the regression equation below:

$$\begin{aligned} r_{ls(\hat{\tau}(5))}(t) - r_{ls(\hat{\tau}(1))}(t) =& \alpha + \beta_1 \left(Mkt(t) - r_{RF}(t)\right) + \beta_2 SMB_{full}(t) + \beta_3 HML_{full}(t) + \\ & \beta_4 CMA_{full}(t) + \beta_5 RMW_{full}(t) + \beta_6 WML_{full}(t) + \\ & \beta_7 r_{ls\left(X^{(1)}(5)\right)}(t) + \beta_8 r_{ls\left(X^{(1)}(1)\right)}(t) + \beta_9 r_{ls\left(X^{(2)}(5)\right)}(t) + \\ & \beta_{10} r_{ls\left(X^{(2)}(1)\right)}(t) + \beta_{11} r_{ls\left(X^{(3)}(5)\right)}(t) + \beta_{12} r_{ls\left(X^{(3)}(1)\right)}(t) + \epsilon(t) \end{aligned} \tag{4.20}$$

, where we use $r_{ls(\hat{\tau}(5))}(t)$ to denote returns of the P5-P1 long-short strategy formed based on treatment restricted to $\hat{\tau}$ quintile 5 in month t. Similarly, $r_{ls(\hat{\tau}(1))}(t)$ denotes the long-short strategy formed based on the treatment but restricted to $\hat{\tau}$ quintile 1 in month t. The left hand side is therefore our long-short of long-short returns. On the right hand side, we first have the FF-5 factors plus momentum. We use the subscript "full" to denote that the standard factors are applied to the full universe. $r_{ls(X^{(i)}(j))}(t)$ denotes the long-short strategy return restricted to X^i quintile j during month t, where $i = 1, 2, 3$ and $j = 1, 5$. We want to test whether α in equation (4.20) is still significantly positive after adding the six additional controls. Recall that the

siginificant FF-5-MOM alphas shown in Tables 4.3, 4.4, and 4.5 are obtained by the following regression:

$$r_{ls(\hat{\tau}(5))}(t) - r_{ls(\hat{\tau}(1))}(t) = \alpha + \beta_1 Mkt(t) + \beta_2 SMB_{full}(t) + \beta_3 HML_{full}(t) + \beta_4 CMA_{full}(t) + \beta_5 RMW_{full}(t) + \beta_6 WML_{full}(t) + \epsilon(t). \quad (4.21)$$

This test looks at whether or not we can explain the α in the long-short of long-short strategy by the simple long-short strategies formed on the extreme quintiles of the most three important features, if we can look ahead and know the three ex-post most important features. If the most three important features are highly correlated with $\hat{\tau}$ in the sense that stocks in $\hat{\tau}$ quintile 5 and 1 have considerable overlap with stocks in extreme quintiles by $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$, then α in equation (4.20) is likely to become insignificant. We present the results for the three treatment variables in Table 4.9 below. According to Table 4.7, the top three features alone explain 40% of the loss reductions achieved in training the τ model, which is more than 27% for the value case.

We can see that extreme quintiles by the most important three features take away some of the alpha from the long-short of long-short strategy, but for all three treatment variables, long-short of long-short alphas remain positively significant. Note that the size treatment experiences the greatest reduction in alphas.

4.6 Impact of Data Preprocessing

In this section, we conduct one robustness check to test the role of our cross-sectional standardization denoted by the $\dot{}$ notation in equation (4.15). Feature standardization or normalization is very popular in the machine learning community and is often recommended before training. For our long-short of long-short tests, we are only concerned about the cross-sectional relationship between stock returns and treatment

variables. The time-series treatment effect of the treatments on future stock returns can be different from how the treatments affect future returns in the cross section. Therefore, the time series variations in the data could just add noise to our model training, and we decide to take them away by cross-sectionally demeaning treatments, controls, and stock returns if they are not ratio-based: we subtract from each variable its cross-sectional mean in each month. As mentioned before, this cross-sectional demeaning can be thought of as a preprocessing procedure corresponding to adding time fixed effects in panel regression in linear models. To make the data from different period more comparable, for non-ratio based controls and treatment variables, we further cross-sectionally standardize each variable so that every non-ratio based variable has mean 0 and standard deviation 1 for each month. The $\dot{}$ operation leaves ratio-based variables unchanged since they are already comparable across different months.

Because we are only interested in cross-sectional relationships, we expect our preprocessing to improve long-short of long-short test results. We show the effects of preprocessing in this section by comparing the backtest results without any standardization with our main results. In Tables 4.10, 4.11, and 4.12, we present long-short of long-short results with raw data without any standardization for cases in which value, size, and momentum are used as treatment variables, respectively.

We can see that all three tables have insignificant long-short of long-short average returns and alphas, whereas all long-short of long-short returns and alphas are significantly positive in Tables 4.3, 4.4, and 4.5. This finding shows our cross-sectional standardization lets the model focus on cross-sectional relationship only and boost the performance.

4.7 Conclusions

In this work, we study heterogeneous effects of three well-known and standard firm characteristics on future stock returns in the cross section, controlling for a large number of potential confounders (features) at one time. We design a long-short of long-short test to check whether our fitted models can predict heterogeneity in the treatment characteristics' effects, and the results are both statistically and economically significant. We provide interpretations and visualizations on what the models have captured to explain the heterogeneity in the treatment effects.

	full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.25	-0.06	-0.16	0.26	0.18	0.71**	0.77**
	[0.99]	[-0.14]	[-0.65]	[1.15]	[0.77]	[2.15]	[2.17]
CAPM Alpha	0.29	-0.10	-0.14	0.31	0.27	0.80**	0.89***
	[1.13]	[-0.25]	[-0.54]	[1.34]	[1.16]	[2.37]	[2.63]
FF-3-MOM Alpha	0.49***	0.19	0.04	0.47***	0.40**	0.97***	0.78**
	[3.04]	[0.63]	[0.23]	[2.88]	[2.23]	[3.42]	[2.38]
FF-5-MOM Alpha	0.39**	-0.13	-0.06	0.47***	0.35**	0.85***	0.98***
	[2.30]	[-0.42]	[-0.38]	[3.02]	[1.99]	[2.84]	[3.11]
τ 25%	-0.08	-0.22	-0.08	-0.03	0.01	0.07	
τ 50%	-0.02	-0.14	-0.07	-0.02	0.02	0.11	
τ 75%	0.03	-0.11	-0.05	-0.01	0.03	0.18	
avg τ	-0.02	-0.20	-0.07	-0.02	0.02	0.15	
sd τ	0.17	0.19	0.01	0.01	0.01	0.14	
avg $\dot{Z}$	-0.69	-0.63	-0.69	-0.72	-0.71	-0.69	
avg $\dot{Z}$ in 1st quintile	-1.98	-1.94	-1.89	-1.87	-1.89	-2.23	
avg $\dot{Z}$ in 5th quintile	0.38	0.46	0.27	0.25	0.30	0.54	
avg $\dot{Z}$ P5 - P1	2.36	2.40	2.15	2.11	2.19	2.77	
number of stocks	3536.90	734.32	656.14	718.12	708.52	719.81	
avg. BM ratio	0.84	1.17	0.69	0.66	0.68	0.97	
avg. ME	5.08	3.88	5.11	5.38	4.39	6.52	
avg. Monthly Vol	25.27	27.87	23.76	23.60	19.82	31.05	
persistence		0.73	0.62	0.60	0.63	0.73	

Table 4.3: **Long-short of long-short test results for value as the treatment using equal weighting** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable book-to-market ratios but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.15	-1.43^{**}	-0.04	0.31	0.12	0.11	1.54^{***}
	[0.43]	[-2.48]	[-0.12]	[1.29]	[0.55]	[0.45]	[2.60]
CAPM Alpha	0.14	-1.66^{***}	-0.32	0.14	0.08	0.13	1.79^{***}
	[0.40]	[-2.80]	[-1.11]	[0.54]	[0.36]	[0.55]	[2.91]
FF-3-MOM Alpha	0.03	-1.70^{***}	-0.39	0.06	-0.02	0.00	1.70^{***}
	[0.11]	[-3.03]	[-1.36]	[0.26]	[-0.12]	[-0.01]	[2.94]
FF-5-MOM Alpha	-0.24	-1.86^{***}	-0.49^{*}	-0.12	-0.21	-0.05	1.80^{***}
	[-0.75]	[-2.96]	[-1.70]	[-0.50]	[-1.17]	[-0.30]	[2.85]
τ 25%	-0.32	-0.79	-0.32	-0.16	-0.05	0.03	
τ 50%	-0.13	-0.60	-0.26	-0.13	-0.03	0.05	
τ 75%	-0.01	-0.47	-0.22	-0.10	-0.01	0.07	
avg τ	-0.21	-0.68	-0.27	-0.13	-0.03	0.06	
sd τ	0.30	0.31	0.06	0.03	0.02	0.04	
avg $\dot{Z}$	0.00	-0.88	-0.27	0.15	0.40	0.61	
avg $\dot{Z}$ in 1st quintile	-1.35	-1.86	-1.30	-0.94	-0.69	-0.51	
avg $\dot{Z}$ in 5th quintile	1.43	0.31	0.89	1.33	1.58	1.79	
avg $\dot{Z}$ P5 - P1	2.78	2.17	2.19	2.26	2.28	2.30	
number of stocks	3902.94	781.07	777.73	780.65	777.88	785.61	
avg. BM ratio	0.77	1.09	0.79	0.69	0.63	0.63	
avg. ME	4.74	0.60	1.81	4.44	6.24	10.51	
avg. Monthly Vol	24.65	17.78	19.96	24.36	27.88	33.03	
persistence		0.71	0.51	0.54	0.56	0.71	

Table 4.4: **Long-short of long-short test results for size as the treatment using equal weighting.** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable firm sizes but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.16	-0.61	-0.08	0.03	0.36	0.57	1.18***
	[0.29]	[-0.87]	[-0.15]	[0.06]	[0.64]	[1.08]	[2.72]
CAPM Alpha	0.40	-0.29	0.18	0.22	0.58	0.80*	1.09**
	[0.90]	[-0.51]	[0.44]	[0.50]	[1.25]	[1.72]	[2.57]
FF-3-MOM Alpha	0.14	−0.66*	-0.06	0.03	0.30	0.59*	1.25***
	[0.60]	[-1.67]	[-0.32]	[0.13]	[1.07]	[1.76]	[3.09]
FF-5-MOM Alpha	-0.03	−1.12**	-0.19	-0.02	0.26	0.53	1.65***
	[-0.12]	[-2.48]	[-0.90]	[-0.07]	[0.86]	[1.43]	[3.75]
τ 25%	-0.05	-0.17	-0.05	-0.03	-0.01	0.01	
τ 50%	-0.02	-0.11	-0.04	-0.02	0.00	0.02	
τ 75%	0.00	-0.08	-0.04	-0.02	0.00	0.03	
avg τ	-0.04	-0.14	-0.04	-0.02	-0.01	0.02	
sd τ	0.08	0.10	0.01	0.00	0.00	0.03	
avg $\dot{Z}$	0.00	-0.07	0.03	0.04	0.01	-0.07	
avg $\dot{Z}$ in 1st quintile	-1.05	-1.01	-0.78	-0.89	-1.09	-1.24	
avg $\dot{Z}$ in 5th quintile	1.23	0.83	0.99	1.19	1.49	1.54	
avg $\dot{Z}$ P5 - P1	2.28	1.84	1.78	2.08	2.58	2.78	
number of stocks	3681.14	739.86	706.92	754.83	708.25	771.28	
avg. BM ratio	0.77	0.88	0.70	0.73	0.72	0.80	
avg. ME	4.96	17.16	4.45	1.71	0.96	0.51	
avg. Monthly Vol	25.54	52.31	29.17	18.26	16.42	11.74	
persistence		0.85	0.72	0.66	0.61	0.73	

Table 4.5: **Long-short of long-short test results for momentum as the treatment using equal weighting.** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable trailing returns but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are all equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

Feature	Avg. Relative Gain
logME	0.11
ret.2_12m	0.09
ret.1m	0.07
ret_sigma	0.06
CFP.FF	0.04
EP.FF	0.03
GrIA	0.03
txdi	0.03
turnover	0.03
OpIB	0.03
Total	0.52

Table 4.6: **Top 10 features for the $\hat{\tau}$ model trained in rolling windows when value is the treatment.** Fix any feature, we calculate its feature importance for $\hat{\tau}$ model during each rolling training window and average them to get one single number as average feature importance (Avg. Relative Gain). We rank all features by this criteria and list the top 10 in the table along with the average feature importance. We use the words "Avg." and "Relative" in the column name to highlight that it is averaged across different training windows and also the numbers are normalized contributions to loss reduction by each feature

Feature	Avg. Relative Gain
ret.1m	0.16
ret.2_12m	0.13
ret_sigma	0.11
avg_daily_dVol_million	0.08
dvc	0.05
xi	0.05
dp	0.03
BE	0.03
dlc	0.03
txdb	0.03
Total	0.69

Table 4.7: **Top 10 features for the $\hat{\tau}$ model trained in rolling windows when size is the treatment.** Fix any feature, we calculate its feature importance for $\hat{\tau}$ model during each rolling training window and average them to get one single number as average feature importance (Avg. Relative Gain). We rank all features by this criteria and list the top 10 in the table along with the average feature importance. We use the column name Avg. Relative to highlight that it is averaged across different training windows and also the numbers are normalized contributions to loss reduction by each feature

Feature	Avg. Relative Gain
00 logME	0.20
ret_sigma	0.14
ret.1m	0.12
itcb	0.05
pstkl	0.04
avg_daily_dVol_million	0.04
xi	0.03
OpIB	0.03
turnover	0.02
txp	0.02
Total	0.68

Table 4.8: **Top 10 features for the $\hat{\tau}$ model trained in rolling windows when momentum is the treatment.** Fix any feature, we calculate its feature importance for $\hat{\tau}$ model during each rolling training window and average them to get one single number as average feature importance (Avg. Relative Gain). We rank all features by this criteria and list the top 10 in the table along with the average feature importance. We use the column name Avg. Relative to highlight that it is averaged across different training windows and also the numbers are normalized contributions to loss reduction by each feature

Treatment	value	size	momentum
Avg. Return	0.82^{**}	1.55^{***}	1.19^{***}
	[2.33]	[2.62]	[2.71]
CAPM Alpha	1.00^{***}	0.60^{*}	1.11^{**}
	[3.18]	[1.76]	[2.13]
FF-3-MOM Alpha	1.02^{***}	0.67^{**}	1.24^{**}
	[3.32]	[2.01]	[2.53]
FF-5-MOM Alpha	1.19^{***}	0.65^{**}	1.42^{***}
	[3.97]	[1.99]	[3.03]

Table 4.9: **Robustness checks for long-short of long-short tests using interactions between most important features and the treatment variable.** We redo the time series regression to evaluate alphas with additional controls specified in Equation (4.20). We only report long-short of long-short results here.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.25	0.48	0.15	-0.09	0.09	0.15	-0.32
	[0.99]	[1.20]	[0.53]	[-0.35]	[0.48]	[0.56]	[-0.84]
CAPM Alpha	0.29	0.50	0.20	0.04	0.18	0.22	-0.29
	[1.13]	[1.28]	[0.67]	[0.14]	[0.87]	[0.78]	[-0.75]
FF-3-MOM Alpha	0.49***	0.73**	0.38*	0.16	0.31**	0.35	-0.38
	[3.04]	[2.30]	[1.68]	[0.83]	[2.19]	[1.45]	[-1.03]
FF-5-MOM Alpha	0.39**	0.44	0.19	0.05	0.30*	0.42*	-0.02
	[2.30]	[1.41]	[0.89]	[0.24]	[1.89]	[1.73]	[-0.05]
τ 25%	0.02	-0.07	0.02	0.04	0.05	0.07	
τ 50%	0.04	-0.03	0.02	0.04	0.06	0.08	
τ 75%	0.06	-0.01	0.03	0.05	0.06	0.10	
avg τ	0.03	-0.05	0.02	0.04	0.06	0.09	
sd τ	0.06	0.06	0.01	0.00	0.01	0.03	
avg $\dot{Z}$	-0.69	-0.61	-0.65	-0.69	-0.77	-0.71	
avg $\dot{Z}$ in 1st quintile	-1.98	-2.00	-1.89	-1.84	-1.97	-2.06	
avg $\dot{Z}$ in 5th quintile	0.38	0.58	0.37	0.25	0.22	0.38	
avg $\dot{Z}$ P5 - P1	2.36	2.58	2.25	2.09	2.18	2.44	
number of stocks	3536.90	707.85	705.98	697.96	714.12	710.99	
avg. BM ratio	0.84	1.39	0.79	0.66	0.59	0.75	
avg. ME	5.08	0.61	1.23	2.17	4.76	16.54	
avg. Monthly Vol	25.27	18.37	13.74	12.50	18.70	62.67	
persistence		0.75	0.57	0.56	0.63	0.79	

Table 4.10: **Long-short of long-short test results for value as the treatment using equal weighting (without standardizing any variables).** We apply our R-learner procedure to raw data without any standardization. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable book-to-market ratios but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.15	-0.45	0.91**	0.13	-0.06	0.02	0.47
	[0.43]	[-0.99]	[2.37]	[0.40]	[-0.19]	[0.04]	[1.02]
CAPM Alpha	0.14	-0.59	0.82**	0.15	-0.07	0.04	0.63
	[0.40]	[-1.25]	[2.14]	[0.45]	[-0.20]	[0.08]	[1.45]
FF-3-MOM Alpha	0.03	-0.66	0.73**	0.02	-0.13	-0.06	0.60
	[0.11]	[-1.46]	[2.07]	[0.07]	[-0.37]	[-0.13]	[1.30]
FF-5-MOM Alpha	-0.24	−0.97*	0.53	-0.33	-0.26	-0.29	0.68
	[-0.75]	[-1.86]	[1.44]	[-1.13]	[-0.91]	[-0.62]	[1.41]
τ 25%	-0.02	-0.06	-0.02	-0.01	-0.01	0.00	
τ 50%	-0.01	-0.04	-0.02	-0.01	-0.01	0.01	
τ 75%	0.00	-0.03	-0.02	-0.01	0.00	0.02	
avg τ	-0.01	-0.05	-0.02	-0.01	0.00	0.01	
sd τ	0.03	0.04	0.00	0.00	0.00	0.03	
avg $\dot{Z}$	6.19	5.71	6.60	6.92	6.26	5.33	
avg $\dot{Z}$ in 1st quintile	3.37	3.01	3.96	4.19	3.75	3.09	
avg $\dot{Z}$ in 5th quintile	9.17	8.56	9.32	9.59	8.86	7.70	
avg $\dot{Z}$ P5 - P1	5.79	5.55	5.37	5.40	5.11	4.61	
number of stocks	3902.94	820.86	696.05	831.87	821.85	788.46	
avg. BM ratio	0.77	0.90	0.66	0.62	0.71	0.95	
avg. ME	4.74	3.71	6.52	7.16	4.16	1.52	
avg. Monthly Vol	24.65	33.55	29.37	24.06	18.99	16.91	
persistence		0.48	0.38	0.45	0.49	0.63	

Table 4.11: **Long-short of long-short test results for size as the treatment using equal weighting (without standardizing any variables).** We apply our R-learner procedure to raw data without any standardization. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable firm sizes but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.16	-0.19	0.12	0.18	0.48	0.34	0.53
	[0.29]	[-0.28]	[0.29]	[0.40]	[1.00]	[0.67]	[1.15]
CAPM Alpha	0.40	0.11	0.26	0.39	0.69*	0.53	0.42
	[0.90]	[0.19]	[0.68]	[1.01]	[1.67]	[1.22]	[0.95]
FF-3-MOM Alpha	0.14	-0.19	0.01	0.11	0.47*	0.37	0.56
	[0.60]	[-0.47]	[0.05]	[0.52]	[1.83]	[1.04]	[1.32]
FF-5-MOM Alpha	-0.03	-0.54	0.02	-0.03	0.35	0.15	0.68
	[-0.12]	[-1.16]	[0.10]	[-0.11]	[1.27]	[0.37]	[1.44]
τ 25%	0.00	-0.01	0.00	0.00	0.01	0.01	
τ 50%	0.00	-0.01	0.00	0.00	0.01	0.01	
τ 75%	0.01	0.00	0.00	0.00	0.01	0.01	
avg τ	0.00	-0.01	0.00	0.00	0.01	0.01	
sd τ	0.01	0.02	0.00	0.00	0.00	0.01	
avg $\dot{Z}$	0.09	0.04	0.11	0.13	0.11	0.08	
avg $\dot{Z}$ in 1st quintile	-0.43	-0.50	-0.38	-0.40	-0.41	-0.41	
avg $\dot{Z}$ in 5th quintile	0.76	0.68	0.73	0.84	0.79	0.71	
avg $\dot{Z}$ P5 - P1	1.19	1.18	1.11	1.24	1.19	1.11	
number of stocks	3681.14	738.85	732.83	736.24	736.35	736.87	
avg. BM ratio	0.77	0.76	0.51	0.60	0.79	1.17	
avg. ME	4.96	11.38	6.79	3.60	2.36	0.65	
avg. Monthly Vol	25.54	44.62	33.22	23.68	17.81	8.36	
persistence		0.73	0.67	0.65	0.68	0.80	

Table 4.12: **Long-short of long-short test results for momentum as the treatment using equal weighting (without standardizing any variables).** We apply our R-learner procedure to raw data without any standardization. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable trailing returns but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

Bibliography

[1] David A. Hirshberg and Stefan Wager. Augmented minimax linear estimation. *Working paper*, 2018.

