Title
Consumer Search in the U.S. Auto Industry

Author
Yavorsky, Daniel Ryan

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Consumer Search in the U.S. Auto Industry

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Management

by

Daniel Ryan Yavorsky

2020

ABSTRACT OF THE DISSERTATION

Consumer Search in the U.S. Auto Industry

by

Daniel Ryan Yavorsky
Doctor of Philosophy in Management
University of California, Los Angeles, 2020
Professor Elisabeth Honka, Co-Chair
Professor Peter E. Rossi, Co-Chair

Consumers may become aware of products and their characteristics either passively or actively. While passive awareness may arise through conversations with friends or consumption of advertising, active awareness occurs through a process of canvassing sellers. Consumers often refer to this activity as browsing or shopping; economists refer to this activity as search.

This dissertation begins with a summary of the current economic models of consumer search and the econometric methods used to estimate those models. An emphasis will be placed on models that assume consumers search sequentially for product fit, as opposed to non-sequential search or search for observable product characteristics such as price. The bulk of this dissertation then provides an empirical analysis of consumer search through which I demonstrate identification of – and provide a first estimation of – a key parameter in sequential search models, the standard deviation of product fit. I devote separate chapters to detailed explanations of the data acquired for the analysis; the model, its estimation, and identification of this key parameter; and important post-estimation analysis including an assessment of model fit, calculations of consumer surplus and price elasticities,

and implementation of counterfactual analyses.

The setting of my empirical analysis is the U.S. automotive market. I assemble a unique data set containing individual-level smartphone geolocation data that inform me about dealership visits which I combine with proprietary DMV registration data that inform me of new vehicle purchases. I model consumers' dealership visits and purchase decisions using a discrete choice model of demand with optimal sequential search for product fit. In these models, the benefit of searching is measured by the standard deviation of the product fit. This benefit is parametrically identified by the functional form of the match-value, but in practice it is difficult to jointly estimate along with consumer search costs. I use the distance a consumer must travel to visit a dealership as an exogenous search cost shifter, which enables me to identify and estimate the standard deviation of the product fit. My results show that the benefit provided by dealerships to consumers is substantial and that failure to estimate the standard deviation of the product fit leads to biased estimates of search costs and consumer surplus, as well as to inaccurate predictions regarding the number of searches that consumers conduct. These effects can impact managerial decisions, as demonstrated through my counterfactual analyses.

The dissertation of Daniel Ryan Yavorsky is approved.

Brett William Hollenbeck

Melvin Keith Chen

Peter E. Rossi, Committee Co-Chair

Elisabeth Honka, Committee Co-Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am extremely grateful for my family, my academic advisors, and the colleagues and friends I have developed while at UCLA.

In particular, I thank my wife Alison for her monumental effort and support throughout the doctoral program. She listened when I could only think out loud, she offered rational advice when my instinct was to behave irrationally, she lent a non-institutional perspective when I was nearsighted, and she helped me refocus when I was distracted. My progress toward this degree would have simultaneously been much faster and much slower without her, and for both I am thankful. I am also thankful for my mother Lorie, brothers Nicholas and Matthew, sisters in-law Alexandra and Juliana, parents in-law Ron and Su, and extended family and friends too numerous to enumerate, who reminded me that one's work is not the only important aspect of life and who shared their time, opened their homes, and invited me on their adventures to ensure not every hour of my week was spent at a computer.

I am also deeply indebted to the professors and staff at UCLA, none more so than my academic advisors. Elisabeth Honka demonstrated unwavering commitment to my success and provided guidance on a near-daily basis. Peter Rossi helped me develop professionally and dedicated much time and effort to learning about my research in order to provide invaluable feedback; his knowledge and passion for research is truly inspiring. Keith Chen made many of my research projects possible both by providing access to data and through his enthusiastic support. Brett Hollenbeck provided excellent feedback and suggestions, and was always eager to receive updates on my work. I am incredibly fortunate to have worked with my advisors.

Other UCLA faculty and staff have graciously offered career advice, opened their homes for social gatherings, or assisted with funding opportunities; in short, they have made UCLA a cohesive community that has profoundly impacted my life. For this I thank Anand Bodapati, Randy Bucklin, Aimee Drolet, Dominique Hanssens, Chad Hazlett, Hal Hershfield,

Sylvia Hristakeva, Cassie Mogilner Holmes, Stephan Seiler, Franklin Shaddy, Suzanne Shu, Sanjay Sood, Stephen Spiller, Andres Terech, Robert Zeithammer, and Shi Zhang, as well as Craig Jessen, Kristin Christian, Lydia Heyman, April Barfield, Christian Yirgu, Ameyalli Martinez, and Crystal Hwang.

I want to also acknowledge my fellow doctoral students with whom I shared this journey: most significantly Geoff Zheng, Marco Testoni, Darren Aiello, and Taylor Corcoran; my amazing cohort that includes Prashant Chintapalli, Ali Fattahi, Elicia John, Araz Khobadakhshian, Linda Nguyen, Jieun Pai, Bruno Pellegrino, Anna Saez de Tejada Cuenca, and Michael Tang; and the fantastic students who have come before or will follow shortly behind, which includes Keunwoo Kim, Wayne Taylor, Charlene Chu, Li Jiang, Jonathan Lim, Marissa Sharif, Mirei Takashima, Christian Blanco, Mahyar Kargar, Chady Gemayel, Nimesh Patel, Alex Fabisiak, Ashley Angulo, Bennett Chiles, Kalyan Rallabandi, Kate Christensen, Bobby Nyotta, Kira Stearns, Julia Levine, Sherry He, Jon Bogard, Joseph Reiff, David Zimmerman, Ipek Demirdag, David Dolifka, Daniel Mirny, Malena de la Fuente, Yilin Zhuo, and Mahsa Paridar. You are awesome.

I would also like to thank the Morrison Family Center for Marketing Studies and Data Analytics for generous financial support as well as Safegraph, the Texas Department of Motor Vehicles, and VinAudit for access to data. In addition, Jae Hyen Chung, Sergei Koulayev, Raluca Ursu, participants at the 2018 University of Houston Doctoral Consortium and 2019 American Marketing Association Summer Meetings, as well as quite a few of the aforementioned UCLA students and faculty have provided helpful or supportive comments and suggestions; thank you all.

A modified version of the research presented in this dissertation is in preparation for publication. The modified manuscript submitted for publication will be co-authored with Elisabeth Honka and Keith Chen.

VITA

2002–2006	B.A. Economics and Mathematics, Claremont McKenna College
2006–2014	Consultant, Cornerstone Research
2011–2014	M.B.A. Management, UCLA Anderson School of Management
2012–2014	Charterholder, CFA Institute
2014–2018	Anderson Fellowship, UCLA Anderson School of Management
2015–2020	Teaching and Research Assistant, UCLA Anderson School of Management

CHAPTER 1

Academic Literature

1.1 Structural Models of Consumer Search

Early economic models assume consumers have full information. That is, all consumers are perfectly informed about all attributes of all products at all times. While convenient for modeling basic features of a market, such an assumption is clearly unrealistic. Consumers may have some information about some products, and if they wish to know more, they must expend effort to gather additional information. Understanding how consumers search for products and eventually chose a product to purchase is the subject of a growing literature in economics and marketing.