[2] Eduardo Abi Jaber and Omar El Euch. Multi-factor approximation of rough volatility models. *arXiv preprint arXiv:1801.10359*, 2018.

[3] Greg M Allenby and Peter E Rossi. Marketing models of consumer heterogeneity. *Journal of econometrics*, 89(1-2):57–78, 1998.

[4] Yakov Amihud. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56, 2002.

[5] D.W.K. Andrews and X. Shi. Inference based on conditional moment inequality models. *Econometrica*, 2013.

[6] Isaiah Andrews, Matthew Gentzkow, and Jesse M Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, pages 1553–1592, 2017.

[7] Andrew Ang, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang. The cross-section of volatility and expected returns. *The Journal of Finance*, 61(1):259–299, 2006.

[8] Kalidas Ashok and Moshe E Ben-Akiva. Alternative approaches for real-time estimation and prediction of time-dependent origin–destination flows. *Transportation Science*, 34(1):21–36, 2000.

[9] Susan Athey, Julia Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 2018.

[10] Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint*, 2017.

[11] Jaume Barcelö, Lidin Montero, Laura Marqués, and Carlos Carmona. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation research record*, 2175(1):19–27, 2010.

[12] Ole E Barndorff-Nielsen, P Reinhard Hansen, Asger Lunde, and Neil Shephard. Realized kernels in practice: Trades and quotes. *Econometrics Journal*, 12(3):C1–C32, 2009.

[13] Christian Bayer, Peter Friz, and Jim Gatheral. Pricing under rough volatility. *Quantitative Finance*, 16(6):887–904, 2016.

[14] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 2014b.

[15] Moshe Ben-Akiva, Michel Bierlaire, Didier Burton, Haris N Koutsopoulos, and Rabi Mishalani. Network state estimation and prediction for real-time traffic management. *Networks and spatial economics*, 1(3-4):293–318, 2001.

[16] Mikkel Bennedsen, Asger Lunde, and Mikko S. Pakkanen. Decoupling the short- and long-term behavior of stochastic volatility. *Working Paper*, 2016.

[17] Lorenzo Bergomi. Smile dynamics iv. 2009.

[18] Lorenzo Bergomi and Julien Guyon. Stochastic volatility's orderly smiles. *Risk*, 25(5):60, 2012.

[19] Steve Berry, Oliver B Linton, and Ariel Pakes. Limit theorems for estimating the parameters of differentiated product demand systems. *The Review of Economic Studies*, 71(3):613–654, 2004.

[20] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile Prices in Market Equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

[21] Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of political Economy*, 112(1):68–105, 2004.

[22] Steven T. Berry and Philip A. Haile. Identification in Differentiated Products Markets Using Market Level Data. *Econometrica*, 82(5):1749–1797, 2014.

[23] Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55, 2009.

[24] Nicholas Buchholz. Spatial equilibrium, search frictions and efficient regulation in the taxi industry. *Working Paper*, 2016.

[25] Mark M Carhart. On persistence in mutual fund performance. *The Journal of Finance*, 52(1):57–82, 1997.

[26] Mark M. Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52:57–82, March 1997.

[27] Peter Carr and Liuren Wu. Variance risk premiums. *The Review of Financial Studies*, 22(3):1311–1341, 2008.

[28] Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *working paper*, 2019.

[29] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 785–794, 2016.

[30] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased Machine Learning for Treatment and Structural parameters. *Econometrics Journal*, 2018.

[31] Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1), 2015.

[32] John H. Cochrane. Presidential address: Discount rates. *Journal of Finance*, 66(4), 2011.

[33] Christopher T Conlon and Julie Holland Mortimer. Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics*, 5(4):1–30, 2013.

[34] Jennifer Conrad, Robert F Dittmar, and Eric Ghysels. Ex ante skewness and expected stock returns. *The Journal of Finance*, 68(1):85–124, 2013.

[35] Kent Daniel and Sheridan Titman. Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance*, LII(1), March 1997.

[36] Paul DeMaio. Bike-sharing: History, impacts, models of provision, and future. *Journal of public transportation*, 12(4):3, 2009.

[37] Omar El Euch, Masaaki Fukasawa, Jim Gatheral, and Mathieu Rosenbaum. Short-term at-the-money asymptotics under stochastic volatility models. *arXiv preprint arXiv:1801.08675*, 2018.

[38] Omar El Euch, Masaaki Fukasawa, and Mathieu Rosenbaum. The microstructural foundations of leverage effect and rough volatility. *Finance and Stochastics*, 22(2):241–280, 2018.

[39] Paul B Ellickson and Sanjog Misra. Supermarket pricing strategies. *Marketing science*, 27(5):811–828, 2008.

[40] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[41] Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *The Journal of Financial Economics*, 116(1):1–22, 2015.

[42] Eugene F Fama and James D MacBeth. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636, 1973.

[43] Guanhao Feng, Stefano Giglioz, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *Working paper*, 2019.

[44] Martin Forde and Hongzhong Zhang. Asymptotics for rough stochastic volatility models. *SIAM Journal on Financial Mathematics*, 8(1):114–145, 2017.

[45] Guillaume R Frechette, Alessandro Lizzeri, and Tobias Salz. Frictions in a competitive, regulated market evidence from taxis. *Working Paper*, 2016.

[46] Daniel Freund, Shane G Henderson, and David B Shmoys. Minimizing multimodular functions and allocating capacity in bike-sharing systems. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 186–198. Springer, 2017.

[47] Joachim Freybergery, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. *Working paper*, Jan 2019.

[48] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[49] Masaaki Fukasawa. Asymptotic analysis for stochastic volatility: martingale expansion. *Finance and Stochastics*, 15(4):635–654, 2011.

[50] Masaaki Fukasawa. Short-time at-the-money skew and rough fractional volatility. *Quantitative Finance*, 17(2):189–198, 2017.

[51] Amit Gandhi, Zhentong Lu, and Xiaoxia Shi. Estimating Demand for Differentiated Products with Zeroes in Market Share Data. *Working Paper*, 2017.

[52] Josselin Garnier and Knut Sølna. Option pricing under fast-varying and rough stochastic volatility. *Annals of Finance*, 14(4):489–516, 2018.

[53] Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. Volatility is rough. *Quantitative Finance*, 18(6):933–949, 2018.

[54] Greater London Authority. The mayor's vision for cycling in london: an olympic legacy for all londoners. *Greater London Authority, London*, 2013.

[55] Shihao Gu, Bryan T. Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *working paper*, 2019.

[56] Jose A Guajardo, Morris A Cohen, and Serguei Netessine. Service competition and product quality in the us automobile industry. *Management Science*, 62(7):1860–1877, 2015.

[57] Campbell R. Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. *Review of Financial Studies*, 29, January 2016.

[58] Pu He, Fanyin Zheng, Elena Belavina, and Karan Girotra. Customer preference and station network in the london bike share system. *Working Paper*, 2019.

[59] Harrison Hong, Terence Lim, and Jeremy C. Stein. Bad news travels slowly: size, analyst coverage, and the profitability of momentum strategies. *Journal of Finance*, LV(1), Feb 2000.

[60] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48(1):65–91, March 1993.

[61] Diana Jorge and Gonçalo Correia. Carsharing systems demand estimation and defined operations: a literature review. *European Journal of Transport and Infrastructure Research*, 13(3), 2013.

[62] Paul Jusselin and Mathieu Rosenbaum. No-arbitrage implies power-law market impact and rough volatility. *arXiv preprint arXiv:1805.07134*, 2018.

[63] Ashish Kabra, Elena Belavina, and Karan Girotra. Bike-share systems: Accessibility and availability. *Working Paper*, 2016.

[64] Giulia Livieri, Saad Mouti, Andrea Pallavicini, and Mathieu Rosenbaum. Rough volatility: evidence from option prices. *IISE Transactions*, pages 1–21, 2018.

[65] David O. Lucca and Emanuel Moench. The pre-FOMC announcement drift. *The Journal of Finance*, LXX(1), 2015.

[66] Puneet Manchanda, Peter E Rossi, and Pradeep K Chintagunta. Response modeling with nonrandom marketing-mix variables. *Journal of Marketing Research*, 41(4):467–478, 2004.

[67] Benoit B Mandelbrot and John W Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4):422–437, 1968.

[68] Peter Midgley. Bicycle-sharing schemes: enhancing sustainable mobility in urban areas. *United Nations, Department of Economic and Social Affairs*, pages 1–12, 2011.

[69] Andrés Musalem, Marcelo Olivares, Eric T Bradlow, Christian Terwiesch, and Daniel Corsten. Structural estimation of the effect of out-of-stocks. *Management Science*, 56(7):1180–1197, 2010.

[70] Aviv Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.

[71] New York City Department of City Planning. Bike-share: Opportunities in new york city. Technical report, City of New York, 2009.

[72] Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

[73] Xinque Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *working paper*, 2019.

[74] Eoin O'Mahony and David B Shmoys. Data analysis and optimization for (citi) bike sharing. In *AAAI*, pages 687–694, 2015.

[75] Mitchell A Petersen. Estimating standard errors in finance panel data sets: comparing approaches. *The Review of Financial Studies*, 22(1):435–480, 2009.

[76] Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H. Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 2018.

[77] Thomas W Quan and Kevin R Williams. Product variety, across-market demand heterogeneity, and the value of online retail. *Working Paper*, 2017.

[78] Peter M. Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 1988.

[79] G Samorodnitsky and M Taqqu. *Non-Gaussian Stable Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, London, 1994.

[80] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *In Proceedings of the 34th International Conference on Machine Learning*, 2017.

[81] Divya Singhvi, Somya Singhvi, Peter I Frazier, Shane G Henderson, Eoin O'Mahony, David B Shmoys, and Dawn B Woodard. Predicting bike usage for new york city's bike sharing system. In *AAAI Workshop: Computational Sustainability*, 2015.

[82] James H Stock and Motohiro Yogo. Testing for weak instruments in linear iv regression. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, page 80, 2005.

[83] Che-Lin Su and Kenneth L Judd. Constrained optimization approaches to estimation of structural models. *Econometrica*, 80(5):2213–2230, 2012.

[84] Karunakaran Sudhir. Competitive pricing behavior in the auto market: A structural analysis. *Marketing Science*, 20(1):42–60, 2001.

[85] Jameson L Toole, Serdar Colak, Bradley Sturt, Lauren P Alexander, Alexandre Evsukoff, and Marta C González. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58:162–177, 2015.

[86] Transport for London. Cycling revolution london. Technical report, Mayor of London, 2010.

[87] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical*, 2018.

[88] James Woodcock, Marko Tainio, James Cheshire, Oliver O'Brien, and Anna Goodman. Health effects of the london bicycle sharing system: health impact modelling study. *Bmj*, 348:g425, 2014.

[89] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 1 edition, 2001.

[90] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[91] Yuhang Xing, Xiaoyan Zhang, and Rui Zhao. What does the individual option volatility smirk tell us about future equity returns. *Journal of Financial and Quantitative Analysis*, 45:641–662, June 2010.

Appendices

Appendix A

Appendix for Chapter 1

A.1 Calculation of Elevation Features

A standard measure for the degree of inclination in transportation is slope grade, defined as $Grade = ||\tan\alpha||$, where the angle α is as marked in Figure A.1.

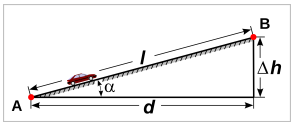

Figure A.1: Slope Grade Illustration

We can calculate *grade* for route ij with the data from Google Maps. For route ij, Google Maps allows us to retrieve the elevation at $N_{ij}+1$ roughly equal distance points (including i and j) along the route, where N_{ij} can be set by us. In other words, the route ij is divided into N_{ij} equal-distance segments. We set the maximum N_{ij} allowed by Google for each ij, and the length of a segment is around 25 meters for all ij.

Let L_{ij} be the set of segments for route ij.[1] There are N_{ij} segments in it: $|L_{ij}| = N_{ij}$. We partition the set L_{ij} into ascending segments and descending segments based on whether the elevation is increasing or decreasing along each segment. We have $L_{ij} = L_{ij}^{up} \cup L_{ij}^{down}$. $Grade_l$ can be calculated for each segment $l \in L_{ij}$ using the elevation at the starting and ending point of segment l, whose difference will give us

[1]Note both route and segment are directional.

Δh in Figure A.1, and the length of each segment $l \approx d$ in the data. We compute two elevation features for a route ij:

1) $AvgAscendGrade_{ij} := \frac{\sum_{l \in L_{ij}^{up}} Grade_l}{|L_{ij}^{up}|}$: average grade among the ascending segments along route ij.

2) $AscendPercentage_{ij} := \frac{|L_{ij}^{up}|}{|L_{ij}|}$: proportion of ascending segments on route ij.

Note that if we define similarly average descending grade and descend percentage for route ij, then by definition we have

$$AvgAscendGrade_{ij} = AvgDescendGrade_{ji},$$

and

$$DescendPercentage_{ij} = 1 - AscendPercentage_{ij}.$$

So we only include $AvgAscendGrade_{ij}$ and $AscendPercentage_{ij}$ in the analysis. The range of $AvgAscendGrade_{ij}$ in London's station network is between 0 and 0.04, which is consistent with the anecdotal evidence that London is relatively flat and suitable for biking in general.[2]

A.2 Implementation Details

We follow the model specified in Section 1.4.4 and use MPEC algorithm proposed in [83]. Our optimization problem is: [3]

$$\begin{aligned} & \min \qquad \eta' W \eta \\ s.t. \quad & \eta = \mathcal{Z}'(\delta - X^l(X^{l'}\mathcal{Z}W\mathcal{Z}'X^l)^{-1}X^{l'}\mathcal{Z}W\mathcal{Z}'\delta) \\ & q_{IJ} = \sum_{\{kl:IJ \in C_{kl}\}} w'_{kl}\alpha \frac{exp(\delta_{IJ} + X'^{nl}_{IJkl}\gamma)}{exp(X'_{0kl}\theta) + \sum_{I'J' \in C_{kl}} exp(\delta_{I'J'} + X'^{nl}_{I'J'kl}\gamma)} \quad \forall IJ \end{aligned} \tag{A.1}$$

[2] One could argue that since sharp downhill can also be challenging for biking, we should include $AvgAscendGrade_{ji}$ in the utility function for route ij. Since the maximum grade is 0.04, the descending slope should not matter.

[3] We solve the same optimization problem for morning and evening rush hour separately

The solver optimizes over $x = [\eta', \delta', \gamma', \theta', \alpha']'$. q_{IJ} is the total trips observed on route IJ in year 2014. $\mathcal{Z}$ is the instruments matrix including the exogenous covariates and the additional instruments specified in (1.10) and (1.11). W is the weight matrix. X^l are the covariates that only depend on IJ. X^{nl} are the covariates that depend on both IJ and kl. Note that in the actual implementation we do not have β from (1.3) in x since we can easily solve for them $\beta^* = (X^{l'}\mathcal{Z}W\mathcal{Z}'X^l)^{-1}X^{l'}\mathcal{Z}W\mathcal{Z}'\delta^*$ using δ^* from the final optimal solution x^*. More details are given below:

We use MPEC algorithm proposed in [83]. Our optimization problem is:

$$\begin{aligned} &\min \qquad \eta' W \eta \\ s.t. \quad &\eta = \mathcal{Z}'(\delta - X^l(X^{l'}\mathcal{Z}W\mathcal{Z}'X^l)^{-1}X^{l'}\mathcal{Z}W\mathcal{Z}'\delta) \\ &q_{IJ} = \sum_{\{kl:IJ\in C_{kl}\}} w'_{kl}\alpha \frac{exp(\delta_{IJ} + X'^{nl}_{IJkl}\gamma)}{exp(X'_{0kl}\theta) + \sum_{I'J'\in C_{kl}} exp(\delta_{I'J'} + X'^{nl}_{I'J'kl}\gamma)} \end{aligned} \tag{A.2}$$

We denote the long vector x as the vertical concatenation of all variables fed: $x = [\eta', \delta', \gamma', \theta', \alpha']'$, which is the actual vector of variables fed in to the solver. q_{IJ} are the actual observed trip count on route IJ throughout 2014. $\mathcal{Z}$ is the instruments matrix. W is the weighting matrix for different moment conditions. X^l are the covariates that only depend on IJ and correspond to linear parameters β. X^{nl} are the covariates that depend on both IJ and kl and correspond to nonlinear parameters γ. Note that in the actual implementation we do not have linear parameter β in x since we can easily solve for them $\beta* = (X^{l'}\mathcal{Z}W\mathcal{Z}'X^l)^{-1}X^{l'}\mathcal{Z}W\mathcal{Z}'\delta^*$ using δ^* from the final optimal solution x^*. Note that nonlinear parameters of interests, θ, α and γ are in x and optimized by the solver.

We denote the total number of products as N, which is equal to the number of unobserved characteristics ξ. We have $N = 9784$ for the evening rush hour and $N = 8764$ for the morning rush hour in our block model. We choose ipopt as our solver and use R package ipoptr, which is an R interface to ipopt solver. We code the core parts of the calculation-mainly the evaluation of quadratic objective function, N

nonlinear constraint and their analytic Jacobian and Hessian matrices in C++ with the help of Rcpp package blending cpp code into R. Note that a full dense Hessian matrix alone would have dimensions close to $10^4 \times 10^4$, and we use a sparse matrix implementation, which reduces the numbers we need to calculate down to around $2 * 10^6$. The same sparse implementation is applied to Jacobian matrices of objective function and all constraints. We can also see the effectiveness of our block model in reducing the computation burden and memory requirement. If we were to solve the problem using routes between stations instead of routes between station blocks, we would have a Hessian matrix 1000 times bigger than the one in the block model. We use the default convergence tolerance in ipopt. In our implementation, we manage to solve the problem using 45G memory on a standard Linux server. The algorithm normally converges in three to seven days. We use solution from a pure Logit model plus uniformly random perturbation as the starting point for x_0, and we start multiple jobs (more than 20) for different random initial values at the same time. Some jobs converge to a local infeasible point but the final x^* are all identical, which suggests that local minimum is not an issue in our estimation.

A.3 Robustness Checks for Reduced-Form Regressions

We present here the estimation results for different specifications of the reduced-form regressions. We try different breakpoints in route distance, i.e. the b in term $\max\{\log d(i,j) - \log b, 0\}$. We use 1.4km and 1.6km for b, and the results are similar to the main results where $b = 1.5km$. We also modify R and D values for our instruments in Equations (1.10) and (1.11). The reduced form results are also robust to those perturbations, which suggests that our proposed instruments are robust to the choice of b, R, and D.

A.3.1 Robustness checks for different breakpoints in route distance

We present the results for different breakpoints in route distance in Tables A.1 and A.2.

BREAKPOINT (KM)	1.4	1.4	1.5	1.5	1.6	1.6
	OLS	IV	OLS	IV	OLS	IV
O. Bike Avail.	0.05	0.66	0.05	0.69	0.05	0.76
%	(0.03)	(0.19)	(0.03)	(0.21)	(0.03)	(0.19)
D. Dock Avail.	0.50	1.75	0.50	1.47	0.50	1.35
%	(0.03)	(0.20)	(0.04)	(0.16)	(0.04)	(0.15)
Dist. 1	0.12	0.12	0.12	0.10	0.08	0.08
In log km	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Dist. 2	-0.41	-0.40	-0.37	-0.36	-0.33	-0.32
In log km	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
R^2	0.20	-	0.20	-	0.20	-
CRAGG-DONALD STATISTIC	-	83.65	-	83.64	-	83.64

Table A.1: Reduced-form regressions with different distance breakpoints for morning rush hour

BREAKPOINT (KM)	1.4	1.4	1.5	1.5	1.6	1.6
	OLS	IV	OLS	IV	OLS	IV
O. Bike Avail.	0.12	1.76	0.13	1.57	0.11	1.67
%	(0.02)	(0.15)	(0.03)	(0.17)	(0.02)	(0.15)
D. Dock Avail.	-0.45	2.28	-0.46	1.87	-0.45	1.77
%	(0.02)	(0.13)	(0.03)	(0.15)	(0.02)	(0.13)
Dist. 1	0.18	0.20	0.14	0.16	0.12	0.14
In log km	(0.02)	(0.02)	(0.03)	(0.04)	(0.02)	(0.02)
Dist. 2	-0.46	-0.39	-0.42	-0.36	-0.41	-0.31
In log km	(0.02)	(0.02)	(0.03)	(0.04)	(0.02)	(0.02)
R^2	0.33	-	0.33	-	0.33	-
CRAGG-DONALD STATISTIC	-	99.38	-	99.37	-	99.37

Table A.2: Reduced-form regressions with different distance breakpoints for evening rush hour

A.3.2 Robustness checks for different instrumental variable specifications

We present the results in Tables A.3 and A.4. Note that the OLS results also change when we modify R because Z_{ij} depend on R. We can see that in all specifications, 1) both bike and dock availability coefficients are significantly positive, and 2) IV regressions significantly increase the bike and dock availability coefficients from the OLS case, which is similar to the base case presented in the main text for $R = 600, 800, 1000m$ and $D = 7000m$.

Radius R (km)	$0.6, 0.8, 1$		$0.5, 0.7, 0.9$		$0.5, 0.7, 0.9$	
Threshold D (km)	6.5	6.5	6.5	6.5	7.0	7.0
	OLS	IV	OLS	IV	OLS	IV
O. Bike Avail.	0.01	1.08	0.05	0.71	0.05	0.87
%	(0.03)	(0.16)	(0.03)	(0.16)	(0.03)	(0.18)
D. Dock Avail.	0.49	1.60	0.50	1.28	0.50	1.24
%	(0.03)	(0.14)	(0.03)	(0.13)	(0.03)	(0.14)
Dist. 1	0.12	0.10	0.13	0.12	0.13	0.09
In log km	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Dist. 2	-0.37	-0.36	-0.38	-0.37	-0.38	-0.37
>1.5km, in log km	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
R^2	0.20	-	0.20	-	0.20	-
Cragg-Donald Statistic	-	98	-	101	-	88

Table A.3: Robustness checks for morning rush hour reduced-form regressions

Radius R (km)	$0.6, 0.8, 1$		$0.5, 0.7, 0.9$		$0.5, 0.7, 0.9$	
Threshold D (km)	6.5	6.5	6.5	6.5	7.0	7.0
	OLS	IV	OLS	IV	OLS	IV
O. Bike Avail.	0.11	1.57	0.10	1.51	0.10	1.36
%	(0.02)	(0.17)	(0.02)	(0.12)	(0.02)	(0.11)
D. Dock Avail.	-0.56	1.87	-0.52	1.68	-0.52	2.00
%	(0.02)	(0.15)	(0.02)	(0.10)	(0.02)	(0.10)
Dist. 1	0.14	0.16	0.13	0.14	0.13	0.13
In log km	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Dist. 2	-0.42	-0.36	-0.43	-0.39	-0.43	-0.41
>1.5km, in log km	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
R^2	0.33	-	0.33	-	0.33	-
Cragg-Donald Statistic	-	91	-	107	-	108

Table A.4: Robustness checks for evening rush hour reduced-form regressions

A.4 Robustness Checks for Structural Estimation

A.4.1 Structural Estimation Results Without Elevation Features

We present in Tables A.5 and A.6 the estimation results for the structural model without the two elevation features: average ascending grade and percentage of ascending segments. The results do not change much and our qualitative conclusions still hold.

A.4.2 Structural estimation results without elevation features and walking distance

We present in Tables A.7 and A.8 the estimation results for the structural model without walking distance.

Non-linear para.		Linear para.	
O. Pop. Density	0.020	Intercept	-7.658
(Ppl per hec)	(0.015)		(0.876)
D. Pop. Density	-0.040	S. Station Count	0.896
	(0.018)	(In log)	(0.055)
O. Google Plc.	0.020	E. Station Count	0.753
(Num per 4 hec)	(0.017)		(0.031)
D. Google Plc.	0.301	S. Bike Avail.	1.653
	(0.170)	(%)	(0.260)
O. Walking Dist.	-3.376	E. Dock Avail.	1.738
(1km)	(0.791)		(0.206)
D. Walking Dist.	-0.963	Route Dist. 1	1.588
	(0.495)	(In log)	(0.267)
O.D. Driving Dist.	0.255	Route Dist. 2	-2.919
(1km)	(0.053)	(>1.5km, in log)	(0.361)
O.D. Transit Dist.	0.040		
(1km)	(0.057)		
Pseudo-R^2	0.667		

Table A.5: Demand estimates: morning rush hour without elevation features

A.4.3 Structural estimation without elevation and outside option features

Next, we present structural estimation results without outside option features for customers who want to travel from O to D, in Tables A.9 and A.10.

A.5 Improving Availability for Morning Rush Hour

Non-linear para.		Linear para.	
O. Pop. Density	-0.045	Intercept	-4.218
(Ppl per hec)	(0.004)		(0.487)
D. Pop. Density	0.006	S. Station Count	0.911
	(0.007)	(In log)	(0.024)
O. Google Plc.	0.181	E. Station Count	0.913
(Num per 4 hec)	(0.024)		(0.052)
D. Google Plc.	0.035	S. Bike Avail.	0.877
	(0.011)	(%)	(0.186)
O. Walking Dist.	-2.058	E. Dock Avail.	0.831
(1km)	(0.286)		(0.282)
D. Walking Dist.	-4.589	Route Dist. 1	2.044
	(0.687)	(In log)	(0.201)
O.D. Driving Dist.	0.141	Route Dist. 2	-4.013
(1km)	(0.031)	(>1.5km, in log)	(0.268)
O.D. Transit Dist.	0.079		
(1km)	(0.033)		
Pseudo-R^2	0.791		

Table A.6: Demand estimates: evening rush hour without elevation features

We present the results in Figures A.2 and A.3 the predicted percentage system usage increase during the morning rush hour after increasing the bike and dock availability by 0.05, respectively.

A.6 Summary Statistics of Data

We have collected several data sets with very rich and diverse features related to the travel demand across the city. We show more aspects of the raw data in this section and present more summary statistics during our exploratory analysis.

A.6.1 Plot of Coverage Area

We plot the station locations in Figure A.4 on the map of the Greater London administrative area, with its 33 boroughs. It shows that the bike share system covers

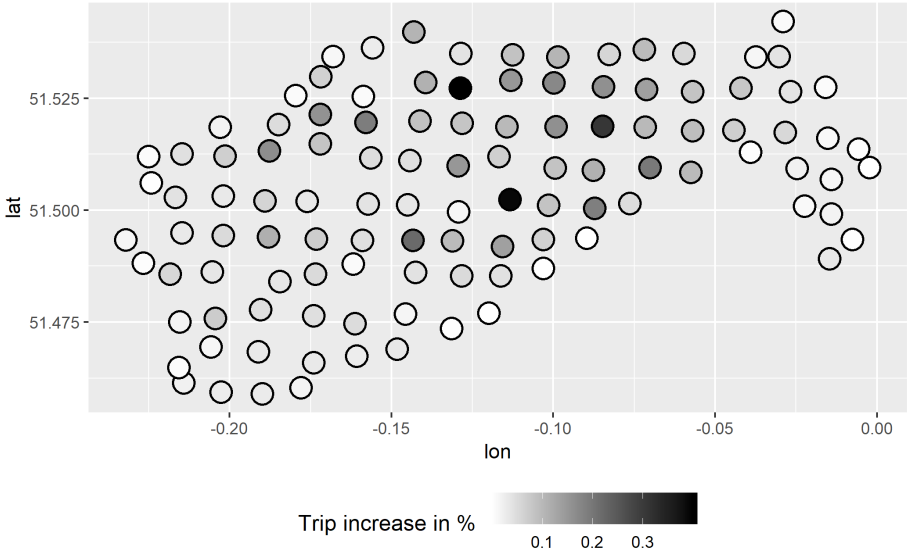

Figure A.2: Predicted morning rush hour usage increase after improving bike availability by 0.05

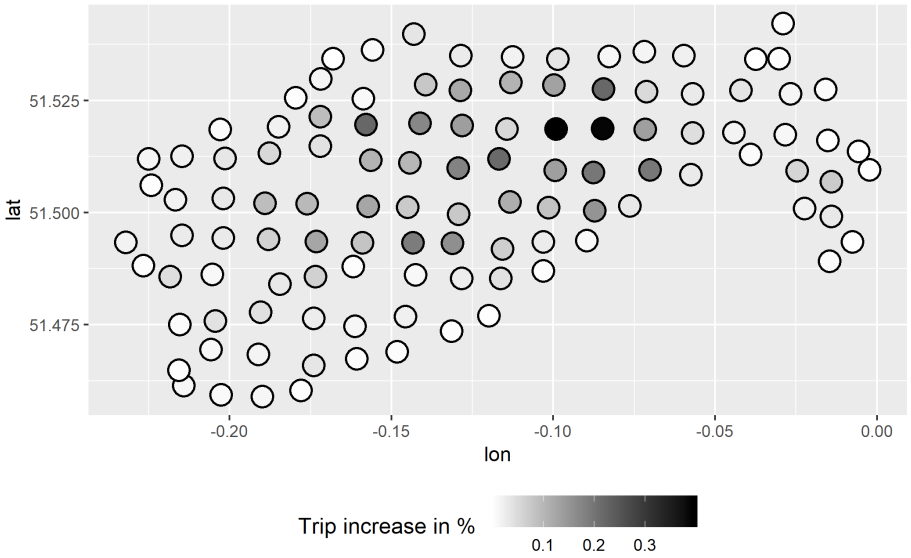

Figure A.3: Predicted morning rush hour usage increase after improving dock availability by 0.05

Nonlinear para.		Linear para.	
O. Pop. Density	0.017	Intercept	-10.524
(Ppl per hec)	(0.015)		(0.655)
D. Pop. Density	-0.040	S. Station Count	0.778
	(0.019)	(In log)	(0.025)
O. Google Plc.	0.027	E. Station Count	0.745
(Num per 4 hec)	(0.019)		(0.030)
D. Google Plc.	0.304	S. Bike Avail.	1.607
	(0.177)	(%)	(0.259)
O. Walking Dist.	NA	E. Dock Avail.	1.758
(1km)			(0.198)
D. Walking Dist.	NA	Route Dist. 1	1.303
		(In log)	(0.280)
O.D. Driving Dist.	0.165	Route Dist. 2	-2.374
(1km)	(0.076)	(>1.5km, in log)	(0.398)
O.D. Transit Dist.	0.201		
(1km)	(0.088)		
Pseudo-R^2	0.664		

Table A.7: Demand estimates: morning rush hour without elevation and walking distance

only the central boroughs of the city.

A.6.2 Google Places Data

1. Details of the Google Place Categories that we use

   Based on the Google place data scraped in Feb 2017, we have, in total, 97 types. We group them into ten categories in our analysis and list them below:

   Food This category has the following six types: food, meal delivery, meal takeaway, restaurant, cafe and bakery

   Religion This category has the following five types: church, Mosque, synagogue, place of worship, and Hindu temple

Nonlinear Para.		Linear Para.	
O. Pop. Density	-0.042	Intercept	-9.022
(Ppl per hec)	(0.004)		(0.387)
D. Pop. Density	0.006	S. Station Count	0.916
	(0.007)	(In log)	(0.024)
O. Google Plc.	0.192	E. Station Count	0.923
(Num per 4 hec)	(0.034)		(0.054)
D. Google Plc.	0.035	S. Bike Avail.	0.887
	(0.007)	(%)	(0.209)
O. Walking Dist.	NA	E. Dock Avail.	0.840
(1km)			(0.272)
D. Walking Dist.	NA	Route Dist. 1	1.964
		(In log)	(0.201)
O.D. Driving Dist.	0.121	Route Dist. 2	-3.981
(1km)	(0.035)	(>1.5km, in log)	(0.265)
O.D. Transit Dist.	0.080		
(1km)	(0.043)		
Pseudo-R^2	0.787		

Table A.8: Demand estimates: evening rush hour without elevation and walking distance

Health This category has the following eight types: dentist, doctor, health, gym, hospital, pharmacy, physiotherapist and spa

Entertainment This category has the following 14 types: amusement park, aquarium, art gallery, bar, bowling alley, movie rental, movie theatre, museum, night club, painter, park, rv park, casino and zoo

Stores This category has the following 17 types: clothing store, convenience store, department store, furniture store, hardware store, home goods store, jewelry store, liquor store, pet store, shoe store, shopping mall, store, book store, electronics store, bicycle store, grocery or supermarket, and florist

Finance This category has three types: accounting, bank and finance

Nonlinear para.		Linear para.	
O. Pop. Density	0.020	Intercept	-9.653
(Ppl per hec)	(0.084)		(0.911)
D. Pop. Density	-0.046	S. Station Count	0.789
	(0.123)	(In log)	(0.073)
O. Google Plc.	0.022	E. Station Count	0.780
(Num per 4 hec)	(0.112)		(0.053)
D. Google Plc.	0.299	S. Bike Avail.	1.583
	(0.104)	(%)	(0.163)
O. Walking Dist.	-3.128	E. Dock Avail.	1.683
(1km)	(0.416)		(0.117)
D. Walking Dist.	-0.280	Route Dist. 1	1.553
	(0.135)	(In log)	(0.361)
O.D. Driving Dist.	NA	Route Dist. 2	-2.549
(1km)		(>1.5km, in log)	(0.374)
O.D. Transit Dist.	NA		
(1km)			
Pseudo-R^2	0.667		

Table A.9: Demand estimates: morning rush hour without elevation and outside option features

Government and Offices This category has the following 12 types: city hall, embassy, courthouse, lawyer, local government office, real estate agency, travel agency, insurance agency, moving company.

We tried to include commercial types from Google places in This category, but it turns out that Google places data do not have a clean office building or commercial type. This is the best we can do and, as presented in the data section, all the groups are highly correlated with each other, suggesting that we have captured where people work in our analysis by using all of the group types, whether using them separately or just using the total place count.

Education This category has the following three types: library, school and university

Nonlinear Para.		Linear Para.	
O. Pop. Density	-0.048	Intercept	-5.316
(Ppl per hec)	(0.018)		(0.444)
D. Pop. Density	0.002	S. Station Count	1.030
	(0.024)	(In log)	(0.031)
O. Google Plc.	0.187	E. Station Count	1.320
(Num per 4 hec)	(0.051)		(0.047)
D. Google Plc.	0.039	S. Bike Avail.	0.911
	(0.047)	(%)	(0.584)
O. Walking Dist.	-1.923	E. Dock Avail.	0.935
(1km)	(0.506)		(0.385)
D. Walking Dist.	-4.909	Route Dist. 1	2.162
	(0.537)	(In log)	(0.214)
O.D. Driving Dist.	NA	Route Dist. 2	-4.774
(1km)		(>1.5km, in log)	(0.218)
O.D. Transit Dist.	NA		
(1km)			
Pseudo-R^2	0.790		

Table A.10: Demand estimates: evening rush hour without elevation and outside option features

Other transportation This category has the following four types: bus station, subway station, train station, transit station. We use this type to try to capture in the structural model the average substitution or complementary effect.

Others This category has the following 27 types: airport, ATM, beauty salon, campground, car dealer, car rental, car repair, car wash, cemetery, electrician, establishment, fire station, funeral home, gas station, general contractor, hair care, laundry, locksmith, lodging, parking, plumber, police, post office, roof contractor, stadium, storage and veterinary care.