This literature has roots in empirical observations about firms and consumers that contradict the full information assumption. In economics, consumer search dates back to the classic work of Stigler (1961) who reported the presence and persistence of price dispersion of homogeneous goods. As emphasized by the "paradox" presented in Diamond (1971), firms will not rationally set different prices of a homogeneous good if they face demand from fully-informed consumers because, in such a case, all consumers would purchase from the low-priced firm. One explanation, therefore, is that consumers are not fully informed and must search to gain additional information. In marketing, consumer search was first explored in the context of analyzing consumers' consideration sets. The pioneering work of Hauser and Wernerfelt (1990) and Roberts and Lattin (1991) reported that consumers consider only a small fraction of available alternatives, indicating that it may be too costly for consumers

to become fully informed.

The observed price-setting behavior of firms and the limited consideration set sizes reported by consumers motivated the development of models that could rationalize both price dispersion and inexhaustive search.[1] Among the structural econometric models based on those economic models of consumer search, a delineation is often made over what the consumer is searching for and the method of search. The following two sections address each in turn. Other important considerations for these models involve product differentiation (i.e., are products homogeneous, or vertically and/or horizontally differentiated?) and the unit of analysis (i.e., do the data capture individual-level behavior or market-wide relationships?). These considerations will be addressed in my review of the estimation approaches and empirical applications of search models in Section 1.2.2. I close this chapter with a review of recent extensions to the standard search models in Section 1.3 and a summary of the application of these models to the auto industry in Section 1.4. See Honka, Hortacsu, and Wildenbeest (2019) for an additional contemporary review of structural search model in economics and marketing.

1.1.1 What the Consumer Searches For

Structural econometric models of consumer search follow the discrete choice random utility framework popular in demand estimation (see, e.g., Train (2009)), where the consumer chooses to consume the (known or searched) product offering her the highest utility. However, unlike the classic discrete choice model where the consumer is assumed to have full information, in a structural econometric search model the consumer must search to resolve her uncertainty about the product (and commonly chooses only from among those products searched). From the perspective of the consumer, it matters little what product feature (or set of features) has uncertainty and is to be resolved through search. However, the distinc-

[1] Baye et al. (2006), Ratchford et al. (2008), Armstrong (2017), and Anderson, Renault et al. (2018) provide comprehensive surveys of the theoretical research on consumer search.

tion is important to the researcher, who either observes or does not observe the feature(s) over which the consumer searches.

The classic example of search for an observable (to the researcher) feature is a homogeneous product market where the consumer must search to discover the price of the product at each retailer, as in Stigler (1961) who considers commodity product markets and McCall (1970) who models job offers for unemployed workers as a function of only the wage. A recent example that analyzes search over a set of product features can be found in Chen and Yao (2017), who assess consumer search for hotels on an online booking platform and model the click-through from the page of search results to a specific product listing as the revelation of a large set of hotel characteristics, including price, star rating, and distance to the city center, among others.

In contrast, a consumer may search over one or more product features unobserved by the researcher. Such a situation is referred to as the consumer's search for "product fit" or "match value" following the concept introduced in Wolinsky (1986). This framework is most appropriate when modeling consumer search for horizontally differentiated products or products with components that are difficult to measure and (possibly) differ across purchase occasions. For example, Dong et al. (2020) model consumer search for cosmetic moisturizers, the purchase of which may depend on the dryness, sensitivity, and texture of the consumer's skin at the time of purchase – all important considerations when choosing a moisturizer, but very difficult to capture in data. As a second example, in the subsequent chapters of this dissertation, I model consumer search for a new car, the choice of which may depend on how much the consumer likes the look of the vehicle, her comfort in it, or how courteous and helpful the dealership staff is during her dealership visit.

The distinction among whether the searched features are observable to the researcher significantly impacts the formulation and estimation of the model. In addition, and importantly, observing the searched features also enables the researcher to distinguish between the two most commonly assumed methods of search (simultaneous and sequential, discussed

next). For example, De los Santos, Hortacsu, and Wildenbeest (2012) uses price data from online book retailers to determine that consumers follow the simultaneous search method; Honka and Chintagunta (2017) find a similar result when analyzing price search for auto insurance. The results from both studies rely on the pattern and search-length of observed price realizations relative to the market-wide distribution of prices, an approach unavailable with an unobserved product characteristic, such as match value.

1.1.2 How the Consumer Searches

As just alluded to, structural econometric models of consumer search often assume one of two search methods: a sequential search method or a non-sequential search method. The latter is commonly referred to as a fixed-sample size or simultaneous[2] search method. Under the latter method, consumers commit to searching a fixed set of products before they begin their search process, whereas under a sequential search method, the consumer decides after each successive search occasion whether continue searching or to stop searching and purchase a product.[3] Often the empirical setting dictates the assumed method of search. For example, simultaneous search fits the situation of high school students applying to college because (typically) all college applications are submitted prior to the student learning her admission decisions; sequential search better represents a hungry consumer at a food court who, upon observing the offerings of one dining establishment may elect to view the next or to stop and place her lunch order.

Each method offers its own estimation challenges. Under simultaneous search, the consumer identifies the subset of products that maximizes his expected maximum utility. This

[2]Note that simultaneous search does not mean that all products are sampled simultaneously, but rather that the consumer commits to sample all products prior to making her purchase decision.

[3]Hybrid theoretical models of sequential search occasions where multiple products may be simultaneous searched on each occasion have been presented. Morgan and Manning (1985) show that this combination dominates either pure simultaneous or pure sequential search. However, I am not aware of any empirical research that has discovered how to take their hybrid search model to data.

involves enumeration of all combinatorically possible subsets of products to search, the complexity of which grows exponentially with the number of available products. Specifically, for J available products, the number of search sets is 2^J. Researchers have circumvented this curse of dimensionality by assessing markets with relatively few products (e.g., Mehta, Rajiv, and Srinivasan (2003)) or making assumptions about first-order or second-order stochastic dominance among the utility distributions, as in Vishwanath (1992) and Chade and Smith (2006). The latter approach reduces the complexity of the calculation to one that grows linearly with the number of potentially searchable products (the number of search sets is J), making the computations feasible for a large number of products but at the cost of interpretability. Honka (2014) demonstrates that the same reduction in complexity is achievable by assuming that the means or variances of the utility distribution are identical, which better connects the necessary mathematical assumptions used to circumvent the curse of dimensionality to plausible aspects of the price distribution in the product market. In addition, it is worth noting that no study has yet empirically analyzed simultaneous search for product fit and there may be challenges to such an implementation that are yet to be discovered.