2. Counts by categories

We lay out 200m by 200m squares uniformly on the coverage area and count

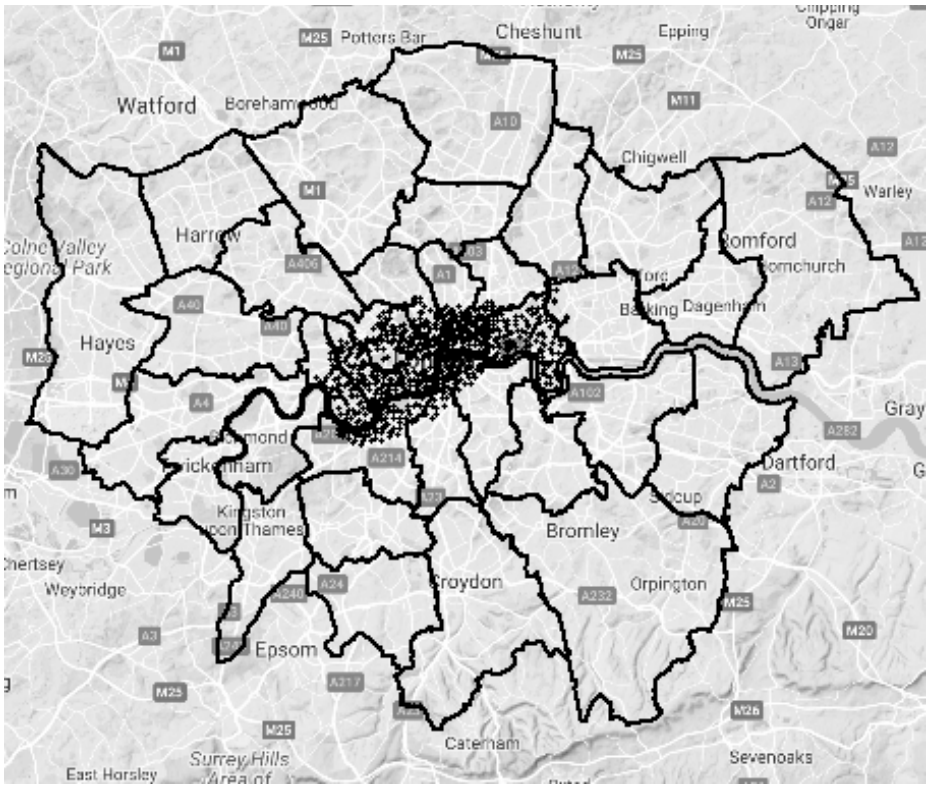

Figure A.4: Greater London and Bike Stations

how many Google Places of each category there are in each square. Table A.11 shows the summary statistics of Google place counts.

	mean	sd	min	25%	median	75%	max
Total Count	30	51	0	4	12	33	799
Food	4	8	0	0	1	5	100
Religion	0	1	0	0	0	0	8
Health	5	17	0	0	1	5	641
Entertainment	2	4	0	0	1	3	66
Store	5	13	0	0	1	5	280
Finance	3	8	0	0	0	2	108
Gov Offices	3	15	0	0	0	2	680
Education	1	1	0	0	0	1	18
Transportation	1	2	0	0	1	2	11
Others	6	9	0	1	3	8	193

Table A.11: Summary for Google Places within Uniform 200m by 200m Squares

3. Correlation of Google Places counts

	Total	Food	Reli.	Heal.	Enter.	Store	Fin.	Gov.	Edu.	Trans.	Others
Total Count	1.00	0.72	0.28	0.60	0.68	0.72	0.69	0.55	0.23	0.36	0.80
Food	0.72	1.00	0.24	0.24	0.76	0.58	0.45	0.20	0.16	0.38	0.64
Religion	0.28	0.24	1.00	0.09	0.21	0.15	0.21	0.14	0.21	0.15	0.27
Health	0.60	0.24	0.09	1.00	0.21	0.22	0.28	0.13	0.13	0.12	0.34
Entertainment	0.68	0.76	0.21	0.21	1.00	0.55	0.53	0.21	0.14	0.29	0.58
Store	0.72	0.58	0.15	0.22	0.55	1.00	0.46	0.19	0.12	0.25	0.57
Finance	0.69	0.45	0.21	0.28	0.53	0.46	1.00	0.27	0.11	0.26	0.63
Gov Offices	0.55	0.20	0.14	0.13	0.21	0.19	0.27	1.00	0.08	0.12	0.29
Education	0.23	0.16	0.21	0.13	0.14	0.12	0.11	0.08	1.00	0.11	0.23
Transportation	0.36	0.38	0.15	0.12	0.29	0.25	0.26	0.12	0.11	1.00	0.32
Others	0.80	0.64	0.27	0.34	0.58	0.57	0.63	0.29	0.23	0.32	1.00

Table A.12: Correlation Matrix for Google Places including Total Place Counts

Table A.12 shows how correlated the counts are across different categories. We can see that in general the correlation is pretty high.

A.6.3 Census Data

We present in Tables A.13 and A.14 the summary statistics of population density for both the 430 covered LSOAs and all 4835 LSOAs from the Census data.

We can see from the Area rows that the LSOAs in the coverage areas are generally smaller than the average LSOA. In terms of population density, the LSOAs in the coverage area are denser than the average LSOA. But comparing the Max columns, we see that, currently, the station network does miss some extremely dense LSOAs, which implie that there would likely be substantial demand increase if the system were expanded.

A.6.4 Route-level Usage

Route level usage and distance statistics are presented in Table A.15 and Table A.16.

Covered LSOAs	units	Mean	sd	Min	25%	median	75%	Max
all ages	num	1702	330	985	1480	1664	1892	3081
working ages	num	1294	289	608	1083	1250	1482	2400
Area	Hectare(10^4m^2)	17	19	4	9	12	18	192
All Ages Den	num/10^4m^2	143	67	6	98	139	182	399
Working Age Den	num/10^4m^2	108	50	5	74	105	139	295

Table A.13: Census Data Summary For 430 Covered LSOAs

All LSOAs	units	Mean	sd	Min	25%	median	75%	Max
all ages	num	1691	264	985	1530	1654	1817	4933
working ages	num	1167	229	608	1010	1128	1281	4235
Area	Hectare(10^4m^2)	33	63	2	13	20	32	1580
All Ages Den	num/10^4m^2	96	61	1	52	83	128	685
Working Age Den	num/10^4m^2	68	46	1	34	57	92	440

Table A.14: Census Data Summary For All 4835 LSOAs

All Routes	Mean	sd	Min	25%	median	75%	Max
Route Trip Counts	17	61	0	0	2	12	6707
Route Distance (m)	5403	3052	27	3057	4934	7274	16338

Table A.15: Route Level Usage and Distance Distribution For All Routes

	Mean	sd	Min	25%	median	75%	Max
Trip counts by route	27	76	1	2	7	25	6707
Route distance (m)	5403	3052	27	2341	3620	5130	15998

Table A.16: Route Level Usage and Distance Distribution for Routes with Positive Usage

Appendix B

Appendix for Chapter 2

B.1 Additional Simulation Results

Similar to the simulation results shown in Chapter 2, the results of this section is preliminary. We show more thorough simulation results in this section under various settings. We vary the true intercept β_0 from -10, -12 to -14 and the true standard deviation of ξ, σ_ξ from 0.5, 1.0 to 1.5. The most difficult estimation problem is the case where $\beta_0 = -14$ and $\sigma_\xi = 1.5$. We can see that in that setting, although no method is perfect but our two-stage weighting estimator outperforms other estimators overall. We didn't vary the number of markets T and number of products per market. They are kept at $T = 25$ and $J_t = 50, t = 1, 2, ..., T$. First stage setup is exactly the same as specified in Chapter 2.

Avg % zeros =26.6%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
β_0	-10	-9.276	-9.302	-9.884	-10.081
		(0.010)	(0.009)	(0.032)	(0.008)
$\bar{\beta}_1$	1	0.830	0.890	0.982	0.982
		(0.003)	(0.002)	(0.007)	(0.003)
$\bar{\beta}_2$	1	0.895	0.885	0.967	0.988
		(0.003)	(0.002)	(0.010)	(0.003)
$\bar{\beta}_3$	1	0.900	0.891	0.964	0.989
		(0.003)	(0.002)	(0.011)	(0.003)
$\bar{\beta}_4$	1	0.896	0.884	0.957	0.996
		(0.003)	(0.002)	(0.011)	(0.003)
$\bar{\beta}_5$	1	0.896	0.884	0.981	0.994
		(0.003)	(0.002)	(0.011)	(0.003)
λ_1	0.5	0.670	0.451	0.482	0.530
		(0.005)	(0.002)	(0.009)	(0.004)
λ_2	0.5	0.492	0.446	0.459	0.496
		(0.005)	(0.005)	(0.017)	(0.004)
λ_3	0.5	0.492	0.454	0.446	0.505
		(0.005)	(0.005)	(0.017)	(0.004)
λ_4	0.5	0.491	0.458	0.460	0.487
		(0.004)	(0.004)	(0.016)	(0.004)
λ_5	0.5	0.486	0.449	0.479	0.492
		(0.005)	(0.005)	(0.019)	(0.004)

Table B.1: Simulation Results from 1000 Runs for $\beta_0 = -10$ and $\sigma_\xi = 0.5$. The utility function of customers are specified by $U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$ where the slope is given by β_0 and does not have random coefficient part. We have five product features. All 5 β_{im}'s are distributed as $\mathcal{N}(1, 0.5)$, i.e., true values of $\bar{\beta}_i = 1, i = 1, 2, ..., 5$ and true values of standard deviation $\lambda_i = 0.5, i = 1, 2, ..., 5$. We have $T = 25$ markets and $J_t = 50$ products per market. ξ is generated by normal distribution with mean 0 and standard deviation σ_ξ). The parameters we need to estimate and care about is β_0 and $\bar{\beta}_i, \lambda_i, i = 1, 2, ..., 5$. We apply different estimators and compare their results against the true values that generated the simulated data. We simulate 1000 data sets and average the parameter estimates for all estimators across 1000 runs. Standard errors in the parenthesis are calculated from the 1000 simulation runs. From left to right, we show true values, estimates of BLP, bound estimator from [51], our two-stage bound estimator and two-stage weighting estimator, respectively. We also show the average percentage of 0-sale products from 1000 runs in the first cell of the table to show how severe the long-tail pattern is in the simulation setting. For the simulated standard normal variables used in the estimation of our random coefficient model, we directly use the true values that generated the data since simulation error is not the focus of this study.

Avg % zeros =30.0%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
β_0	-10	-8.921	-9.063	-9.628	-10.188
		(0.014)	(0.011)	(0.043)	(0.015)
$\bar{\beta}_1$	1	0.778	0.859	0.939	0.926
		(0.003)	(0.002)	(0.010)	(0.005)
$\bar{\beta}_2$	1	0.841	0.843	0.936	0.971
		(0.005)	(0.004)	(0.017)	(0.005)
$\bar{\beta}_3$	1	0.845	0.842	0.935	0.969
		(0.004)	(0.004)	(0.017)	(0.005)
$\bar{\beta}_4$	1	0.839	0.840	0.898	0.977
		(0.004)	(0.004)	(0.018)	(0.005)
$\bar{\beta}_5$	1	0.846	0.851	0.951	0.990
		(0.004)	(0.004)	(0.016)	(0.005)
λ_1	0.5	0.666	0.399	0.419	0.570
		(0.008)	(0.005)	(0.012)	(0.007)
λ_2	0.5	0.471	0.413	0.400	0.490
		(0.007)	(0.007)	(0.023)	(0.007)
λ_3	0.5	0.490	0.407	0.411	0.509
		(0.007)	(0.009)	(0.026)	(0.006)
λ_4	0.5	0.476	0.416	0.386	0.480
		(0.008)	(0.008)	(0.026)	(0.007)
λ_5	0.5	0.475	0.421	0.437	0.490
		(0.007)	(0.007)	(0.025)	(0.007)

Table B.2: Simulation Results from 1000 Runs for $\beta_0 = -10$ and $\sigma_\xi = 1.0$. The utility function of customers are specified by $U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$ where the slope is given by β_0 and does not have random coefficient part. We have five product features. All 5 β_{im}'s are distributed as $\mathcal{N}(1, 0.5)$, i.e., true values of $\bar{\beta}_i = 1, i = 1, 2, ..., 5$ and true values of standard deviation $\lambda_i = 0.5, i = 1, 2, ..., 5$. We have $T = 25$ markets and $J_t = 50$ products per market. ξ is generated by normal distribution with mean 0 and standard deviation σ_ξ). The parameters we need to estimate and care about is β_0 and $\bar{\beta}_i, \lambda_i, i = 1, 2, ..., 5$. We apply different estimators and compare their results against the true values that generated the simulated data. We simulate 1000 data sets and average the parameter estimates for all estimators across 1000 runs. Standard errors in the parenthesis are calculated from the 1000 simulation runs. From left to right, we show true values, estimates of BLP, bound estimator from [51], our two-stage bound estimator and two-stage weighting estimator, respectively. We also show the average percentage of 0-sale products from 1000 runs in the first cell of the table to show how severe the long-tail pattern is in the simulation setting. For the simulated standard normal variables used in the estimation of our random coefficient model, we directly use the true values that generated the data since simulation error is not the focus of this study.

Avg % zeros =32.8%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
β_0	-10	-8.447	-8.712	-9.105	-10.390
		(0.019)	(0.015)	(0.038)	(0.020)
$\bar{\beta}_1$	1	0.722	0.807	0.856	0.859
		(0.004)	(0.003)	(0.009)	(0.009)
$\bar{\beta}_2$	1	0.773	0.784	0.797	0.952
		(0.006)	(0.005)	(0.022)	(0.008)
$\bar{\beta}_3$	1	0.773	0.781	0.822	0.947
		(0.006)	(0.006)	(0.023)	(0.007)
$\bar{\beta}_4$	1	0.764	0.775	0.834	0.953
		(0.006)	(0.006)	(0.020)	(0.007)
$\bar{\beta}_5$	1	0.777	0.791	0.838	0.973
		(0.007)	(0.006)	(0.020)	(0.007)
λ_1	0.5	0.633	0.341	0.338	0.594
		(0.010)	(0.006)	(0.012)	(0.012)
λ_2	0.5	0.450	0.351	0.335	0.493
		(0.011)	(0.012)	(0.027)	(0.012)
λ_3	0.5	0.471	0.347	0.329	0.509
		(0.010)	(0.013)	(0.029)	(0.010)
λ_4	0.5	0.455	0.353	0.333	0.473
		(0.011)	(0.012)	(0.029)	(0.010)
λ_5	0.5	0.462	0.362	0.366	0.494
		(0.010)	(0.010)	(0.028)	(0.010)

Table B.3: Simulation Results from 1000 Runs for $\beta_0 = -10$ and $\sigma_\xi = 1.5$. The utility function of customers are specified by $U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$ where the slope is given by β_0 and does not have random coefficient part. We have five product features. All 5 β_{im}'s are distributed as $\mathcal{N}(1, 0.5)$, i.e., true values of $\bar{\beta}_i = 1, i = 1, 2, ..., 5$ and true values of standard deviation $\lambda_i = 0.5, i = 1, 2, ..., 5$. We have $T = 25$ markets and $J_t = 50$ products per market. ξ is generated by normal distribution with mean 0 and standard deviation σ_ξ). The parameters we need to estimate and care about is β_0 and $\bar{\beta}_i, \lambda_i, i = 1, 2, ..., 5$. We apply different estimators and compare their results against the true values that generated the simulated data. We simulate 1000 data sets and average the parameter estimates for all estimators across 1000 runs. Standard errors in the parenthesis are calculated from the 1000 simulation runs. From left to right, we show true values, estimates of BLP, bound estimator from [51], our two-stage bound estimator and two-stage weighting estimator, respectively. We also show the average percentage of 0-sale products from 1000 runs in the first cell of the table to show how severe the long-tail pattern is in the simulation setting. For the simulated standard normal variables used in the estimation of our random coefficient model, we directly use the true values that generated the data since simulation error is not the focus of this study.

Avg % zeros =41.8%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
β_0	-12	-10.538	-10.848	-11.600	-11.980
		(0.019)	(0.012)	(0.041)	(0.012)
$\bar{\beta}_1$	1	0.522	0.856	0.954	0.933
		(0.009)	(0.003)	(0.011)	(0.006)
$\bar{\beta}_2$	1	0.858	0.847	0.938	0.980
		(0.003)	(0.003)	(0.012)	(0.003)
$\bar{\beta}_3$	1	0.865	0.849	0.947	0.980
		(0.003)	(0.003)	(0.011)	(0.003)
$\bar{\beta}_4$	1	0.857	0.844	0.925	0.996
		(0.004)	(0.003)	(0.011)	(0.003)
$\bar{\beta}_5$	1	0.861	0.849	0.941	1.000
		(0.004)	(0.003)	(0.011)	(0.003)
λ_1	0.5	0.791	0.430	0.463	0.550
		(0.010)	(0.003)	(0.008)	(0.006)
λ_2	0.5	0.472	0.433	0.417	0.494
		(0.007)	(0.006)	(0.019)	(0.005)
λ_3	0.5	0.468	0.433	0.413	0.499
		(0.007)	(0.005)	(0.020)	(0.005)
λ_4	0.5	0.467	0.439	0.435	0.467
		(0.007)	(0.005)	(0.021)	(0.005)
λ_5	0.5	0.474	0.433	0.428	0.484
		(0.006)	(0.006)	(0.020)	(0.004)

Table B.4: Simulation Results from 1000 Runs for $\beta_0 = -12$ and $\sigma_\xi = 0.5$. The utility function of customers are specified by $U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$ where the slope is given by β_0 and does not have random coefficient part. We have five product features. All 5 β_{im}'s are distributed as $\mathcal{N}(1, 0.5)$, i.e., true values of $\bar{\beta}_i = 1, i = 1, 2, ..., 5$ and true values of standard deviation $\lambda_i = 0.5, i = 1, 2, ..., 5$. We have $T = 25$ markets and $J_t = 50$ products per market. ξ is generated by normal distribution with mean 0 and standard deviation σ_ξ). The parameters we need to estimate and care about is β_0 and $\bar{\beta}_i, \lambda_i, i = 1, 2, ..., 5$. We apply different estimators and compare their results against the true values that generated the simulated data. We simulate 1000 data sets and average the parameter estimates for all estimators across 1000 runs. Standard errors in the parenthesis are calculated from the 1000 simulation runs. From left to right, we show true values, estimates of BLP, bound estimator from [51], our two-stage bound estimator and two-stage weighting estimator, respectively. We also show the average percentage of 0-sale products from 1000 runs in the first cell of the table to show how severe the long-tail pattern is in the simulation setting. For the simulated standard normal variables used in the estimation of our random coefficient model, we directly use the true values that generated the data since simulation error is not the focus of this study.

Avg % zeros =42.8%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
β_0	-12	-10.128	-10.561	-11.283	-12.019
		(0.023)	(0.016)	(0.045)	(0.017)
$\bar{\beta}_1$	1	0.495	0.841	0.949	0.881
		(0.011)	(0.005)	(0.012)	(0.009)
$\bar{\beta}_2$	1	0.799	0.805	0.885	0.963
		(0.005)	(0.004)	(0.017)	(0.005)
$\bar{\beta}_3$	1	0.804	0.802	0.875	0.956
		(0.005)	(0.005)	(0.017)	(0.005)
$\bar{\beta}_4$	1	0.795	0.797	0.851	0.976
		(0.005)	(0.005)	(0.013)	(0.005)
$\bar{\beta}_5$	1	0.805	0.813	0.866	0.993
		(0.005)	(0.004)	(0.016)	(0.004)
λ_1	0.5	0.754	0.377	0.377	0.561
		(0.012)	(0.005)	(0.012)	(0.009)
λ_2	0.5	0.455	0.406	0.383	0.497
		(0.010)	(0.008)	(0.024)	(0.007)
λ_3	0.5	0.466	0.391	0.374	0.509
		(0.007)	(0.010)	(0.024)	(0.006)
λ_4	0.5	0.450	0.401	0.391	0.462
		(0.011)	(0.008)	(0.023)	(0.008)
λ_5	0.5	0.442	0.404	0.425	0.478
		(0.009)	(0.007)	(0.025)	(0.007)

Table B.5: Simulation Results from 1000 Runs for $\beta_0 = -12$ and $\sigma_\xi = 1.0$. The utility function of customers are specified by $U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$ where the slope is given by β_0 and does not have random coefficient part. We have five product features. All 5 β_{im}'s are distributed as $\mathcal{N}(1, 0.5)$, i.e., true values of $\bar{\beta}_i = 1, i = 1, 2, ..., 5$ and true values of standard deviation $\lambda_i = 0.5, i = 1, 2, ..., 5$. We have $T = 25$ markets and $J_t = 50$ products per market. ξ is generated by normal distribution with mean 0 and standard deviation σ_ξ). The parameters we need to estimate and care about is β_0 and $\bar{\beta}_i, \lambda_i, i = 1, 2, ..., 5$. We apply different estimators and compare their results against the true values that generated the simulated data. We simulate 1000 data sets and average the parameter estimates for all estimators across 1000 runs. Standard errors in the parenthesis are calculated from the 1000 simulation runs. From left to right, we show true values, estimates of BLP, bound estimator from [51], our two-stage bound estimator and two-stage weighting estimator, respectively. We also show the average percentage of 0-sale products from 1000 runs in the first cell of the table to show how severe the long-tail pattern is in the simulation setting. For the simulated standard normal variables used in the estimation of our random coefficient model, we directly use the true values that generated the data since simulation error is not the focus of this study.