Under sequential search, the consumer faces a dynamic utility optimization problem. At each step in the search sequence, the consumer must decide whether to stop searching and purchase among the searched products, or to continue searching acting optimally in the subsequent period. This problem was famously solved and simplified by Weitzman (1979), who shows that each product can be assigned an index value (commonly referred to as a reservation utility value) that facilitates optimal behavior via three decision rules: the consumer continues searches in decreasing order of reservation utility values (the Selection Rule) until a realized utility value exceeds the best outstanding reservation utility value (the Stopping Rule), at which time the consumer stops searching and purchases the product with the best realized utility value (the Choice Rule). As a result of Weitzman's rules, the computation complexity of sequential search models is substantially reduced. However, the researcher must still calculate reservation utility values and fit the observed search and

purchase behavior of consumers to Weitzman's rules, as discussed in the next section.

It is worth noting that, for either method, the computational complexity is impacted by the informativeness of the data. For example, in Honka and Chintagunta (2017) where the authors estimate a sequential price-search model, the data reveal only which products were searched, but not the order of search. This yields another combinatorically challenging problem of order $K!$ (where K is the number of observed searches by the consumer), albeit generally milder than initial problem of determining search set inclusion encountered in the simultaneous search model of order 2^J, but daunting nonetheless for any observed search sequences of substantial length. By contrast, when the researcher observes the order of search in a model characterized by Weitzman's rules, there are only J possible search sets because the order of search is known. Moreover, once the number of searches is determined, there is only the one realized search path and thus no curse of dimensionality.

1.2 Sequential Search for Product Fit

Sequential search is the modeling approach taken in subsequent chapters of this dissertation and thus will be the focus of the majority of the remaining literature review in this chapter. Emphasis is placed on estimation approaches rather than specific applications or empirical results.

1.2.1 Model Specification

To fix ideas, I first recount the theoretical model of Weitzman (1979) and subsequently expand the framework to the econometric specification of Kim, Albuquerque, and Bronnenberg (2010).

Weitzman (1979) supposes there are $j = 1, \ldots, J$ available products with utility u_j drawn from probability distributions $F_j(u_j)$ along with a "fallback" reward of $u_0 = 0$ if no alternative is selected. Discovering product j's utility costs c_j. The consumer's goal is to maximize

the expected net utility, which is the product's gross utility less the sum of the search costs incurred. As indicated above, optimal consumer behavior follows the three Weitzman rules, which depend on an index value known as a reservation utility. The reservation utility (z_j) is implicitly defined as the value that equates the marginal benefit and marginal cost of search:

$$\underbrace{\int_{z_j}^{\infty} u\, f_j(u)\, du}_{\text{Marginal Benefit}} \quad = \quad \underbrace{c_j}_{\text{Marg. Cost}} \tag{1.1}$$

The result from Weitzman (1979) is powerful because it facilitates analysis of differentiated products (and simplifies to handle homogenous products as well). For example, each product may be differentiated in its before-search observable value by specifying utility as

$$u_j = \delta_j + \varepsilon_j \tag{1.2}$$

with δ_j known to the consumer or it may be differentiated by a product-specific search cost c_j. In addition or alternatively, each realized value may result from a product-specific distribution $f_j(u_j)$, where greater mass in the upper tail of $f_j(u_j)$ results in higher reservation values z_j (holding everything else constant).

Kim, Albuquerque, and Bronnenberg (2010) take the Weitzman (1979) framework to data, allowing product utilities and search costs to be individual specific (i.e., u_{ij} and c_{ij}). They specify the component of utility known prior to search (δ_{ij}) as a combination of product features ($\boldsymbol{x}_j$) and individual-specific preference parameters ($\boldsymbol{\beta}_i$):

$$u_{ij} = \boldsymbol{x}_j' \boldsymbol{\beta}_i + \varepsilon_{ij} \tag{1.3}$$

Search costs are modeled as a lognormally distributed random effect, with $\mathbf{w_j}$ a vector of cost attributes and $\boldsymbol{\gamma}_i$ a vector of individual-specific search-cost sensitivity parameters: $c_{ij} = \mathbf{w_j}'\boldsymbol{\gamma}_i$.

In addition, the authors assume that match-value realizations for all consumers and all products are drawn from independent and identically distributed normal distributions ($\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$) where they fix $\sigma_{ij} = 1$ (in the subsequent chapters of this dissertation, I show that this is not necessary, particularly when the researcher has access to an exogenous cost-shifting variable). Kim, Albuquerque, and Bronnenberg (2010) show how to calculate reservation utilities under the assumption of normally distributed match-values. They first equate expected marginal benefit to marginal cost following Equation (1.1), which can be integrated given the normal distributional assumption to yield

$$\frac{c_{ij}}{\sigma_{ij}} = \phi(\zeta_{ij}) + \zeta_{ij}\left(1 - \Phi(\zeta_{ij})\right) \tag{1.4}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the normal probability density function and cumulative distribution function, respectively. And second, they implicitly solve Equation (1.4) for ζ_{ij} and use ζ_{ij} to recover reservation utilities through the relationship

$$z_{ij} = \delta_{ij} + \zeta_{ij}\sigma_{ij}. \tag{1.5}$$

The framework provided by Weitzman (1979) and Kim, Albuquerque, and Bronnenberg (2010) underlies contemporary research on the estimation and application of structural econometric models of sequential search.

1.2.2 Methods of Estimation and Empirical Applications

Unique data sets and advances in estimation approaches have resulted in numerous methods by which to fit variants of the model presented in Kim, Albuquerque, and Bronnenberg (2010). In this section, I review many of them.

Crude Frequency MSLE

Chen and Yao (2017) analyze click-stream data of online hotel bookings provided by a travel website. In their sequential search model, consumers search for attribute levels (not match values) by using refinement tools on the website and by clicking through to product-specific webpages. To make the multi-attribute search model tractable, consumer are assumed to know the joint multivariate-distribution of the combination of the website's refinement methods and the hotels' attributes. With the exception of what the consumer is searching for, the model of Chen and Yao (2017) closely follows the Kim, Albuquerque, and Bronnenberg (2010) framework presented above.

To estimate the model, the authors construct the likelihood function, decomposing the joint probability of search and purchase into the marginal probability of search and the conditional probability of purchase given search. Let j^* denote the purchased product and S_{ij} denote the set of searched alternatives by consumer i prior to searching product j, then the likelihood is

$$\begin{aligned} L &= \prod_{i=1}^{N} \Pr_i(\text{purchase}|\text{search}) \times \Pr_i(\text{search}) \\ &= \prod_{i=1}^{N} \Big(\Pr_i(u_{ij^*} \geq u_{ij},\, \forall j \in S_i) \Big) \times \Big(\Pr_i(z_{ij} \geq u_{ir},\, \forall r \in S_{ij}) \times \Pr(z_{ij} \geq z_{it},\, \forall t \notin S_{ij}) \Big) \end{aligned} \tag{1.6}$$

In words, the likelihood function is the mathematical representation of Weitzman's rules. The first term in the second line is the probability that the chosen alternative has the maximum utility among all searched alternatives (Weitzman's Choice Rule). The second term is the probability on each search occasion that the searched alternative's reservation utility exceeds the realized utility values of all products searched before it (Weitzman's Continuation/Stopping Rule). And the third term is the probability that the reservation utility of the product searched on each search occasion exceeds the reservation utilities of all

other unsearched products (Weitzman's Selection Rule).