Avg % zeros =44.6%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
β_0	-12	-9.567	-10.161	-10.627	-12.132
		(0.026)	(0.018)	(0.042)	(0.025)
$\bar{\beta}_1$	1	0.463	0.807	0.880	0.810
		(0.011)	(0.005)	(0.011)	(0.013)
$\bar{\beta}_2$	1	0.728	0.745	0.780	0.950
		(0.007)	(0.005)	(0.020)	(0.008)
$\bar{\beta}_3$	1	0.724	0.743	0.785	0.930
		(0.006)	(0.005)	(0.020)	(0.006)
$\bar{\beta}_4$	1	0.714	0.735	0.767	0.963
		(0.007)	(0.006)	(0.022)	(0.007)
$\bar{\beta}_5$	1	0.731	0.758	0.781	0.986
		(0.007)	(0.006)	(0.020)	(0.007)
λ_1	0.5	0.710	0.312	0.276	0.594
		(0.014)	(0.006)	(0.013)	(0.012)
λ_2	0.5	0.425	0.347	0.313	0.493
		(0.014)	(0.012)	(0.029)	(0.010)
λ_3	0.5	0.437	0.333	0.347	0.509
		(0.012)	(0.014)	(0.027)	(0.009)
λ_4	0.5	0.426	0.348	0.341	0.446
		(0.014)	(0.013)	(0.026)	(0.014)
λ_5	0.5	0.415	0.352	0.290	0.472
		(0.011)	(0.012)	(0.026)	(0.009)

Table B.6: Simulation Results from 1000 Runs for $\beta_0 = -12$ and $\sigma_\xi = 1.5$. The utility function of customers are specified by $U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$ where the slope is given by β_0 and does not have random coefficient part. We have five product features. All 5 β_{im}'s are distributed as $\mathcal{N}(1, 0.5)$, i.e., true values of $\bar{\beta}_i = 1, i = 1, 2, ..., 5$ and true values of standard deviation $\lambda_i = 0.5, i = 1, 2, ..., 5$. We have $T = 25$ markets and $J_t = 50$ products per market. ξ is generated by normal distribution with mean 0 and standard deviation σ_ξ). The parameters we need to estimate and care about is β_0 and $\bar{\beta}_i, \lambda_i, i = 1, 2, ..., 5$. We apply different estimators and compare their results against the true values that generated the simulated data. We simulate 1000 data sets and average the parameter estimates for all estimators across 1000 runs. Standard errors in the parenthesis are calculated from the 1000 simulation runs. From left to right, we show true values, estimates of BLP, bound estimator from [51], our two-stage bound estimator and two-stage weighting estimator, respectively. We also show the average percentage of 0-sale products from 1000 runs in the first cell of the table to show how severe the long-tail pattern is in the simulation setting. For the simulated standard normal variables used in the estimation of our random coefficient model, we directly use the true values that generated the data since simulation error is not the focus of this study.

Avg % zeros =59.1%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
β_0	-14	-11.618	-12.154	-13.244	-13.925
		(0.035)	(0.021)	(0.047)	(0.019)
$\bar{\beta}_1$	1	0.158	0.820	0.951	0.867
		(0.027)	(0.007)	(0.013)	(0.014)
$\bar{\beta}_2$	1	0.828	0.793	0.899	0.989
		(0.006)	(0.004)	(0.011)	(0.004)
$\bar{\beta}_3$	1	0.826	0.792	0.904	0.980
		(0.005)	(0.004)	(0.011)	(0.004)
$\bar{\beta}_4$	1	0.825	0.787	0.892	1.025
		(0.005)	(0.003)	(0.012)	(0.004)
$\bar{\beta}_5$	1	0.826	0.791	0.904	1.031
		(0.005)	(0.004)	(0.012)	(0.004)
λ_1	0.5	0.829	0.397	0.430	0.574
		(0.017)	(0.004)	(0.007)	(0.009)
λ_2	0.5	0.474	0.413	0.427	0.490
		(0.009)	(0.007)	(0.021)	(0.005)
λ_3	0.5	0.472	0.427	0.426	0.502
		(0.010)	(0.007)	(0.017)	(0.006)
λ_4	0.5	0.481	0.421	0.388	0.458
		(0.009)	(0.007)	(0.021)	(0.006)
λ_5	0.5	0.466	0.417	0.425	0.463
		(0.009)	(0.007)	(0.017)	(0.006)

Table B.7: Simulation Results from 1000 Runs for $\beta_0 = -14$ and $\sigma_\xi = 0.5$. The utility function of customers are specified by $U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$ where the slope is given by β_0 and does not have random coefficient part. We have five product features. All 5 β_{im}'s are distributed as $\mathcal{N}(1, 0.5)$, i.e., true values of $\bar{\beta}_i = 1, i = 1, 2, ..., 5$ and true values of standard deviation $\lambda_i = 0.5, i = 1, 2, ..., 5$. We have $T = 25$ markets and $J_t = 50$ products per market. ξ is generated by normal distribution with mean 0 and standard deviation σ_ξ). The parameters we need to estimate and care about is β_0 and $\bar{\beta}_i, \lambda_i, i = 1, 2, ..., 5$. We apply different estimators and compare their results against the true values that generated the simulated data. We simulate 1000 data sets and average the parameter estimates for all estimators across 1000 runs. Standard errors in the parenthesis are calculated from the 1000 simulation runs. From left to right, we show true values, estimates of BLP, bound estimator from [51], our two-stage bound estimator and two-stage weighting estimator, respectively. We also show the average percentage of 0-sale products from 1000 runs in the first cell of the table to show how severe the long-tail pattern is in the simulation setting. For the simulated standard normal variables used in the estimation of our random coefficient model, we directly use the true values that generated the data since simulation error is not the focus of this study.

Avg % zeros =59.4%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
β_0	-14	-11.044	-11.795	-12.750	-13.902
		(0.038)	(0.024)	(0.067)	(0.026)
$\bar{\beta}_1$	1	0.119	0.814	0.951	0.830
		(0.026)	(0.008)	(0.018)	(0.018)
$\bar{\beta}_2$	1	0.759	0.744	0.809	0.961
		(0.007)	(0.006)	(0.014)	(0.005)
$\bar{\beta}_3$	1	0.763	0.748	0.820	0.956
		(0.007)	(0.005)	(0.015)	(0.006)
$\bar{\beta}_4$	1	0.755	0.739	0.801	1.000
		(0.007)	(0.006)	(0.016)	(0.006)
$\bar{\beta}_5$	1	0.762	0.754	0.829	1.031
		(0.007)	(0.006)	(0.020)	(0.006)
λ_1	0.5	0.804	0.346	0.345	0.570
		(0.017)	(0.005)	(0.013)	(0.011)
λ_2	0.5	0.444	0.387	0.382	0.494
		(0.012)	(0.011)	(0.025)	(0.008)
λ_3	0.5	0.450	0.369	0.358	0.509
		(0.012)	(0.012)	(0.024)	(0.008)
λ_4	0.5	0.446	0.383	0.410	0.442
		(0.013)	(0.011)	(0.024)	(0.010)
λ_5	0.5	0.431	0.392	0.350	0.456
		(0.010)	(0.009)	(0.024)	(0.008)

Table B.8: Simulation Results from 1000 Runs for $\beta_0 = -14$ and $\sigma_\xi = 1.0$. The utility function of customers are specified by $U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$ where the slope is given by β_0 and does not have random coefficient part. We have five product features. All 5 β_{im}'s are distributed as $\mathcal{N}(1, 0.5)$, i.e., true values of $\bar{\beta}_i = 1, i = 1, 2, ..., 5$ and true values of standard deviation $\lambda_i = 0.5, i = 1, 2, ..., 5$. We have $T = 25$ markets and $J_t = 50$ products per market. ξ is generated by normal distribution with mean 0 and standard deviation σ_ξ). The parameters we need to estimate and care about is β_0 and $\bar{\beta}_i, \lambda_i, i = 1, 2, ..., 5$. We apply different estimators and compare their results against the true values that generated the simulated data. We simulate 1000 data sets and average the parameter estimates for all estimators across 1000 runs. Standard errors in the parenthesis are calculated from the 1000 simulation runs. From left to right, we show true values, estimates of BLP, bound estimator from [51], our two-stage bound estimator and two-stage weighting estimator, respectively. We also show the average percentage of 0-sale products from 1000 runs in the first cell of the table to show how severe the long-tail pattern is in the simulation setting. For the simulated standard normal variables used in the estimation of our random coefficient model, we directly use the true values that generated the data since simulation error is not the focus of this study.

Avg % zeros =61.2%	True	BLP	Gandhi et al. (2017)	Two-stage Bound Estimator	Two-stage Weighting Estimator
β_0	-14	-10.437	-11.325	-11.912	-13.979
		(0.039)	(0.029)	(0.060)	(0.034)
$\bar{\beta}_1$	1	0.150	0.790	0.892	0.812
		(0.026)	(0.009)	(0.015)	(0.025)
$\bar{\beta}_2$	1	0.681	0.684	0.726	0.946
		(0.008)	(0.007)	(0.017)	(0.008)
$\bar{\beta}_3$	1	0.685	0.688	0.690	0.938
		(0.007)	(0.006)	(0.018)	(0.009)
$\bar{\beta}_4$	1	0.670	0.677	0.714	0.984
		(0.007)	(0.007)	(0.021)	(0.008)
$\bar{\beta}_5$	1	0.686	0.692	0.713	1.022
		(0.008)	(0.007)	(0.021)	(0.007)
λ_1	0.5	0.730	0.281	0.237	0.564
		(0.019)	(0.007)	(0.013)	(0.016)
λ_2	0.5	0.402	0.344	0.376	0.475
		(0.017)	(0.014)	(0.027)	(0.013)
λ_3	0.5	0.420	0.329	0.330	0.512
		(0.016)	(0.015)	(0.025)	(0.011)
λ_4	0.5	0.397	0.333	0.340	0.413
		(0.016)	(0.013)	(0.028)	(0.014)
λ_5	0.5	0.414	0.353	0.330	0.464
		(0.014)	(0.013)	(0.025)	(0.012)

Table B.9: Simulation Results from 1000 Runs for $\beta_0 = -14$ and $\sigma_\xi = 1.5$. The utility function of customers are specified by $U_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_{im} X_{jt}^{(m)} + \xi_{jt} + \varepsilon_{ijt}$ where the slope is given by β_0 and does not have random coefficient part. We have five product features. All 5 β_{im}'s are distributed as $\mathcal{N}(1, 0.5)$, i.e., true values of $\bar{\beta}_i = 1, i = 1, 2, ..., 5$ and true values of standard deviation $\lambda_i = 0.5, i = 1, 2, ..., 5$. We have $T = 25$ markets and $J_t = 50$ products per market. ξ is generated by normal distribution with mean 0 and standard deviation σ_ξ). The parameters we need to estimate and care about is β_0 and $\bar{\beta}_i, \lambda_i, i = 1, 2, ..., 5$. We apply different estimators and compare their results against the true values that generated the simulated data. We simulate 1000 data sets and average the parameter estimates for all estimators across 1000 runs. Standard errors in the parenthesis are calculated from the 1000 simulation runs. From left to right, we show true values, estimates of BLP, bound estimator from [51], our two-stage bound estimator and two-stage weighting estimator, respectively. We also show the average percentage of 0-sale products from 1000 runs in the first cell of the table to show how severe the long-tail pattern is in the simulation setting. For the simulated standard normal variables used in the estimation of our random coefficient model, we directly use the true values that generated the data since simulation error is not the focus of this study.

Appendix C

Appendix for Chapter 3

C.1 Filtering of Option Data

We apply some filtering rules when computing implied roughness to avoid using questionable data from illiquid options. We largely follow the rules in [91], which are quite standard in the empirical literature on options. We require the following features:

- Underlying stock volume for that day > 0;
- Underlying stock price for that day $> \$5$;
- Implied volatility of the option $\geq 3\%$ and $\leq 200\%$;
- The option's open interest > 0;
- The option's volume can be 0 but has to be non-missing;
- The option has time to maturity $\tau \geq 5$ and $\tau \leq 365$ calendar days.

In addition, when estimating non-parametrically the ATM skew for each time-to-maturity τ, we set the minimal number of implied volatilities needed to measure the ATM skew for a particular time-to-maturity (for a particular stock on a particular day) at four.

When running a regression of the ATM skew term structure to estimate an implied H, we use the following filtering rules: The minimal number of ATM skews along the dimension of time-to-maturity (for a particular stock on a particular day) is three,

meaning that there must be at least three points in the regression

$$\log \phi(\tau) = c + (H - 1/2) \log \tau + \epsilon.$$

For each day, we apply these filters to call and put options separately. If for a stock, both calls and puts pass the filtering rules on a given day, we use the average implied roughness $(H^{call} + H^{put})/2$ as the implied measure on that day; otherwise we use whichever type of option passes, and if neither passes the filters, we mark the value as NA for that stock on that day.

When forming monthly portfolios, we need to aggregate daily implied roughness measures into a monthly measure. We include a stock only if it has more than 15 non-NA daily implied roughness estimates for that month. (This is similar to what is used by [7].) Otherwise, we mark the implied measure for that stock and month as NA. These restrictions define our implied universe of stock-month pairs.

In estimating daily realized variance $\hat{\sigma}_d^2$ in Section 3.2.1, we use trade data only and we apply the data cleaning steps in [12].

Appendix D

Appendix for Chapter 4

D.1 Feature List

We list all features used in Table D.1. For most features involving balance sheet items from Compustat such as BM.FF and EP.FF in Table D.1, we exactly follow the convention in [40]: we require certain delay when aligning Compustat data and CRSP data to make sure that the 10-K data are available to the investors when we feed them into R-learner procedure. This is more conservative and a relaxation should only improve our results.

In the residualization step specified in Equation (4.15), we control the potential confounders by subtracting $m(x)$ and $e(x)$ from response and treatment. Predicting returns are very hard. $m(x)$ model usually cannot do a good job and thus most of the variations in response, or cross-sectionally standardized next month returns, are left in the residuals. However, it is different for the $e(x)$ model. If it happens that our controls can predict the treatment too well so that $\dot{Z} - \hat{e}(\dot{X})$ have little variation left, then it might become hard for the R-learner to detect any meaningful effect. For example, if we include feature "AME.FF" and "ABE.m" in the controls when book-to-market ratio is the treatment, the residualized treatment $\dot{Z} - \hat{e}(\dot{X})$ has very small variance and very close to 0. We try to remove such covariates whenever we see that $e(x)$ model does too good of a job predicting treatment $\dot{Z}$ to a point the root-mean-squared-error (RMSE) loss falls below 0.05. To provide a reference to that number, since we cross-sectionally standardize treatment to have mean 0 and

standard deviation 1 for each month, if we just use 0 for prediction we will achieve a RMSE loss of 1. We mark those exclusions of features for different treatment cases in Table D.2.

Features	Description
ME′	Market equity
ret_sigma	Past month daily return volatility
ret.1m	Past month return
turnover∗	Stock turnover defined as volume-to-market-cap ratio during past month
avg_daily_dVol_million′	Average daily dollar volume during the past month
avg_daily_Vol_thousand′	Average daily volume during the past month
BE	Book Equity based on Fama-French definition
OpProf	Operating profits
GrProf	Gross profits
Cflow	Free cash flow
AstChg∗	Percentage change of total asset from the previous 10K
BM.FF∗	Book-to-market ratio based on Fama French definition
OpIB∗	Operating profit to book equity ratio
GrIA∗	Gross profit to total asset ratio
EP.FF∗	Earning-to-price ratio based on Fama French definition
DP.likeFF∗	Dividend yield based on Fama French definition
AME.FF∗	Asset divided by market equity based on Fama French definition
ABE.m∗	Asset divided by book equity.
CFP.FF∗	Cashflow divided by market equity based on Fama French definition
at′	Total asset
pstkl′	Preferred stock liquidating value
txdi	Income taxes - deferred
txdb	Deferred taxes
itcb	Investment tax credit
revt′	Total revenue
xint	Interests expense
xi	Extraordinary items (in the income statement)
dvc′	Dividend for common/ordinary equity (Cash Flow items)
act′	Total current assets
che	Cash and short-term investments
dlc	Total debt in current liabilities
txp	Income taxes payable
dp	Depreciation and amortization
invt′	Total inventories
ni	Net income
FFIC	Fama-French industry categories based on French's website

Table D.1: **Feature list.** We list here all features collected for our empirical exercise. One feature from the list is selected to be the treatment variable and the rest is used as controls. Features with a $*$ after their names are ratio-based which are not standardized. The rest features are standardized cross-sectionally before feeding into our R-learner procedure. We take log of some non-ratio-based features before standardization to make the distribution more like normal distribution. We mark those features by $'$. Values of 0 would become NA after log transformation.

Treatment Variable	Controls excluded
value	AME.FF
	ABE.m
size	NULL
momentum	NULL

Table D.2: Features excluded for certain treatments. We document here in this table if we ever exclude any features from full feature list (Table D.1) due to e model is fitted too well. The criterion we use is when best cross-validation loss from our hyperparameter grid search gives a CV loss under 0.05 for e model.

D.2 Detailed Portfolio Sorting Results for Each $\hat{\tau}$ Quintile

In this appendix subsection, we list the detailed single sort results on treatment variable restricted to each of the five $\hat{\tau}$ quintiles. We show the results for value, size and momentum as treatment variable in Table D.3, D.4 and D.5, respectively.

D.3 Restricted τ Model on Top-3 Ex-post Features

In this subsection, we repeat our R-learner procedure but restrict ourselves to only the top-3 most important features (ex-post) for the τ model. Recall that there is still significant portion of the HTE is explained by the rest of features that are probably less commonly studied. In fact, the non-top-3 features combined explain the majority of HTE (73% for value, 60% for size and 54% for momentum). Thus we expect that performance of our long-short of long-short tests should take a hit if if we only use the ex-post top 3 features in our τ model. We emphasize that even if the top-3 features can explain all the performance, our procedure still has merits since the top 3 important features are not known beforehand and only identified by our R-learner in the rolling windows.

From the result of section 4.5.2, our procedure has a lot more to offer than a simple interaction of expost most important features and treatment variables. We are then interested in exploring whether our heterogeneity results are mainly driven by the top-3 features if we control the top-3 features in a complicated non-linear fashion instead of using simple interactions. To implement this idea, we rerun our entire procedure but force training of τ model to only use the top-3 ex-post important features. We emphasise again that the purpose of this study is to understand what drives the results and the result is not achievable in practice since the top-3 features are computed using all rolling windows and thus are available only ex-post.

D.3.0.1 Value as the Treatment

We can see that the results change a lot by comparing with Table 4.3. In particular, our long-short of long-short test result is much weaker and in fact is not significantly positive anymore. This shows that for value as treatment, the non top-3 features really contribute a lot and if we only use the top-3 features, the predictive power actually is completely lost. This is somewhat expected since the top-3 features together only explain 27% of the loss reduction achieved by τ model according to Table 4.6.

D.3.0.2 Size as the Treatment

Again, we compare the results in Table D.7 and Table 4.4. We can see that by only including the top-3 features, the predictive power of long-short of long-short test still holds: the average returns and alphas are positive and statistically significant. However, the magnitude of average return and alphas decrease a bit. So do the t-stats of average return and alphas. For example, the FF-5-MOM alpha decrease by around 20% from 1.80% per month to 1.47% per month. It is interesting to note that the size responders, $\hat{\tau}$ 1 quintile in this case since we want the quintile with the most negative $\hat{\tau}$'s, have average $\hat{\tau}$ of -0.56 compared with -0.68 before. The difference in average $\hat{\tau}$

between $\hat{\tau}$ 5 and $\hat{\tau}$ 1 quintile decrease from 0.74 in Table 4.4 to 0.57 in Table D.7. That is roughly a 20% decrease again. This shows that the model's ability to capture heterogeneity decreases due to the loss of non-top-3 features and that it is reflected in the $\hat{\tau}$'s.

D.3.0.3 Momentum as the Treatment

We compare the results in Table D.8 and Table 4.5. Surprisingly the long-short of long-short tests have a much higher average return 1.57% compared with 1.18% per month before when using all features. The average return also have a much higher t-stats 4.30 too. For FF-5-MOM alpha, we do see similar results as before. When restricting τ model to the top-3 expost features, we have a FF-5-MOM alpha of 1.69% per month compared with 1.65% per month on all features. We do see slightly more statistical significance of the FF-5-MOM alpha now with compared 4.15 with 3.75 before. By focusing on the top-3 features, it seems that we are able to achieve essentially the same alpha with a higher t-stats. That seems to suggests that the results in Table 4.5 are mainly driven by top 3 features and the rest features will add noise to the model for our long-short of long-short test of momentum as treatment. That might not be too surprising considering the fact that the top 3 features alone explain 46% of loss reduction achieved by τ model, which is almost 50% and is the most comparing with cases where value and size are used as treatment.

D.4 Restricted τ Model on Top-10 Ex-post Features

We repeat the analysis in section D.3 but instead of restricting training of τ model to the top 3 ex-post most important features, we include the top 10. We show the

results for value, size and momentum as treatment variable in Table D.9, D.10 and D.11, respectively.

D.5 Results with Only Demeaning Variables in the Cross Section

In our main results from section 4.4, we standardize all non-ratio-based variables in the cross section in each month so that all non-ratio-based variables have 0 mean and standard deviation 1 for each month. In section 4.6, we apply R-learner procedure to raw data without any standardization to show the impact of our pre-processing procedure. In this subsection of the appendix, we present results with something in between: we only demean all non-ratio-based variables (including returns since returns are non-ratio-based) in the cross section such that all variables have 0 mean during each month. As in our main results, ratio-based variables are unchanged during the pre-processing. We show the results for value, size and momentum as treatment variables in Tables D.12, D.13, and D.14, respectively.

D.6 Value Weighted Results for Long-short of Long-short Tests

We repeat the analysis in section 4.4.2 but instead of reporting results with all equal-weighted portfolios, we show results with all value-weighted portfolios in our long-short of long-short tests. This should alleviate the concerns on overweighting small stocks in trading strategies. However, as discussed before, our procedure assigns equal weights to all data points during training and we should expect a performance decrease if the testing distribution is different from the training one. The decrease in performance should be particularly large in our case since we are utilizing highly flex-

ible machine learning algorithm gradient boosting, which tries to exploit all sorts of relationships in the training data. We show the results for value, size and momentum as treatment variables in Table D.15, D.16 and D.17, respectively.

D.7 Training vs. Test Results for Long-short of Long-short Tests

In the last section of the appendix, we present the "in-sample" average returns of our long-short of long-short strategies, compared with our main results which is conducted in a predictive fashion without look-ahead bias. Tables D.18, D.19, and D.20 show the results with value, size, and momentum as the treatment variable, respectively. Each of the three tables have three different long-short of long-short results: "training", "test oneshot", and "test rolling". For "training", we calculate the average returns of P5-P1 long-short trading strategies restricted to each $\hat{\tau}$ quintile, for the period of the first rolling window for training (199601-200512) before our out-of-sample test period starts. The $\hat{\tau}$ quintiles used in forming long-short of long-short strategies are determined by one single τ model fitted using data in the same time window (199601-200512). Since we form the five $\hat{\tau}$ quintiles in the 10-year window used in the training of the τ model, we consider this as an in-sample test for our long-short of long-short strategies. For "test oneshot", we keep the one single τ model trained from the first rolling window 199601-200512 and repeat the same calculation for our entire out-of-sample test period 200601-201812. This is an out-of-sample test because we backtest the strategies in the test period 200601-201812 while the τ models determining $\hat{\tau}$ quintiles are fitted using periods that are non-overlapping (199601-200512). This "oneshot" out-of-sample test is unrealistic and unlikely to work well because it only fits the model once with data before year 1996. For "test rolling", we repeat the calculations using the same method as in our main results: the average returns of

long-short of long-short strategies are calculated for our out-of-sample test period 200601-201812 with annual refittings using the previous 10 years as training data. For each year in the test period, $\hat{\tau}$ quintiles are determined by a τ model trained using the 10-year rolling window before that year. In total we train 13 τ models since we have 13 10-year rolling windows. All return numbers are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively.

We can see that across all three treatment variables, average returns of our long-short of long-short strategies in training set have much better performance than our main results, "test rolling", which is expected since training results are in-sample and not implementable in practice. One interesting observation is that the gaps between our in-sample and out-of-sample results are huge: for value as the treatment variable, the average return of the long-short of long-short strategy is 1.72% in "training" vs. 0.77% in "test rolling"; for size as the treatment, the return is 4.35% for the "training" and 1.54% for the "test rolling", whereas for momentum as the treatment, 5.15% for " training" and 1.18% for "test rolling". The large gap suggests that there might be some overfitting in our model training and a more careful hyperparameter tuning for controlling overfitting might shrink the gap between in-sample and out-of-sample rolling results and improve out-of-sample performance even more.

Lastly, we note that for all three teatment varibles, "test oneshot" does not perform well in that the average returns of long-short of long-short strategies are smaller than "test rolling" and not significantly positive. This confirms our intuition that the financial market evolves over time and it is hard for models trained using data from 199601 to 200512 to generalize well into the next 20 years. Our yearly refitting helps a lot in performance, which is obvious from comparing "test oneshot" with "test rolling".

		$\dot{Z}$ 1	$\dot{Z}$ 2	$\dot{Z}$ 3	$\dot{Z}$ 4	$\dot{Z}$ 5	$P5-P1$
Panel 1	Mean	1.26*	1.18*	1.03*	0.97	1.20	-0.06
$\hat{\tau}$ 1		[1.95]	[1.78]	[1.69]	[1.47]	[1.53]	[-0.14]
	CAPM Alpha	0.26	0.21	0.07	0.04	0.17	-0.10
		[0.76]	[0.57]	[0.23]	[0.10]	[0.38]	[-0.25]
	FF-3-MOM Alpha	0.35	0.36	0.32	0.35	0.54*	0.19
		[1.36]	[1.34]	[1.62]	[1.33]	[1.73]	[0.63]
	FF-5-MOM Alpha	0.76***	0.62**	0.53***	0.47*	0.63*	-0.13
		[2.75]	[2.20]	[2.64]	[1.71]	[1.88]	[-0.42]
	avg $\dot{Z}$	-1.94	-0.98	-0.53	-0.15	0.46	
	avg. BM ratio	0.27	0.58	0.89	1.25	2.85	
	avg. ME	6.47	4.56	3.92	2.69	1.75	
	avg. Monthly Vol	33.01	26.70	26.46	27.43	25.74	
	persistence	0.33	0.35	0.34	0.33	0.30	
Panel 2	Mean	0.99**	0.78*	0.82*	0.75*	0.82*	-0.16
$\hat{\tau}$ 2		[2.16]	[1.73]	[1.84]	[1.70]	[1.73]	[-0.65]
	CAPM Alpha	0.17	-0.04	0.01	-0.03	0.03	-0.14
		[0.95]	[-0.33]	[0.03]	[-0.17]	[0.16]	[-0.54]
	FF-3-MOM Alpha	0.19	0.03	0.12	0.13	0.22	0.04
		[1.58]	[0.30]	[1.21]	[1.35]	[1.59]	[0.23]
	FF-5-MOM Alpha	0.34***	0.12	0.15	0.17*	0.28*	-0.06
		[3.18]	[1.13]	[1.52]	[1.73]	[1.92]	[-0.38]
	avg $\dot{Z}$	-1.89	-1.01	-0.59	-0.24	0.27	
	avg. BM ratio	0.19	0.39	0.59	0.81	1.46	
	avg. ME	8.72	6.82	4.57	3.12	2.34	
	avg. Monthly Vol	32.12	26.86	23.75	17.66	18.39	
	persistence	0.42	0.43	0.43	0.42	0.44	
Panel 3	Mean	0.60	0.66	0.72*	0.80*	0.86*	0.26
$\hat{\tau}$ 3		[1.40]	[1.61]	[1.67]	[1.91]	[1.86]	[1.15]
	CAPM Alpha	-0.20	-0.11	-0.07	0.06	0.11	0.31
		[-1.31]	[-0.78]	[-0.47]	[0.30]	[0.50]	[1.34]
	FF-3-MOM Alpha	−0.19*	-0.06	0.05	0.21*	0.28*	0.47***
		[-1.93]	[-0.56]	[0.53]	[1.85]	[1.74]	[2.88]
	FF-5-MOM Alpha	−0.16*	-0.07	0.06	0.19*	0.32**	0.47***
		[-1.65]	[-0.73]	[0.66]	[1.80]	[2.08]	[3.02]
	avg $\dot{Z}$	-1.87	-1.06	-0.64	-0.28	0.25	
	avg. BM ratio	0.18	0.37	0.56	0.79	1.41	
	avg. ME	9.32	7.46	5.02	3.02	2.09	
	avg. Monthly Vol	31.09	28.20	21.07	16.39	21.25	
	persistence	0.43	0.44	0.45	0.45	0.47	
Panel 4	Mean	0.57	0.55	0.58	0.54	0.75	0.18
$\hat{\tau}$ 4		[1.25]	[1.19]	[1.29]	[1.30]	[1.58]	[0.77]
	CAPM Alpha	-0.25	-0.27	-0.22	-0.19	0.02	0.27
		[-1.44]	[-1.62]	[-1.31]	[-1.16]	[0.08]	[1.16]
	FF-3-MOM Alpha	−0.22**	−0.20*	-0.11	-0.08	0.18	0.40**
		[-2.09]	[-1.88]	[-0.88]	[-0.64]	[0.99]	[2.23]
	FF-5-MOM Alpha	-0.14	-0.14	-0.07	-0.08	0.21	0.35**
		[-1.36]	[-1.30]	[-0.60]	[-0.62]	[1.23]	[1.99]
	avg $\dot{Z}$	-1.89	-1.05	-0.63	-0.26	0.30	
	avg. BM ratio	0.19	0.38	0.57	0.81	1.47	
	avg. ME	7.67	5.80	3.47	2.74	2.27	
	avg. Monthly Vol	26.38	23.81	18.04	14.17	16.70	
	persistence	0.41	0.41	0.42	0.42	0.44	
Panel 5	Mean	-0.27	0.22	0.31	0.37	0.45	0.71**
$\hat{\tau}$ 5		[-0.46]	[0.41]	[0.63]	[0.76]	[0.75]	[2.15]
	CAPM Alpha	−1.18***	−0.68**	−0.50**	−0.40*	-0.38	0.80**
		[-3.83]	[-2.44]	[-2.09]	[-1.80]	[-1.09]	[2.37]
	FF-3-MOM Alpha	−1.14***	−0.63***	−0.41**	-0.28	-0.17	0.97***
		[-4.95]	[-3.21]	[-2.27]	[-1.44]	[-0.54]	[3.42]
	FF-5-MOM Alpha	−0.91***	−0.50***	−0.32*	-0.20	-0.06	0.85***
		[-4.31]	[-2.71]	[-1.67]	[-1.08]	[-0.19]	[2.84]
	avg $\dot{Z}$	-2.23	-1.10	-0.56	-0.12	0.54	
	avg. BM ratio	0.17	0.43	0.71	1.09	2.46	
	avg. ME	8.74	9.84	6.61	5.17	2.25	
	avg. Monthly Vol	33.83	38.44	31.94	23.77	27.26	
	persistence	0.28	0.35	0.35	0.34	0.29	

Table D.3: Detailed long-short test results restricted to different $\hat{\tau}$ quintiles for value as the treatment. We calculate average equal weighted returns of portfolios sorted based on treatment but restricted to each τ quintile. We also report alpha and additional information about each portfolio. Return and alpha numbers are monthly in percentage. Inside square brackets are t-stats with Newey-West standard errors. One, two and three stars represent significance level 10%, 5%, and 1% respectively.

		$\dot{Z}$ 1	$\dot{Z}$ 2	$\dot{Z}$ 3	$\dot{Z}$ 4	$\dot{Z}$ 5	$P5-P1$
Panel 1	Mean	2.01**	-0.18	-0.18	0.01	0.58	−1.43**
$\hat{\tau}$ 1		[2.41]	[-0.27]	[-0.26]	[0.01]	[0.89]	[-2.48]
	CAPM Alpha	1.15*	−1.05**	−1.13***	−1.04***	-0.51	−1.66***
		[1.75]	[-2.24]	[-2.69]	[-3.09]	[-1.58]	[-2.80]
	FF-3-MOM Alpha	1.37**	−0.96**	−0.99***	−0.82***	-0.33	−1.70***
		[2.48]	[-2.36]	[-2.82]	[-3.44]	[-1.42]	[-3.03]
	FF-5-MOM Alpha	1.76***	-0.62	−0.71**	−0.63***	-0.10	−1.86***
		[2.84]	[-1.42]	[-1.97]	[-2.58]	[-0.49]	[-2.96]
	avg $\dot{Z}$	-1.86	-1.35	-0.97	-0.52	0.31	
	avg. BM ratio	2.20	1.17	0.92	0.68	0.48	
	avg. ME	0.01	0.03	0.07	0.20	2.67	
	avg. Monthly Vol	3.20	4.10	5.71	11.02	64.87	
	persistence	0.20	0.36	0.45	0.52	0.55	
Panel 2	Mean	0.72	0.86*	0.52	0.65	0.69	-0.04
$\hat{\tau}$ 2		[1.45]	[1.67]	[0.91]	[1.16]	[1.28]	[-0.12]
	CAPM Alpha	0.07	0.06	-0.41	-0.33	-0.25	-0.32
		[0.23]	[0.21]	[-1.60]	[-1.42]	[-1.17]	[-1.11]
	FF-3-MOM Alpha	0.19	0.19	-0.23	-0.19	-0.19	-0.39
		[0.67]	[0.86]	[-1.38]	[-1.35]	[-1.21]	[-1.36]
	FF-5-MOM Alpha	0.45	0.43**	-0.03	-0.07	-0.04	−0.49*
		[1.61]	[2.00]	[-0.16]	[-0.50]	[-0.28]	[-1.70]
	avg $\dot{Z}$	-1.30	-0.73	-0.32	0.13	0.89	
	avg. BM ratio	1.40	0.92	0.69	0.54	0.42	
	avg. ME	0.04	0.12	0.30	0.80	7.78	
	avg. Monthly Vol	1.20	2.84	6.93	15.39	73.43	
	persistence	0.53	0.56	0.59	0.60	0.55	
Panel 3	Mean	0.51	0.73	0.78	0.78	0.82*	0.31
$\hat{\tau}$ 3		[1.08]	[1.48]	[1.50]	[1.53]	[1.77]	[1.29]
	CAPM Alpha	-0.17	-0.09	-0.12	-0.13	-0.03	0.14
		[-0.61]	[-0.41]	[-0.55]	[-0.76]	[-0.21]	[0.54]
	FF-3-MOM Alpha	-0.06	0.06	0.05	-0.01	0.00	0.06
		[-0.26]	[0.38]	[0.38]	[-0.08]	[0.01]	[0.26]
	FF-5-MOM Alpha	0.15	0.09	0.14	0.00	0.03	-0.12
		[0.65]	[0.62]	[1.10]	[-0.04]	[0.26]	[-0.50]
	avg $\dot{Z}$	-0.94	-0.31	0.12	0.56	1.33	
	avg. BM ratio	1.14	0.76	0.62	0.50	0.43	
	avg. ME	0.10	0.33	0.84	2.12	18.84	
	avg. Monthly Vol	1.55	4.39	9.80	20.77	85.27	
	persistence	0.58	0.55	0.55	0.51	0.44	
Panel 4	Mean	0.55	0.70	0.85*	0.70	0.66*	0.12
$\hat{\tau}$ 4		[1.18]	[1.43]	[1.82]	[1.58]	[1.67]	[0.55]
	CAPM Alpha	-0.18	-0.15	0.01	-0.12	-0.10	0.08
		[-0.77]	[-0.68]	[0.03]	[-0.89]	[-1.28]	[0.36]
	FF-3-MOM Alpha	-0.05	0.04	0.17**	-0.02	-0.07	-0.02
		[-0.28]	[0.32]	[1.97]	[-0.18]	[-0.98]	[-0.12]
	FF-5-MOM Alpha	0.11	0.09	0.17*	-0.05	-0.10	-0.21
		[0.63]	[0.79]	[1.90]	[-0.49]	[-1.34]	[-1.17]
	avg $\dot{Z}$	-0.69	-0.06	0.36	0.79	1.58	
	avg. BM ratio	1.00	0.68	0.56	0.48	0.44	
	avg. ME	0.16	0.50	1.20	2.99	26.36	
	avg. Monthly Vol	1.69	4.79	10.46	22.43	100.02	
	persistence	0.54	0.54	0.50	0.48	0.44	
Panel 5	Mean	0.68	0.86*	0.89**	0.86**	0.78**	0.11
$\hat{\tau}$ 5		[1.53]	[1.92]	[2.10]	[2.14]	[2.20]	[0.45]
	CAPM Alpha	-0.05	0.04	0.09	0.10	0.08	0.13
		[-0.23]	[0.22]	[0.62]	[0.83]	[0.93]	[0.55]
	FF-3-MOM Alpha	0.09	0.23**	0.20*	0.14	0.09	0.00
		[0.56]	[2.00]	[1.92]	[1.29]	[1.01]	[-0.01]
	FF-5-MOM Alpha	0.11	0.15	0.19*	0.12	0.05	-0.05
		[0.67]	[1.34]	[1.71]	[1.10]	[0.61]	[-0.30]
	avg $\dot{Z}$	-0.51	0.15	0.58	1.03	1.79	
	avg. BM ratio	0.95	0.69	0.57	0.50	0.44	
	avg. ME	0.29	0.99	2.33	5.97	42.97	
	avg. Monthly Vol	3.76	7.90	15.65	31.39	106.42	
	persistence	0.45	0.43	0.39	0.34	0.28	

Table D.4: Detailed long-short test results restricted to different $\hat{\tau}$ quintiles for size as the treatment. We calculate average equal weighted returns of portfolios sorted based on treatment but restricted to each τ quintile. We also report alpha and additional information about each portfolio. Return and alpha numbers are monthly in percentage. Inside square brackets are t-stats with Newey-West standard errors. One, two and three stars represent significance level 10%, 5%, and 1% respectively.