The authors proceed with simulated maximum likelihood estimation. The simulation is carried out by the crude frequency approach. Specifically, for each consumer, the authors draw a 100 length-P vectors of values for the P random coefficient parameters (β_i and γ_i). For each of those 100 vectors, the authors draw 50 length-K vectors of values for the unobserved preference component ε_i (recall $|S_i| = K_i$). Because, from the perspective of the researcher, the set of searched alternatives is known, each vector of ε_i is drawn from a truncated normal distribution where truncation depends on the values of the reservation utilities for the search products. With the sets of draws in hand for one consumer, the individual likelihood is simply the percentage of time that the Choice and Selection Rules hold.[4]

While results from the author's simulation study demonstrate effective parameter recovery using the crude frequency simulation approach, it is worth noting that, in general, the parameter space of most sequential search models is high-dimensional such that very few draws may satisfy the conditions set forth in the Weitzman Rules, and thus an extremely large (perhaps computationally infeasible) number of draws may be required to obtain accurate parameter estimates.

Chen and Yao (2017) use exclusions restrictions to separately estimate utility and search cost parameters (an approach I follow in later chapters of this dissertation). In their approach, indirect utility is a function of hotel characteristics and price while search costs are a function of the consumer's time constraint (measured by days until check-in) and hotel "slot" position on the website's page of search results. Chen and Yao (2017) find that consumer ratings a large impact on consumer utility and that search costs are significant. They also find that consumers demonstrate considerable variation in preference and search cost,

[4]As footnoted in Chen and Yao (2017) and expanded upon in Kim, Albuquerque, and Bronnenberg (2017), the truncation in the probability of purchase given search implies that the Stopping Rule holds and thus does not need to be calculated separately.

as measured by high dispersion in the distributions of individual-level coefficients, and that the website's refinement tools (e.g., filtering or ordering search results) make the market less concentrated because they help heterogeneous consumers find hotels that match their preferences better; matches that would be too costly without refinement tools. They use their fitted model to assess the effect of the website's refinement tools on consumer welfare and show that refinement tools lead to consumers making more searches and achieving higher gross utility, but that overall consumer welfare (net utility, or gross utility net of search costs) may decrease as consumers disproportionately make excessive searches using the refinement tools.

Logit-Smoothed Accept/Reject MSLE

Consumer search for online hotel bookings are also investigated in Ursu (2018). The author assesses how rankings affect search when the rankings are endogenous (consumers pay more attention to highly ranked products, and the booking platform ranks the most relevant products at the top). To handle the endogeneity, she utilizes a field experiment run by the booking platform that exogenously varies hotel rankings on the website.

Ursu (2018) follows the sequential search for product fit modeling framework of Kim, Albuquerque, and Bronnenberg (2010). Unlike Chen and Yao (2017), however, Ursu (2018) employs a logit-smoother accept/reject simulator to estimate the model via maximum simulated likelihood. She specifies the individual likelihood as the joint probability that Weitzman's Selection, Stopping, and Choice Rules hold:

$$\begin{aligned} L_i &= \Pr\left(u_{ij^*} \geq u_{ij},\, \forall j \in S_i \bigcap z_{ij} \geq u_{ir},\, \forall r \in S_{ij} \bigcap z_{ij} \geq z_{it},\, \forall t \notin S_{ij}\right) \\ &= \int I(\text{conditions})\phi(\varepsilon)\, d\varepsilon \end{aligned} \tag{1.7}$$

Estimation proceeds by re-writing each of the three conditions as an expression that is greater than zero when the condition is met (e.g., the Choice Rule is expressed as $\nu_1 =$

$u_{ij^*} - \max_{j \in S} u_{ij} \geq 0$) and feeding the conditions into the logit formula:

$$L_i^q = \frac{1}{1 + \sum_{l=1}^{3} e^{-} \nu_{il}^q \lambda} \tag{1.8}$$

where q indexes the random draws and λ is a scaling parameter that controls the "smoothness" of the logit transformation. High λ values shape the logit function as a stair-step and thus have limited smoothing while low λ yield a gentle slope with substantial smoothing.

The simulated likelihood for one consumer is then the average of the logit-smoothed value across draws: $L_i = Q^{-1} \sum_{q=1}^{Q} L_i^q$.

Smoothing in the simulated maximum likelihood calculation facilitates optimization by a gradient-based optimizer that would otherwise struggle or fail when faced with the discrete jumps in likelihood values that result from small changes in parameters. These jumps occur at the threshold when parameter values result in a reordering of reservation utilities or different choices (j^*) among searched alternatives.

Ursu (2018) uses the logit-smoothed accept/reject simulated maximum likelihood method to estimate her model with exogenous ranking variation. She finds a lower position effect (i.e., that it is less costly for a consumer to look further down the list of Expedia's search results) than most other hotel industry literature. She uses her model to assess the effect of changing the ranking scheme on the website to a utility-based ranking. Under this counterfactual, consumers search less yet are better-matched with hotels that align with their preferences, leading to large improvements in consumer welfare. This finding demonstrates the large potential positive effect on consumer welfare achievable through better product ranking, and emphasizes the usefulness of random variation in product ranking in order for intermediaries to learn about their consumers and improve their product offering.

GHK Simulation Used in MSLE

Jiang et al. (2019) use a structural sequential search model to study retargeting ad-

vertising effectiveness. While their framework builds upon the modeling approach of Kim, Albuquerque, and Bronnenberg (2010), their estimation approach cleverly relies on the GHK importance sampling method of Geweke, Keane, and Runkle (1994). Specifically, the authors estimate their sequential search model by first iteratively drawing values of β_i from truncated normal distributions, where truncation results from Weitzman's Selection Rule (i.e., $z_{ij} \geq z_{i,j+1}$ implies that $\delta_{i,j} \geq \delta_{i,j+1}$, thereby imposing restrictions on plausible values of β_i in order to match the observed search sequence in the data). Having satisfied the Search Rule, the authors then iteratively draw values of ε_{ij} to satisfy the Continuation and Choice Rules.

As with the logit-smoothed accept/reject simulation approach used in Ursu (2018), the simulated likelihood function using the GHK method proposed by Jiang et al. (2019) is smooth and differentiable, and thus amenable to optimization with gradient-based approaches. In addition, the GHK method can obtain an approximation of the likelihood function with high accuracy without requiring a large number of draws, even when the dimensionality of integration is high.[5] This is a substantial benefit and will likely lead to wide adoption of this method in future research employing a maximum simulated likelihood estimation approach.