		$\dot{Z}$ 1	$\dot{Z}$ 2	$\dot{Z}$ 3	$\dot{Z}$ 4	$\dot{Z}$ 5	$P5-P1$
Panel 1	Mean	1.57*	0.75*	0.87**	0.90***	0.96**	-0.61
$\hat{\tau}$ 1		[1.85]	[1.66]	[2.39]	[2.87]	[2.44]	[-0.87]
	CAPM Alpha	0.55	-0.01	0.19	0.27**	0.26	-0.29
		[1.00]	[-0.03]	[1.60]	[2.16]	[1.23]	[-0.51]
	FF-3-MOM Alpha	0.92**	0.16	0.27***	0.30***	0.27*	−0.66*
		[2.20]	[1.12]	[3.04]	[2.62]	[1.66]	[-1.67]
	FF-5-MOM Alpha	1.37***	0.20	0.26***	0.26**	0.24	−1.12**
		[2.72]	[1.29]	[2.81]	[2.44]	[1.41]	[-2.48]
	avg $\dot{Z}$	-1.01	-0.34	-0.05	0.21	0.83	
	avg. BM ratio	2.00	0.79	0.61	0.53	0.45	
	avg. ME	4.82	16.86	21.18	23.15	19.78	
	avg. Monthly Vol	42.33	56.88	56.26	55.77	50.30	
	persistence	0.36	0.50	0.55	0.52	0.36	
Panel 2	Mean	0.84	0.83*	0.94**	0.77**	0.76*	-0.08
$\hat{\tau}$ 2		[1.18]	[1.73]	[2.42]	[2.02]	[1.84]	[-0.15]
	CAPM Alpha	-0.18	0.01	0.22	0.06	0.01	0.18
		[-0.48]	[0.04]	[1.53]	[0.43]	[0.03]	[0.44]
	FF-3-MOM Alpha	0.09	0.17	0.35***	0.14	0.02	-0.06
		[0.40]	[1.59]	[3.97]	[1.46]	[0.21]	[-0.32]
	FF-5-MOM Alpha	0.23	0.14	0.34***	0.13	0.04	-0.19
		[1.00]	[1.34]	[3.99]	[1.34]	[0.33]	[-0.90]
	avg $\dot{Z}$	-0.78	-0.29	-0.03	0.26	0.99	
	avg. BM ratio	1.25	0.69	0.60	0.53	0.42	
	avg. ME	2.59	3.96	4.81	5.42	5.49	
	avg. Monthly Vol	36.39	26.69	25.58	25.70	31.48	
	persistence	0.49	0.60	0.64	0.60	0.46	
Panel 3	Mean	0.71	0.88*	1.01**	0.87**	0.74	0.03
$\hat{\tau}$ 3		[1.02]	[1.82]	[2.37]	[2.09]	[1.58]	[0.06]
	CAPM Alpha	-0.31	0.06	0.24	0.13	-0.08	0.22
		[-0.76]	[0.28]	[1.47]	[0.79]	[-0.39]	[0.50]
	FF-3-MOM Alpha	-0.06	0.21*	0.37***	0.23**	-0.03	0.03
		[-0.28]	[1.79]	[3.14]	[1.99]	[-0.20]	[0.13]
	FF-5-MOM Alpha	0.02	0.24**	0.41***	0.28**	0.00	-0.02
		[0.07]	[2.05]	[3.52]	[2.54]	[0.00]	[-0.07]
	avg $\dot{Z}$	-0.89	-0.35	-0.05	0.28	1.19	
	avg. BM ratio	1.26	0.74	0.65	0.56	0.42	
	avg. ME	1.03	1.54	1.92	1.99	2.07	
	avg. Monthly Vol	21.93	18.05	15.80	16.22	19.31	
	persistence	0.55	0.64	0.66	0.64	0.51	
Panel 4	Mean	0.34	0.54	0.73	0.77*	0.70	0.36
$\hat{\tau}$ 4		[0.41]	[0.91]	[1.42]	[1.68]	[1.34]	[0.64]
	CAPM Alpha	−0.76*	-0.40	-0.13	-0.03	-0.18	0.58
		[-1.66]	[-1.47]	[-0.60]	[-0.14]	[-0.70]	[1.25]
	FF-3-MOM Alpha	-0.51	-0.21	0.00	0.05	-0.21	0.30
		[-1.63]	[-1.15]	[-0.02]	[0.39]	[-1.22]	[1.07]
	FF-5-MOM Alpha	-0.33	-0.10	0.07	0.13	-0.07	0.26
		[-1.05]	[-0.59]	[0.49]	[0.93]	[-0.44]	[0.86]
	avg $\dot{Z}$	-1.09	-0.51	-0.13	0.30	1.49	
	avg. BM ratio	1.18	0.78	0.67	0.57	0.39	
	avg. ME	0.52	0.87	1.00	1.16	1.24	
	avg. Monthly Vol	20.60	16.19	13.84	13.49	17.99	
	persistence	0.53	0.67	0.71	0.68	0.52	
Panel 5	Mean	-0.17	0.12	0.35	0.50	0.41	0.57
$\hat{\tau}$ 5		[-0.20]	[0.19]	[0.65]	[1.01]	[0.78]	[1.08]
	CAPM Alpha	−1.24**	−0.83**	−0.48*	-0.29	-0.44	0.80*
		[-2.31]	[-2.37]	[-1.71]	[-1.11]	[-1.50]	[1.72]
	FF-3-MOM Alpha	−1.02**	−0.66**	-0.36	-0.19	−0.43**	0.59*
		[-2.45]	[-2.39]	[-1.63]	[-0.96]	[-2.06]	[1.76]
	FF-5-MOM Alpha	−0.70*	-0.40	-0.17	-0.02	-0.18	0.53
		[-1.65]	[-1.61]	[-0.81]	[-0.10]	[-1.00]	[1.43]
	avg $\dot{Z}$	-1.24	-0.65	-0.24	0.22	1.54	
	avg. BM ratio	1.33	0.87	0.74	0.63	0.44	
	avg. ME	0.22	0.46	0.58	0.64	0.63	
	avg. Monthly Vol	16.38	10.89	8.84	8.91	13.69	
	persistence	0.42	0.61	0.66	0.64	0.42	

Table D.5: Detailed long-short test results restricted to different $\hat{\tau}$ quintiles for momentum as the treatment. We calculate average equal weighted returns of portfolios sorted based on treatment but restricted to each τ quintile. We also report alpha and additional information about each portfolio. Return and alpha numbers are monthly in percentage. Inside square brackets are t-stats with Newey-West standard errors. One, two and three stars represent significance level 10%, 5%, and 1% respectively.

	full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.25	0.33	0.56	0.08	0.01	0.43	0.10
	[0.99]	[0.83]	[1.41]	[0.40]	[0.04]	[1.55]	[0.36]
CAPM Alpha	0.29	0.34	0.67*	0.13	0.07	0.44	0.10
	[1.13]	[0.87]	[1.75]	[0.61]	[0.27]	[1.59]	[0.36]
FF-3-MOM Alpha	0.49***	0.58**	0.83**	0.27**	0.23	0.64***	0.06
	[3.04]	[1.97]	[2.47]	[1.98]	[1.16]	[3.12]	[0.24]
FF-5-MOM Alpha	0.39**	0.38	0.64*	0.20	0.19	0.63***	0.25
	[2.30]	[1.23]	[1.86]	[1.51]	[0.91]	[2.83]	[0.94]
τ 25%	-0.09	-0.26	-0.09	-0.04	0.00	0.04	
τ 50%	-0.03	-0.17	-0.08	-0.03	0.01	0.07	
τ 75%	0.01	-0.13	-0.06	-0.02	0.02	0.13	
avg τ	-0.04	-0.23	-0.08	-0.03	0.01	0.11	
sd τ	0.17	0.19	0.02	0.01	0.01	0.12	
avg $\dot{Z}$	-0.69	-0.66	-0.72	-0.74	-0.72	-0.68	
avg $\dot{Z}$ in 1st quintile	-1.98	-2.05	-1.96	-1.95	-1.95	-2.01	
avg $\dot{Z}$ in 5th quintile	0.38	0.55	0.29	0.22	0.24	0.42	
avg $\dot{Z}$ P5 - P1	2.36	2.60	2.25	2.17	2.19	2.44	
number of stocks	3536.90	810.88	639.92	734.91	679.13	928.43	
avg. BM ratio	0.84	1.12	0.71	0.65	0.69	0.95	
avg. ME	5.08	1.26	2.37	3.68	7.43	9.61	
avg. Monthly Vol	25.27	18.34	17.03	18.64	25.10	40.97	
persistence		0.48	0.31	0.39	0.40	0.51	

Table D.6: **Long-short of long-short test results for value as the treatment using equal weighting (τ model trained with top 3 ex-post most important features).** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. m and e models are trained the same way as before while for τ model, we take the ex-post top3 features according to Table 4.6 and only include those three features in the training of τ model. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable book-to-market ratios but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.15	-1.00	0.28	-0.07	-0.02	0.32	1.32**
	[0.43]	[-1.62]	[0.83]	[-0.28]	[-0.10]	[1.43]	[2.39]
CAPM Alpha	0.14	−1.33**	0.08	-0.22	-0.19	0.17	1.49***
	[0.40]	[-2.10]	[0.22]	[-0.90]	[-0.78]	[0.73]	[2.64]
FF-3-MOM Alpha	0.03	−1.29**	0.03	-0.30	-0.23	0.11	1.40***
	[0.11]	[-2.22]	[0.08]	[-1.25]	[-0.98]	[0.54]	[2.61]
FF-5-MOM Alpha	-0.24	−1.52**	-0.25	−0.42*	−0.41**	-0.04	1.47**
	[-0.75]	[-2.24]	[-0.71]	[-1.73]	[-2.00]	[-0.22]	[2.31]
τ 25%	-0.27	-0.64	-0.27	-0.12	-0.07	-0.02	
τ 50%	-0.10	-0.49	-0.22	-0.11	-0.06	0.00	
τ 75%	-0.05	-0.40	-0.17	-0.09	-0.05	0.02	
avg τ	-0.19	-0.56	-0.22	-0.11	-0.06	0.01	
sd τ	0.24	0.27	0.06	0.02	0.01	0.06	
avg $\dot{Z}$	0.00	-0.72	-0.16	0.17	0.31	0.41	
avg $\dot{Z}$ in 1st quintile	-1.35	-1.80	-1.34	-1.07	-0.96	-0.90	
avg $\dot{Z}$ in 5th quintile	1.43	0.50	1.06	1.45	1.62	1.78	
avg $\dot{Z}$ P5 - P1	2.78	2.30	2.40	2.52	2.58	2.67	
number of stocks	3902.94	782.83	770.19	753.90	792.95	803.07	
avg. BM ratio	0.77	1.27	0.70	0.63	0.62	0.61	
avg. ME	4.74	0.65	2.20	4.84	6.51	9.32	
avg. Monthly Vol	24.65	22.40	21.65	24.11	26.15	28.60	
persistence		0.61	0.36	0.30	0.33	0.45	

Table D.7: **Long-short of long-short test results for size as the treatment using equal weighting (τ model trained with top 3 ex-post most important features).** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. m and e models are trained the same way as before while for τ model, we take the ex-post top3 features according to Table 4.7 and only include those three features in the training of τ model. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable firm sizes but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.16	-0.98	0.20	-0.08	0.63	0.59	1.57***
	[0.29]	[-1.62]	[0.44]	[-0.18]	[1.12]	[1.04]	[4.30]
CAPM Alpha	0.40	-0.73	0.39	0.03	0.83*	0.82*	1.55***
	[0.90]	[-1.38]	[0.93]	[0.07]	[1.74]	[1.67]	[4.19]
FF-3-MOM Alpha	0.14	−0.99***	0.22	-0.15	0.48*	0.54	1.53***
	[0.60]	[-2.66]	[0.80]	[-0.57]	[1.78]	[1.57]	[4.04]
FF-5-MOM Alpha	-0.03	−1.27***	0.00	-0.24	0.48*	0.43	1.69***
	[-0.12]	[-3.05]	[0.01]	[-0.92]	[1.68]	[1.12]	[4.14]
τ 25%	-0.04	-0.08	-0.04	-0.02	-0.02	-0.01	
τ 50%	-0.02	-0.07	-0.03	-0.02	-0.02	-0.01	
τ 75%	-0.01	-0.05	-0.03	-0.02	-0.01	0.00	
avg τ	-0.03	-0.08	-0.03	-0.02	-0.02	0.00	
sd τ	0.05	0.07	0.01	0.00	0.00	0.03	
avg $\dot{Z}$	-0.01	-0.01	0.04	0.04	-0.02	-0.13	
avg $\dot{Z}$ in 1st quintile	-1.05	-0.94	-0.73	-0.84	-1.07	-1.30	
avg $\dot{Z}$ in 5th quintile	1.23	0.84	0.96	1.15	1.39	1.60	
avg $\dot{Z}$ P5 - P1	2.28	1.78	1.69	1.99	2.46	2.90	
number of stocks	3681.14	776.09	653.47	711.77	703.79	908.87	
avg. BM ratio	0.77	0.78	0.67	0.70	0.83	0.88	
avg. ME	4.96	19.49	2.12	1.23	0.77	0.51	
avg. Monthly Vol	25.54	61.56	17.02	14.41	16.20	15.23	
persistence		0.79	0.49	0.44	0.40	0.58	

Table D.8: **Long-short of long-short test results for momentum as the treatment using equal weighting (τ model trained with top 3 ex-post most important features).** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. m and e models are trained the same way as before while for τ model, we take the ex-post top3 features according to Table 4.8 and only include those three features in the training of τ model. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable trailing returns but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are all equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.25	0.33	-0.21	-0.13	0.21	0.17	-0.16
	[0.99]	[0.89]	[-0.56]	[-0.53]	[0.94]	[0.53]	[-0.49]
CAPM Alpha	0.29	0.36	-0.17	-0.03	0.27	0.22	-0.14
	[1.13]	[0.97]	[-0.45]	[-0.11]	[1.19]	[0.66]	[-0.42]
FF-3-MOM Alpha	0.49***	0.57*	-0.09	0.07	0.41**	0.41*	-0.15
	[3.04]	[1.80]	[-0.25]	[0.34]	[2.41]	[1.71]	[-0.46]
FF-5-MOM Alpha	0.39**	0.23	-0.25	0.02	0.34**	0.32	0.09
	[2.30]	[0.72]	[-0.68]	[0.13]	[2.07]	[1.27]	[0.25]
τ 25%	-0.08	-0.26	-0.08	-0.04	-0.01	0.04	
τ 50%	-0.04	-0.16	-0.07	-0.03	-0.01	0.07	
τ 75%	0.00	-0.11	-0.06	-0.03	0.01	0.15	
avg τ	-0.04	-0.23	-0.07	-0.03	0.00	0.13	
sd τ	0.17	0.21	0.01	0.01	0.01	0.16	
avg $\dot{Z}$	-0.69	-0.59	-0.59	-0.65	-0.74	-0.84	
avg $\dot{Z}$ in 1st quintile	-1.98	-2.00	-1.74	-1.72	-1.84	-2.34	
avg $\dot{Z}$ in 5th quintile	0.38	0.57	0.33	0.24	0.20	0.46	
avg $\dot{Z}$ P5 - P1	2.36	2.57	2.07	1.96	2.04	2.80	
number of stocks	3536.90	716.89	618.70	693.51	790.39	863.61	
avg. BM ratio	0.84	1.23	0.76	0.69	0.63	0.88	
avg. ME	5.08	1.61	3.30	4.24	6.40	8.74	
avg. Monthly Vol	25.27	22.56	16.53	19.26	25.90	38.17	
persistence		0.66	0.49	0.51	0.55	0.71	

Table D.9: **Long-short of long-short test results for value as the treatment using equal weighting (τ model trained with top 10 ex-post most important features).** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. m and e models are trained the same way as before while for τ model, we take the ex-post top 10 features according to Table 4.6 and only include those three features in the training of τ model. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable book-to-market ratios but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.15	-1.46^{**}	-0.03	0.00	0.14	0.17	1.63^{***}
	[0.43]	[-2.46]	[-0.08]	[-0.02]	[0.59]	[0.70]	[2.78]
CAPM Alpha	0.14	-1.72^{***}	-0.23	-0.20	0.11	0.17	1.89^{***}
	[0.40]	[-2.78]	[-0.65]	[-0.76]	[0.43]	[0.73]	[3.08]
FF-3-MOM Alpha	0.03	-1.75^{***}	-0.30	-0.23	0.03	0.07	1.82^{***}
	[0.11]	[-3.12]	[-0.90]	[-0.93]	[0.13]	[0.34]	[3.24]
FF-5-MOM Alpha	-0.24	-1.94^{***}	-0.46	-0.47^{**}	-0.16	0.02	1.96^{***}
	[-0.75]	[-3.05]	[-1.34]	[-2.03]	[-0.82]	[0.09]	[3.16]
τ 25%	-0.35	-0.82	-0.35	-0.19	-0.07	0.01	
τ 50%	-0.16	-0.62	-0.29	-0.16	-0.05	0.04	
τ 75%	-0.03	-0.51	-0.25	-0.14	-0.03	0.08	
avg τ	-0.24	-0.72	-0.30	-0.16	-0.05	0.05	
sd τ	0.31	0.32	0.06	0.03	0.03	0.06	
avg $\dot{Z}$	0.00	-0.85	-0.25	0.14	0.41	0.56	
avg $\dot{Z}$ in 1st quintile	-1.35	-1.85	-1.27	-0.93	-0.70	-0.55	
avg $\dot{Z}$ in 5th quintile	1.43	0.35	0.90	1.31	1.62	1.78	
avg $\dot{Z}$ P5 - P1	2.78	2.20	2.17	2.24	2.32	2.33	
number of stocks	3902.94	781.37	775.69	779.41	778.32	788.15	
avg. BM ratio	0.77	1.10	0.76	0.69	0.63	0.67	
avg. ME	4.74	0.70	1.90	4.41	6.53	10.06	
avg. Monthly Vol	24.65	18.20	18.03	23.04	29.67	34.04	
persistence		0.67	0.46	0.47	0.48	0.63	

Table D.10: **Long-short of long-short test results for size as the treatment using equal weighting (τ model trained with top 10 ex-post most important features).** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. m and e models are trained the same way as before while for τ model, we take the ex-post top 10 features according to Table 4.7 and only include those three features in the training of τ model. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable firm sizes but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.16	-0.52	0.42	0.05	0.40	0.45	0.96**
	[0.29]	[-0.77]	[0.82]	[0.09]	[0.69]	[0.84]	[2.25]
CAPM Alpha	0.40	-0.23	0.72*	0.33	0.57	0.66	0.89**
	[0.90]	[-0.39]	[1.73]	[0.71]	[1.15]	[1.43]	[2.14]
FF-3-MOM Alpha	0.14	-0.49	0.48**	0.01	0.39	0.45	0.94**
	[0.60]	[-1.17]	[2.02]	[0.05]	[1.33]	[1.38]	[2.29]
FF-5-MOM Alpha	-0.03	−0.93**	0.45*	-0.02	0.27	0.30	1.23***
	[-0.12]	[-2.04]	[1.84]	[-0.08]	[0.87]	[0.86]	[2.77]
τ 25%	-0.04	-0.09	-0.04	-0.03	-0.02	0.00	
τ 50%	-0.02	-0.07	-0.04	-0.02	-0.02	0.00	
τ 75%	-0.01	-0.06	-0.03	-0.02	-0.01	0.02	
avg τ	-0.03	-0.09	-0.04	-0.02	-0.02	0.01	
sd τ	0.05	0.08	0.01	0.00	0.00	0.03	
avg $\dot{Z}$	-0.01	-0.06	0.03	0.02	-0.01	-0.06	
avg $\dot{Z}$ in 1st quintile	-1.05	-1.02	-0.83	-0.93	-1.09	-1.18	
avg $\dot{Z}$ in 5th quintile	1.23	0.87	1.05	1.24	1.43	1.45	
avg $\dot{Z}$ P5 - P1	2.28	1.89	1.88	2.17	2.52	2.63	
number of stocks	3681.14	741.12	722.77	704.21	727.37	823.06	
avg. BM ratio	0.77	0.88	0.66	0.72	0.80	0.78	
avg. ME	4.96	16.49	4.12	2.12	1.12	1.00	
avg. Monthly Vol	25.54	57.92	25.24	17.98	10.79	16.27	
persistence		0.78	0.56	0.49	0.52	0.69	