The empirical setting for Jiang et al. (2019) is the Chinese consumer-to-consumer retail platform Taobao, which did not employ retargeting advertising during the study period. The authors assess two potential retargeting marketing strategies: coupons and seller recommendations. Their counterfactual results show coupons to be less effective than seller recommendations, but that both strategies are effective at increasing sales from consumers that appear to have abandoned the focal product. They then show how an auction pricing mechanism for retargeting advertising could be used by Taobao to redistribute profits from retargeted consumers between sellers on the platform and the platform host.

[5]Chung, Chintagunta, and Misra (2019) compare the accuracy of the crude frequency, logit-smoothed, and GHK methods of estimating the sequential search model and find that the GHK method performs best.

Sequential Search with Aggregate Data

Kim, Albuquerque, and Bronnenberg (2010) study online search for digital camcorders at Amazon.com using only view-rank data for estimation. For any focal product, the data provide all other camcorder products ranked by frequency of same-session viewing with the focal product. Thus, although the data are aggregate, they provide a complex pattern of relationships between products.

To estimate the model, the authors use non-linear least squares on a large set of rank inequalities. In order to do so, the authors first compute the "inclusions probability" for a given product j. Product j will be in the consumer's searched set S_i if the first $j-1$ draws of ε_{ij} result in $u_{ik} < z_{ij}$ for $k < j$. Denote this inclusion probability of such an occurrence as π_{ij}. Now, given the sequential nature of search according to Weitzman's Selection Rule, the probability that products j and $j+l$ occur together in a searched set is simply the probability that later-to-be-searched product $j+l$ is in the set; for an arbitrarily ranked product k, the inclusion probability of both j and k is $\pi_{i,jk} = \min\{\pi_{ij}, \pi_{ik}\}$.

The authors then create the commonality index CI_{jk} for the two products j and k as the ratio

$$CI_{jk} = \frac{n_{jk}}{\sqrt{n_j}\sqrt{n_k}} \tag{1.9}$$

where n_{ij} is the number of consumers who searched both products j and k, calculated as $n_{ij} = \sum_{i=1}^{N} \pi_{i,jk}$ (and analogously for n_j and n_k). The functional form of the index is designed to match the view-rank algorithm employed by Amazon.com.

Finally, the authors create an indicator variable $I_{j,kl}$ if the estimated commonality index ordering $CI_{jk} > CI_{jl}$ matches the observed pairwise ordering in the view rank data $(j,k) > (j,l)$. Estimation then proceeds by non-linear least squares, where the objective function to minimize

$$\sum_{(j,k,l)\in S} [\Pr(I_{j,kl} = 1) - 1]^2 \tag{1.10}$$

Kim, Albuquerque, and Bronnenberg (2010) demonstrates that aggregate view-rank data, in combination with a structural sequential model of search, are sufficient to recover utility preference parameters and search cost parameters.

The authors use the fitted model to assess the effects Amazon.com's product references. Providing easy access to products by means of product references selectively lowers search costs for some products, but not all. Doing so may increase consumer surplus if the product references facilitate the finding of preferred products at a lower cost, but it may also lower consumer surplus if search costs are lowered on the wrong products or if lower search costs result in disproportionately more search. Kim, Albuquerque, and Bronnenberg (2010) find that 99% of simulated consumers benefit from the Amazon.com product references, driven mostly by a decrease in average search costs while product choice remained mostly unchanged. In particular, there are popular products that are chosen frequently if searched, and the Amazon.com product references help consumers find those products at lower cost.

Gibbs Sampler, A Bayesian Approach

Morozov (2019) studies the computer hard drive market, building on the framework of Kim, Albuquerque, and Bronnenberg (2010) detailed above. The goal of the study is to assess the adoption of a recent product introduction, the solid state drive (SSD). To do so, the author augments the model to include awareness. Each consumer may be aware (or not) of the existence of SSD hard drives. Lacking awareness of the SSD products, a consumer would never search and thus never purchase an SSD product.

All empirical applications of the sequential search model discussed thus far have included two components of utility unobserved to the researcher: random coefficients in utility preferences (β_i) and match value (ε_{ij}). Morozov (2019) adds a third component, an additional term of pre-search unobserved preference heterogeneity (η_{ij}) such that his utility specification is

$$u_{ij} = x_j\beta_i + \eta_{ij} + \varepsilon_{ij} \tag{1.11}$$

The author then develops a Gibbs sampler using Markov Chain Monte Carlo (MCMC) simulation methods. This approach removes the optimization step where the likelihood function is maximized with respect to model parameters and replaces it with an algorithm that generates draws from the posterior distribution of parameters given the observed search and purchase decisions of consumers in his data.

The benefit of this Bayesian estimation approach is that it scales easily as the number of products and dimension of the parameter space grow. In addition, maximum likelihood approaches require approximations in order to integrate-out consumer heterogeneity, which is computationally burdensome. And most importantly for search models over large product assortments, simulating the joint search sequence and purchase outcome is a low probability event that would require, in a maximum likelihood approach, a large number of simulation draws to precisely approximate the likelihood. The Bayesian approach succumbs to none of these. However, as shown in Morozov (2019), the Gibbs sampler converges relatively slowly, so that proper mixing of the Markov Chain is sometimes achieved only after several thousand iterations. This autocorrelation among the posterior draws is an issue to be closely assessed when implementing a Bayesian approach, such as the Gibbs sampling method outlined in Morozov (2019).

Morozov (2019) uses the estimated model to explain the gap between market shares of solid state drives and traditional hard disk drives, decomposing the difference into effects of consumer preferences and those of search frictions. He calculates that increasing the storage capacity of SSDs to match HDDs increases SDD market share by 14%, decreasing the price of SSDs to match HDDs increases SDD market share by 13%, and removing search frictions by decreasing search costs increases SDD market shares by 32%. In contrast, increasing SDD awareness yields an estimated 84% increase in SDD market share. Thus, he shows that awareness is an important aspect to include when modeling demand for relatively new products such as SDDs and that information frictions overwhelm feature, price, and search considerations, at least for the hard drive market.

1.3 Extensions to the Sequential Search Model

Although the structural sequential search model can flexibly combine observed and unobserved preference and search cost heterogeneity for search and purchase behavior over horizontally and vertically differentiated products, it is still subject to rather stringent assumptions. A set of current research seeks to relax those assumptions and/or augment the model to accommodate additional important considerations. This section summarizes select areas of that research.

Gradual Learning through Revisits

Gradual learning is introduced to sequential search models by allowing for the natural possibility that the consumer is not only uncertain about what price (or match value) firms offer, but that she is also uncertain about the prevailing offer distribution(s) in the market.[6] Thus, by relaxing the rational expectations assumption – that is, by assuming consumers are not aware of all features, prices, and availability of all products at all times – each search not only reveals some information about the searched product, but permits the consumer to change her beliefs (usually updated in a Bayesian fashion) about the match-value distribution.