Table D.11: **Long-short of long-short test results for momentum as the treatment using equal weighting (τ model trained with top 10 ex-post most important features)** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. m and e models are trained the same way as before while for τ model, we take the ex-post top 10 features according to Table 4.8 and only include those three features in the training of τ model. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable trailing returns but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are all equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.25	0.56	0.01	0.14	0.09	0.35	-0.22
	[0.99]	[1.29]	[0.04]	[0.55]	[0.38]	[1.13]	[-0.53]
CAPM Alpha	0.29	0.52	0.05	0.14	0.12	0.44	-0.09
	[1.13]	[1.23]	[0.17]	[0.56]	[0.53]	[1.41]	[-0.21]
FF-3-MOM Alpha	0.49***	0.81**	0.23	0.34**	0.28*	0.58**	-0.23
	[3.04]	[2.48]	[1.14]	[1.98]	[1.76]	[2.18]	[-0.60]
FF-5-MOM Alpha	0.39**	0.51	0.16	0.33*	0.25	0.50*	-0.01
	[2.30]	[1.55]	[0.78]	[1.73]	[1.50]	[1.84]	[-0.03]
τ 25%	-0.01	-0.05	-0.01	0.00	0.01	0.03	
τ 50%	0.00	-0.03	-0.01	0.00	0.02	0.03	
τ 75%	0.02	-0.02	0.00	0.01	0.02	0.05	
avg τ	0.00	-0.05	-0.01	0.00	0.02	0.04	
sd τ	0.04	0.05	0.00	0.00	0.00	0.03	
avg $\dot{Z}$	0.00	0.13	0.05	-0.04	-0.04	-0.08	
avg $\dot{Z}$ in 1st quintile	-1.29	-1.15	-1.10	-1.19	-1.23	-1.61	
avg $\dot{Z}$ in 5th quintile	1.07	1.24	0.99	0.91	0.95	1.15	
avg $\dot{Z}$ P5 - P1	2.36	2.39	2.08	2.10	2.18	2.76	
number of stocks	3536.90	710.74	701.05	707.78	695.45	721.88	
avg. BM ratio	0.84	1.21	0.76	0.68	0.68	0.85	
avg. ME	5.08	1.85	3.26	4.70	5.94	9.64	
avg. Monthly Vol	25.27	25.95	18.76	19.05	20.88	41.60	
persistence		0.71	0.58	0.57	0.61	0.74	

Table D.12: **Long-short of long-short test results for value as the treatment using equal weighting with cross-sectionally demeaned variables.** We apply our R-learner procedure to cross-sectionally de-meaned data: all variables have mean 0 during each month. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable book-to-market ratios but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.15	-1.19^{**}	0.54	0.43	0.25	0.67^{**}	1.86^{***}
	[0.43]	[-2.03]	[1.49]	[1.47]	[0.73]	[2.09]	[3.52]
CAPM Alpha	0.14	-1.14^{*}	0.43	0.43	0.37	0.74^{**}	1.88^{***}
	[0.40]	[-1.94]	[1.19]	[1.42]	[1.17]	[2.27]	[3.53]
FF-3-MOM Alpha	0.03	-1.27^{**}	0.37	0.34	0.20	0.64^{**}	1.91^{***}
	[0.11]	[-2.52]	[1.18]	[1.23]	[0.73]	[2.00]	[3.92]
FF-5-MOM Alpha	-0.24	-1.42^{**}	0.16	0.08	0.05	0.51	1.93^{***}
	[-0.75]	[-2.42]	[0.53]	[0.29]	[0.17]	[1.51]	[3.56]
τ 25%	-0.05	-0.09	-0.05	-0.03	-0.01	0.00	
τ 50%	-0.02	-0.07	-0.04	-0.02	-0.01	0.01	
τ 75%	0.00	-0.06	-0.04	-0.02	0.00	0.01	
avg τ	-0.03	-0.08	-0.04	-0.02	-0.01	0.01	
sd τ	0.04	0.04	0.01	0.00	0.00	0.01	
avg $\dot{Z}$	0.01	-1.57	-0.16	0.49	0.76	0.51	
avg $\dot{Z}$ in 1st quintile	-2.81	-3.82	-2.59	-1.88	-1.55	-1.61	
avg $\dot{Z}$ in 5th quintile	2.98	1.08	2.65	2.99	3.12	2.47	
avg $\dot{Z}$ P5 - P1	5.79	4.90	5.24	4.87	4.67	4.08	
number of stocks	3902.94	781.03	780.07	780.57	780.16	781.11	
avg. BM ratio	0.77	1.02	0.76	0.70	0.67	0.68	
avg. ME	4.74	3.47	5.34	5.31	6.08	3.48	
avg. Monthly Vol	24.65	19.12	24.08	25.58	29.52	24.81	
persistence		0.76	0.59	0.57	0.61	0.74	

Table D.13: **Long-short of long-short test results for size as the treatment using equal weighting with cross-sectionally demeaned variables.** We apply our R-learner procedure to cross-sectionally de-meaned data: all variables have mean 0 during each month. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable firm sizes but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.16	-0.35	0.47	0.13	0.44	-0.07	0.28
	[0.29]	[-0.54]	[0.83]	[0.25]	[0.98]	[-0.13]	[0.54]
CAPM Alpha	0.40	-0.01	0.72	0.36	0.60	0.15	0.16
	[0.90]	[-0.02]	[1.54]	[0.87]	[1.52]	[0.28]	[0.31]
FF-3-MOM Alpha	0.14	-0.33	0.43*	0.06	0.39*	-0.06	0.27
	[0.60]	[-0.88]	[1.83]	[0.27]	[1.77]	[-0.14]	[0.54]
FF-5-MOM Alpha	-0.03	-0.52	0.28	0.02	0.30	-0.29	0.23
	[-0.12]	[-1.30]	[1.17]	[0.09]	[1.19]	[-0.61]	[0.40]
τ 25%	0.00	-0.03	0.00	0.01	0.01	0.02	
τ 50%	0.01	-0.02	0.00	0.01	0.02	0.02	
τ 75%	0.02	-0.01	0.00	0.01	0.02	0.03	
avg τ	0.01	-0.02	0.00	0.01	0.02	0.03	
sd τ	0.02	0.03	0.00	0.00	0.00	0.02	
avg $\dot{Z}$	-0.01	-0.04	0.00	0.02	0.00	-0.01	
avg $\dot{Z}$ in 1st quintile	-0.54	-0.51	-0.48	-0.48	-0.52	-0.61	
avg $\dot{Z}$ in 5th quintile	0.66	0.47	0.59	0.67	0.69	0.81	
avg $\dot{Z}$ P5 - P1	1.19	0.98	1.07	1.14	1.21	1.42	
number of stocks	3681.14	754.13	691.68	737.34	731.22	766.76	
avg. BM ratio	0.77	0.94	0.72	0.70	0.68	0.77	
avg. ME	4.96	7.65	6.15	5.63	3.34	2.02	
avg. Monthly Vol	25.54	38.38	28.86	25.88	17.70	16.84	
persistence		0.86	0.75	0.75	0.73	0.80	

Table D.14: **Long-short of long-short test results for momentum as the treatment using equal weighting with cross-sectionally demeaned variables.** We apply our R-learner procedure to cross-sectionally de-meaned data: all variables have mean 0 during each month. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable trailing returns but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	-0.29	-1.03^{**}	-0.21	0.25	-0.07	0.01	1.04^{*}
	[-0.99]	[-2.29]	[-0.77]	[0.90]	[-0.27]	[0.02]	[1.95]
CAPM Alpha	-0.45	-1.18^{***}	-0.27	0.17	-0.15	-0.13	1.04^{*}
	[-1.59]	[-2.72]	[-0.99]	[0.59]	[-0.56]	[-0.26]	[1.86]
FF-3-MOM Alpha	-0.19	-0.84^{***}	-0.08	0.33	0.03	0.18	1.02^{*}
	[-1.20]	[-2.59]	[-0.41]	[1.40]	[0.15]	[0.43]	[1.79]
FF-5-MOM Alpha	-0.04	-1.08^{***}	-0.23	0.40^{*}	-0.08	0.31	1.38^{**}
	[-0.29]	[-3.17]	[-1.05]	[1.79]	[-0.40]	[0.75]	[2.48]
τ 25%	-0.08	-0.22	-0.08	-0.03	0.01	0.07	
τ 50%	-0.02	-0.14	-0.07	-0.02	0.02	0.11	
τ 75%	0.03	-0.11	-0.05	-0.01	0.03	0.18	
avg τ	-0.02	-0.20	-0.07	-0.02	0.02	0.15	
sd τ	0.17	0.19	0.01	0.01	0.01	0.14	
avg $\dot{Z}$	-0.69	-0.63	-0.69	-0.72	-0.71	-0.69	
avg $\dot{Z}$ in 1st quintile	-1.98	-1.94	-1.89	-1.87	-1.89	-2.23	
avg $\dot{Z}$ in 5th quintile	0.38	0.46	0.27	0.25	0.30	0.54	
avg $\dot{Z}$ P5 - P1	2.36	2.40	2.15	2.11	2.19	2.77	
number of stocks	3536.90	734.32	656.14	718.12	708.52	719.81	
avg. BM ratio	0.84	1.17	0.69	0.66	0.68	0.97	
avg. ME	5.08	3.88	5.11	5.38	4.39	6.52	
avg. Monthly Vol	25.27	27.87	23.76	23.60	19.82	31.05	
persistence		0.73	0.62	0.60	0.63	0.73	

Table D.15: **Long-short of long-short test results for value as the treatment using value weighting.** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable book-to-market ratios but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are value weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.42	-0.94	0.06	0.33	0.19	-0.04	0.89
	[1.17]	[-1.62]	[0.19]	[1.31]	[0.76]	[-0.15]	[1.44]
CAPM Alpha	0.47	-1.15^{*}	-0.22	0.23	0.25	0.07	1.22^{*}
	[1.37]	[-1.95]	[-0.67]	[0.86]	[0.97]	[0.29]	[1.95]
FF-3-MOM Alpha	0.34	-1.24^{**}	-0.30	0.11	0.10	-0.09	1.14^{*}
	[1.20]	[-2.23]	[-0.97]	[0.46]	[0.52]	[-0.48]	[1.88]
FF-5-MOM Alpha	0.08	-1.28^{**}	-0.38	0.00	-0.07	-0.09	1.19^{*}
	[0.28]	[-2.05]	[-1.18]	[-0.02]	[-0.40]	[-0.51]	[1.82]
τ 25%	-0.32	-0.79	-0.32	-0.16	-0.05	0.03	
τ 50%	-0.13	-0.60	-0.26	-0.13	-0.03	0.05	
τ 75%	-0.01	-0.47	-0.22	-0.10	-0.01	0.07	
avg τ	-0.21	-0.68	-0.27	-0.13	-0.03	0.06	
sd τ	0.30	0.31	0.06	0.03	0.02	0.04	
avg $\dot{Z}$	0.00	-0.88	-0.27	0.15	0.40	0.61	
avg $\dot{Z}$ in 1st quintile	-1.35	-1.86	-1.30	-0.94	-0.69	-0.51	
avg $\dot{Z}$ in 5th quintile	1.43	0.31	0.89	1.33	1.58	1.79	
avg $\dot{Z}$ P5 - P1	2.78	2.17	2.19	2.26	2.28	2.30	
number of stocks	3902.94	781.07	777.73	780.65	777.88	785.61	
avg. BM ratio	0.77	1.09	0.79	0.69	0.63	0.63	
avg. ME	4.74	0.60	1.81	4.44	6.24	10.51	
avg. Monthly Vol	24.65	17.78	19.96	24.36	27.88	33.03	
persistence		0.71	0.51	0.54	0.56	0.71	

Table D.16: **Long-short of long-short test results for size as the treatment using value weighting.** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable firm sizes but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are value weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	Full	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	$5-1$
Mean	0.37	0.36	0.34	-0.05	0.28	0.59	0.24
	[0.64]	[0.51]	[0.64]	[-0.07]	[0.41]	[1.20]	[0.43]
CAPM Alpha	0.77*	0.72	0.68	0.22	0.58	0.76*	0.04
	[1.73]	[1.24]	[1.44]	[0.41]	[1.02]	[1.65]	[0.08]
FF-3-MOM Alpha	0.43**	0.31	0.46	-0.15	0.14	0.52	0.21
	[2.45]	[0.85]	[1.18]	[-0.53]	[0.43]	[1.41]	[0.42]
FF-5-MOM Alpha	0.23	-0.07	0.31	-0.24	0.19	0.50	0.57
	[1.34]	[-0.18]	[0.92]	[-0.87]	[0.55]	[1.31]	[1.16]
τ 25%	-0.05	-0.17	-0.05	-0.03	-0.01	0.01	
τ 50%	-0.02	-0.11	-0.04	-0.02	0.00	0.02	
τ 75%	0.00	-0.08	-0.04	-0.02	0.00	0.03	
avg τ	-0.04	-0.14	-0.04	-0.02	-0.01	0.02	
sd τ	0.08	0.10	0.01	0.00	0.00	0.03	
avg $\dot{Z}$	-0.01	-0.07	0.03	0.04	0.01	-0.07	
avg $\dot{Z}$ in 1st quintile	-1.05	-1.01	-0.78	-0.89	-1.09	-1.24	
avg $\dot{Z}$ in 5th quintile	1.23	0.83	0.99	1.19	1.49	1.54	
avg $\dot{Z}$ P5 - P1	2.28	1.84	1.78	2.08	2.58	2.78	
number of stocks	3681.14	739.86	706.92	754.83	708.25	771.28	
avg. BM ratio	0.77	0.88	0.70	0.73	0.72	0.80	
avg. ME	4.96	17.16	4.45	1.71	0.96	0.51	
avg. Monthly Vol	25.54	52.31	29.17	18.26	16.42	11.74	
persistence		0.85	0.72	0.66	0.61	0.73	

Table D.17: **Long-short of long-short test results for momentum as the treatment using value weighting.** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. In the top panel, we calculate the average returns and alphas of P5-P1 long-short trading strategies based on treatment variable trailing returns but restricted to each $\hat{\tau}$ quintile. All return and alpha numbers are from P5-P1 strategy and are monthly in percentage. All portfolios in the P5-P1 long-short strategies are all value weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively. The second panel shows cross-sectional summary statistics for $\hat{\tau}$. The third panel presents the spread of treatment variable in the cross section for all $\hat{\tau}$ quintiles and for the full universe. The bottom panel shows some additional information about each quintile and the full universe. We include number of stocks, book-to-market ratios, market equity (ME) in billions of dollars, monthly trading volume in millions of shares, and persistence defined as the average percentage of stocks that will remain in the same $\hat{\tau}$ quintile in the next month.

	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	5-1
training	0.00	0.88**	1.34***	1.86***	1.73***	1.72**
	[0.01]	[2.00]	[2.67]	[3.83]	[2.96]	[2.56]
test oneshot	0.06	0.05	-0.07	-0.21	0.56*	0.50
	[0.17]	[0.17]	[-0.30]	[-0.70]	[1.75]	[1.50]
test rolling	-0.06	-0.16	0.26	0.18	0.71**	0.77**
	[-0.14]	[-0.65]	[1.15]	[0.77]	[2.15]	[2.17]

Table D.18: **Long-short of long-short test results for value as the treatment: training vs. test set.** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. For the first two rows "training", we calculate the average returns of P5-P1 long-short trading strategies restricted to each $\hat{\tau}$ quintile, for the period of the first rolling window for training (199601-200512). The $\hat{\tau}$ quintiles used in forming long-short of long-short strategies are determined by one τ model fitted using data in the same time window (199601-200512). For the third and fourth rows, "test oneshot", we repeat the same calculation for our entire out-of-sample test period 200601-201812 and $\hat{\tau}$ quintiles are determined by one single τ model trained from only the first rolling window 199601-200512. For the last two rows, "test rolling", we repeat the calculations using the same method as in our main results: the average returns of long-short of long-short strategies are calculated for our out-of-sample test period 200601-201812 with annual refittings using the previous 10 years as training data. For each year in the test period, $\hat{\tau}$ quintiles are determined by a τ model trained using the 10-year rolling window before the current year. All return numbers are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively.

	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	5-1
train	-4.99^{***}	-1.43^{**}	-0.09	-0.03	-0.63	4.35^{***}
	[-5.23]	[-2.55]	[-0.18]	[-0.05]	[-1.06]	[5.80]
test oneshot	-0.90	-0.11	-0.12	0.05	0.02	0.93
	[-1.43]	[-0.32]	[-0.44]	[0.17]	[0.08]	[1.62]
test rolling	-1.43^{**}	-0.04	0.31	0.12	0.11	1.54^{***}
	[-2.48]	[-0.12]	[1.29]	[0.55]	[0.45]	[2.60]

Table D.19: **Long-short of long-short test results for size as the treatment: training vs. test set.** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. For the first two rows "training", we calculate the average returns of P5-P1 long-short trading strategies restricted to each $\hat{\tau}$ quintile, for the period of the first rolling window for training (199601-200512). The $\hat{\tau}$ quintiles used in forming long-short of long-short strategies are determined by one τ model fitted using data in the same time window (199601-200512). For the third and fourth rows, "test oneshot", we repeat the same calculation for our entire out-of-sample test period 200601-201812 and $\hat{\tau}$ quintiles are determined by one single τ model trained from only the first rolling window 199601-200512. For the last two rows, "test rolling", we repeat the calculations using the same method as in our main results: the average returns of long-short of long-short strategies are calculated for our out-of-sample test period 200601-201812 with annual refittings using the previous 10 years as training data. For each year in the test period, $\hat{\tau}$ quintiles are determined by a τ model trained using the 10-year rolling window before the current year. All return numbers are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively.

	$\hat{\tau}$ 1	$\hat{\tau}$ 2	$\hat{\tau}$ 3	$\hat{\tau}$ 4	$\hat{\tau}$ 5	5-1
train	-3.31^{***}	0.48	1.50^{**}	1.45^{**}	1.84^{**}	5.15^{***}
	[-3.34]	[0.81]	[2.43]	[1.99]	[2.09]	[7.13]
test oneshot	-0.52	0.31	0.60	0.01	0.18	0.69
	[-0.75]	[0.60]	[1.17]	[0.01]	[0.36]	[1.63]
test rolling	-0.61	-0.08	0.03	0.36	0.57	1.18^{***}
	[-0.87]	[-0.15]	[0.06]	[0.64]	[1.08]	[2.72]

Table D.20: **Long-short of long-short test results for momentum as the treatment: training vs. test set.** We apply our R-learner procedure to cross-sectionally standardized data: all variables have mean 0 and standard deviation 1 during each month. For the first two rows "training", we calculate the average returns of P5-P1 long-short trading strategies restricted to each $\hat{\tau}$ quintile, for the period of the first rolling window for training (199601-200512). The $\hat{\tau}$ quintiles used in forming long-short of long-short strategies are determined by one τ model fitted using data in the same time window (199601-200512). For the third and fourth rows, "test oneshot", we repeat the same calculation for our entire out-of-sample test period 200601-201812 and $\hat{\tau}$ quintiles are determined by one single τ model trained from only the first rolling window 199601-200512. For the last two rows, "test rolling", we repeat the calculations using the same method as in our main results: the average returns of long-short of long-short strategies are calculated for our out-of-sample test period 200601-201812 with annual refittings using the previous 10 years as training data. For each year in the test period, $\hat{\tau}$ quintiles are determined by a τ model trained using the 10-year rolling window before the current year. All return numbers are monthly in percentage. All portfolios in the P5-P1 long-short strategies are equal weighted. t-stats with NW standard errors are inside square brackets. One, two and three stars represent 10%, 5%, and 1% significance level respectively.