Chick and Frazier (2012) provides a solution method for dynamically and sequentially deciding which alternatives to sample and when to stop sampling, when the match-value distribution is unknown and learned through search. Their result provides an set of rules analogous to Weitzman's Selection, Stopping, and Choice Rules.

While I find the result from Chick and Frazier (2012) to most naturally study search behavior that where consumer revisit previously searched alternatives, Ursu, Wang, and Chintagunta (2018) cleverly use the result from Chick and Frazier (2012) to study search

[6]Search with learning problems are related to multi-armed bandit problems, but differ in that the consumer does not receive an award upon each search (lever pull) and when the consumer decides to stop search (not pull any additional levers) she receives the best of all previously observed rewards.

duration. In their application, consumers browse an online restaurant review website and each minute of search is treated as a unique search occasion. The authors find that the extensive margin of search to mirror previous findings in the search literature, namely that consumers search few alternatives, but that the intensive margin reveals important information. Search duration is considerable in their data, and consumers that search longer have an increases purchase probability.

In addition, the authors ask whether restaurants should modify the amount of information provided on their restaurant-specific pages of the website, and whether the platform should prioritize in its ranking restaurants that have more information on their restaurant-specific pages. They find that more information on restaurant-specific pages leads to fewer but longer searches, and greater consumer welfare; the magnitude is substantial with the increase in transactions equivalent to a 20% reduction in prices in the absence of a change in information. If the platform prioritizes restaurants with more information on their restaurant-specific pages, transactions again increase but now to lower-priced and generally lower-utility restaurants, leading to an increase in consumer welfare, but to a much lesser extent than in the first counterfactual analysis.

Adding Awareness

A mainstay of introductory marketing concepts is the consumer purchase funnel. The consumer's path to purchase proceeds through discrete activities of awareness, consideration, and choice. The sequential search models described thus far capture consideration and choice. Morozov (2019) augments the model to include awareness.

Specifically, he models the probability p_{ij} that consumer i is aware of product j for all products in the market. The consumer then elects to search over those products of which she is aware. Morozov (2019) finds that limited adoption of new products (such as the solid-state drives in the computer hard drive market) is mostly attributed to limited search and awareness rather than consumer preferences. His result emphasizes the need for marketing

strategies that aim to inform consumers about new products.

A related model is put forth in Honka, Hortacsu, and Vitorino (2017), where the authors model an awareness score for each product as a function of market factors and consumer demographics. They then construct a consumer's individual likelihood as the product of her probability of awareness, her probability of consideration conditional on awareness, and her probability of choice conditional on consideration. The latter two probabilities result from a simultaneous search model. The focus of their analysis is the effect of advertising. They find that advertising predominantly affects consumer awareness, but not consideration and choice. However, as a result of increased awareness, there is a trickle-down effect where consumers then search more and find better alternatives.

Preference Heterogeneity

Dong et al. (2020) makes use of a unique data set containing a panel of search histories and purchase occasions for a set of consumers. Using a sequential search model, they investigate persistence in product choice while accounting for individual-specific preferences. It is a stylized fact that consumers exhibit strong persistence in their product choices. Prior literature has interpreted this persistence as evidence that consumers have strong hetergeneous preferences over existing products. Dong et al. (2020) argues that this must not necessarily be the case. The authors show that search frictions, coupled with even mild heterogeneous preferences, can translate into highly persistent choices.

A major result from their work comes from comparing a standard discrete choice model with consumer-specific preference parameters to an analogous specification that also incorporates consumers' sequential search processes. The authors show that ignoring the search component results in overstatement of the scope for targeted marketing, an underestimate of own-price elasticities, and an overestimate of firm's markups. These results suggest an important role for consumer search when modeling demand, with implications for marketing, pricing, and policy decision making. In particular, the authors assess the potential profit

increases possible from personalized pricing based on each consumer's purchase and search history. When purchase history alone is used to set personalized prices, the firm can achieve a 6% increase in profits; an additional 3% increase is possible when consumers' search histories are also used.

Price Endogeneity

Berry, Levinsohn, and Pakes (1995) famously addresses price endogeneity in a discrete choice model of demand that is estimated with only aggregate data. Moraga-Gonzalez, Sandor, and Wildenbeest (2018) seek to do the same in the context of a sequential search model. To do so, the authors utilize a recent finding and a careful choice of probability distribution.

First, Moraga-Gonzalez, Sandor, and Wildenbeest (2018) uses a recent finding by Armstrong (2017) and Choi, Dai, and Kim (2018), which consists of a methodology to compute purchase decisions without taking into account the myriad search paths consumers may possibly follow. This is important in the setting of Moraga-Gonzalez, Sandor, and Wildenbeest (2018) where the authors use aggregate data that do not reveal the search sequences taken by consumers. Specifically, Armstrong (2017) and Choi, Dai, and Kim (2018) show that the Weitzman's solution to the sequential search problem is equivalent to picking the firm with the highest "effective value" w_{ij}, where an effective value is the minimum of the reservation utility and realized utility values:

$$w_{ij} = \min \{z_{ij}, u_{ij}\} \tag{1.12}$$

Second, the authors cleverly choose distributional assumptions for the match-value distribution and the search cost distribution. Specifically, match values are assumed to be IID T1EV (i.e., independent and identically distributed Type-1 Extreme Value) while search

costs are distributed according to the CDF

$$F_{ij}^{c} = \frac{1 - \exp(-\exp(-B^{-1}(c) - \mu_{ij}))}{1 - \exp(-\exp(-B^{-1}(c)))} \tag{1.13}$$

where $B(\cdot)$ refers to the marginal benefit of search (see Equation (1.1 above) and μ_{ij} is a consumer-product specific location parameter of the search cost distribution. Importantly, the authors show that the search cost CDF is not unlike a normal CDF or T1EV CDF and thus would not impact results any more than the arbitrary use of those common distributions.

The combined distributional choices for match values and search costs yield a closed-form probability that consumer i purchases product j conditional on the consumer's preferences. As a result, the only integral left in the likelihood function is over consumer-specific preferences, and the computational complexity of the model is rather mild.

With their computational approach established, the authors estimate their search model and compare it against a full-information model à la Berry, Levinsohn, and Pakes (1995). They find that the full-information model overestimates price sensitivity and leads to an overestimate of the absolute value of the own-price elasticity.

Future Directions Potentially Guided by Empirical Observations

While potential advancements in the modeling and estimation of search models has been the focus of this literature review, I close by noting that advances in data collection and availability have offered tremendous opportunity for contribution to the search literature.

In particular, Bronnenberg, Kim, and Mela (2016) use web browsing log files to observe URL-level browsing histories and transactions for a panel of consumers over time. Dong et al. (2020) observe a panel of search sequences for a set of consumers purchasing cosmetic products online. Seiler and Pinna (2017) use radio-frequency identification (RFID) tags on shopping carts at a grocery store to observe time spent in front of each product category. And subsequent chapters of this dissertation explore consumer search in the U.S. auto

industry using smartphone geolocations collected by mobile device applications. As Kim, Albuquerque, and Bronnenberg (2010) note, "the premise [of these data] is that we can learn about the preference of consumers by studying their shopping behaviors."

1.4 Search in the U.S. Auto Industry

A few recent studies assess consumer search behavior in the U.S. auto industry.

Albuquerque and Bronnenberg (2012) estimates a model of supply and demand that does not specifically incorporate consumer search behavior, but does utilize the locations of consumers and dealers (and thus the distance between them) to find that consumers have a strong disutility for travel. The focus of their analysis, however, is on consumer substitution patterns and dealership pricing in order to assess counterfactual closures of dealerships, a timely topic following the recession of 2007 when General Motors planned to consolidate its dealer network, reducing the number of U.S. dealerships by approximately 30%.

Palazzolo and Feinberg (2015) take an approach motivated by simultaneous search. Instead of a full structural model, the authors focus on consideration set substitution among products. They generalize the full-information discrete choice model to incorporate the probability that an observed consideration set is optimal while flexibly permitting consideration set substitution among available options. Motivated by then-recent Toyota recalls due to faulty breaking equipment and Tohoku earthquake and tsunami in Japan, the authors use their model to study the impact of vehicle redesigns and recalls on consideration and purchase probabilities.

Murry and Zhou (2019) studies the agglomeration-competition trade-off of dealership co-location using a sequential search model and transaction-level data for new vehicles sold in the state of Ohio. The authors assume that consumers search dealership clusters simultaneously, and like Moraga-Gonzalez, Sandor, and Wildenbeest (2018), observe the locations of the consumers and dealerships as well as the selling dealership, but do not observe consumer

search decisions. By comparing results of the search model to a standard model of full information, the authors find, among other things, that search frictions generate over $300 in mark-ups.

CHAPTER 2

Data

2.1 Data Sources

I combine data from several sources for my empirical analysis. Mobile device geolocation data inform me about consumers' home locations and dealership visits. However, these data do not provide information on the purchased vehicle. Therefore I combine the dealership visit data with data on new car registrations from the Texas DMV. By combining these two data sets, I observe both the search sequence and the purchased vehicle at the individual level. I supplement these data with consumer and vehicle characteristics, and information on the location of auto dealerships and the brands they carry.

2.1.1 Search Data

In partnership with a professor at UCLA, I obtained consumer search data from Safegraph, a company that "provides high-quality location data products" by aggregating location information from various mobile device applications.[1] I was provided with two types of information for each (anonymized) individual mobile device: estimated home locations and dealership visits. The estimated home locations were generated by a proprietary model that is based on the (im)mobility of the device, time of day, and assumed work patterns of device

[1]See https://github.com/YalePrivacyLab/tracker-profiles/blob/master/trackers/SafeGraph.md and https://www.safegraph.com/.

owners. The estimated home locations are stored as a geohash.[2] The dealership visits were generated by Safegraph merging their mobile device geolocation data with their proprietary data set of U.S. dealership geospacial locations. A unique record is a mobile device at a dealership (identified by name and street address) at a specific date and time. Thus, for each mobile device, I observe the visited dealerships and the order of those visits. The data span the three-month time period from November 1, 2016 to January 31, 2017; I limit the data to dealership visits in the state of Texas.

The dealership visit data record all dealership visits, for any reason. As a result, they may capture behavior other than new car searches. To remove errant observations, I exclude all data for a device if the device is observed (i) at any dealership more than 25 times, (ii) at the same dealership more than 10 times, or (iii) at any dealership between 12:00am and 6:00am. These criteria are applied to exclude, for example, the device of someone who works at, delivers vehicles to, or provides janitorial services for an auto dealership. The remaining data consist of approximately 154,000 unique mobile devices making 277,000 dealership visits to one or more of the 1,258 dealerships identified by Safegraph.

Dealership visits of consumers shopping for a new car can spread over days or weeks. This raises the concern that vehicles purchased at the beginning of the three-month time period may have truncated search sequences. To investigate this concern, I calculate the average number of days between the first and last search for consumers whose last search occurred in November 2016 versus during December 2016 and January 2017. I find these values to be 1.1 and 10.2 days, respectively.[3] These numbers suggest that many search histories for consumers who made a purchase in November 2016 are truncated. As a result, I limit my analysis to only those consumers who purchased their vehicles in the latter two

[2]A geohash is a public domain geocoding system that encodes geographic areas into short alphanumeric strings.

[3]If I exclude consumers who only made one search, the mean and median number of days between the first and last search for consumers whose last search occurred during first month are 9.2 and 7, while the mean and median number of days between the first and last search for consumers whose last search occurred during the latter two months are 29.3 and 25.

months ("sample period"). This reduces the sample size to approximately 128,000 consumers making 243,000 dealership visits.

The sequence of data preparation steps is summarized in Table 2.1.

Table 2.1: Filters Applied to Search Data

Filter	Devices Remaining	Pings Remaining
Visits to Dealerships in Texas	157,582	313,806
Remove any Device at any Dealership > 25	157,173	296,889
Remove any Device at Same Dealership > 10	156,439	285,054
Remove any Device at Any Dealership at Night	153,848	277,083
Remove any Device with last Visit $<$ Dec 2016	127,595	243,426

2.1.2 Purchase Data

My second data come from the Texas DMV. I observe all first-time titled or registered vehicles in the state of Texas for the 16-month period spanning June 2016 through September 2017. The data include the Vehicle Identification Number (VIN), the registrant, the registrant's address, the date that the title or registration paperwork was processed by the state of Texas, the gross sales price before any adjustment for a trade-in vehicle, the VIN for the trade-in vehicle if applicable, and the name, city, and state of the previous owner. It is important to note that information on the previous owner identifies the dealership from which a vehicle was purchased ("selling dealership"). While most of this information is available to the public through a Freedom of Information Act (FOIA) request, the personally identifying information is not. It must be obtained through a special request and its use is subject to restrictions.

To focus on new retail vehicle sales for which consumer search data may be available from Safegraph, I limit the Texas DMV data to vehicles that belong to the 35 most popular

brands from model years 2015–2018 with an odometer reading of fewer than 2,000 miles, a price exceeding $5,000, and for which the date of title occurred during the sample period (or less than 14 days after its end).[4] These criteria are used to omit used vehicles, commercial vehicles such as freight trucks, and alternative vehicles such as motorcycles, motor homes, and tractors. In addition, I constrain the Texas DMV data to only vehicles for which the registrant has a non-PO Box home address in the state of Texas; this is a necessary criterion for combining the search and purchase data. Finally, data are excluded if more than one vehicle is titled to the same individual during the sample period or if the registrant is not an individual. Approximately 195,000 vehicles meet these criteria.

Figures 2.1a and 2.1b each provide a map of individuals' home locations. Figure 2.1a displays home locations for searchers from the Safegraph data, while Figure 2.1b displays home locations of buyers from the Texas DMV data. Although the data sources are different, the figures show a similar geographic distribution of consumers across the state and major urban areas are clearly identifiable.

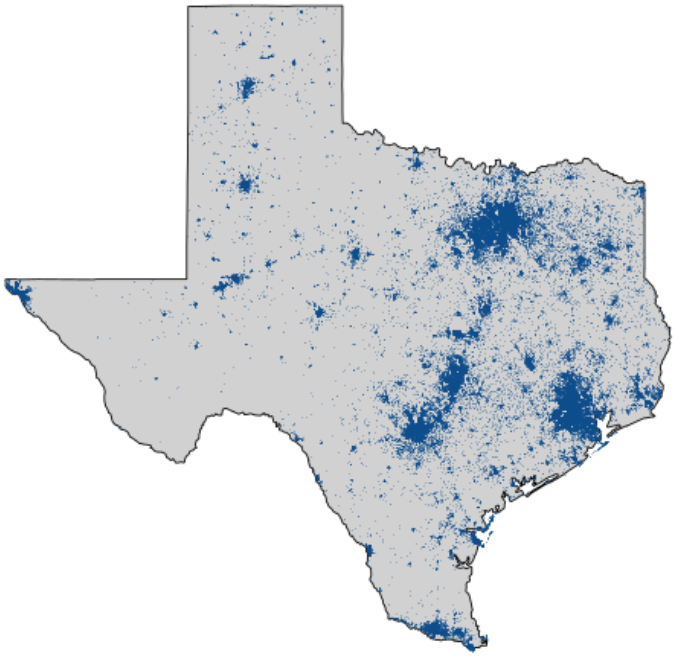
2.1.3 Dealership Data

I manually prepared a third data set consisting of all auto dealerships in the state of Texas and the brands carried at each dealership. This data collection was necessary for two reasons. First, the colloquial and legal definition of a dealership vary and I required a definition compatible with the Safegraph data. I use the term and have organized the data such that a "dealership" represents a distinct geographic area of new vehicle retailing. For example, although Randall Noe Chrysler Dodge and Randall Noe Subaru are legally distinct dealerships, their showrooms share the same building, they have the same street address, and the vehicles in inventory are adjacently located on the same lot. I therefore categorize them

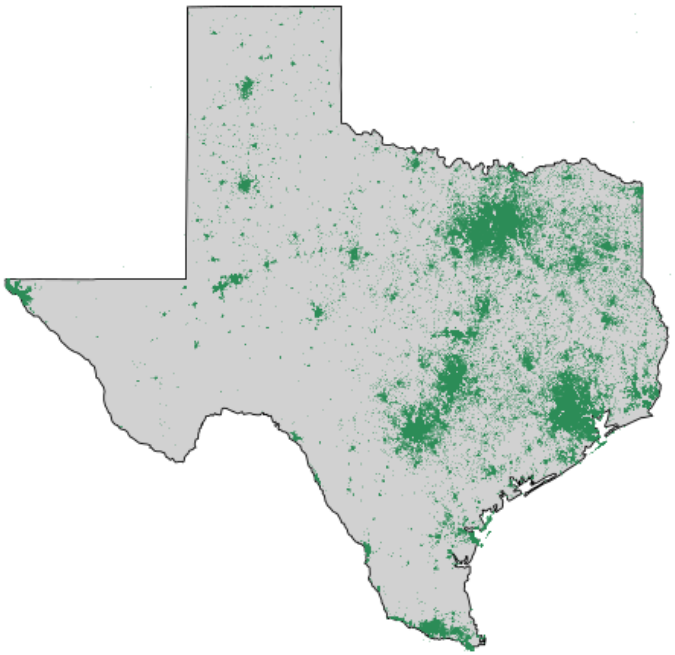
[4]In a conversation with a Texas DMV employee, her experience was that title and registration paperwork is typically processed between one and fourteen days after submission to the DMV or county tax office, and that submission typically occurs immediately upon sale of a vehicle by a dealership.

Figure 2.1: Home Locations

(a) Searcher Data
(from Safegraph)



(b) Buyer Data
(from Texas DMV)



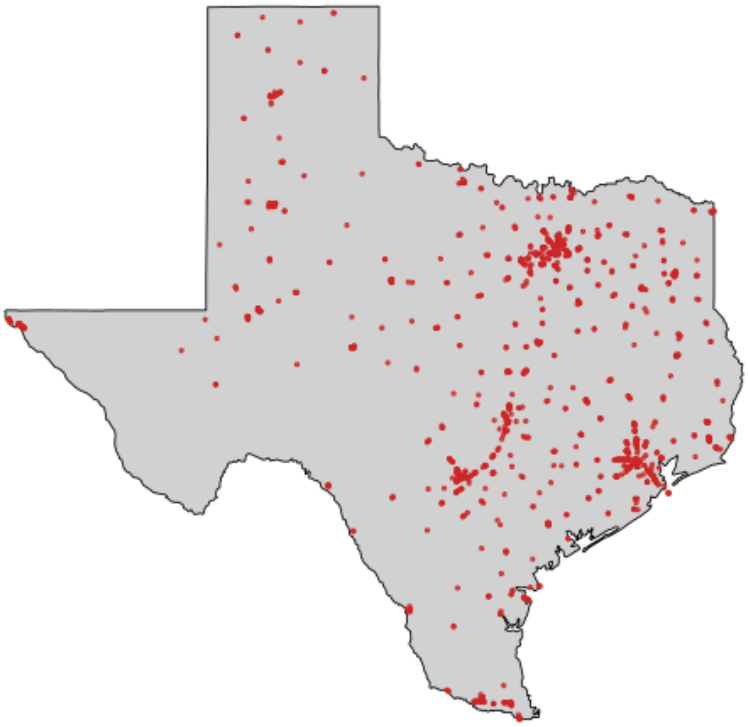
as one dealership. By contrast, Blue Bonnet Motors, located at the intersection of North Interstate 35 Frontage Road and McQueeney Road in New Braunfels, Texas (a suburb of San Antonio), appears to be a single distinct legal entity, but with separate showrooms and street addresses for the Ford, Lincoln, and Jeep brands. I therefore categorize them as separate dealerships. Second, to make reasonable assumptions about vehicles that were searched but not purchased, as required for the structural search model to be introduced later, I require information on the set of brands carried at each dealership. For example, often the General Motors brands Chevrolet, Buick, GMC, and Cadillac are often available at the same dealership.

I identify 1,314 auto dealerships that carry new vehicles in the state of Texas. This number of dealerships is very similar to the number of dealerships (1,309) listed by the Texas Automotive Dealership Association (TADA) as of February 2018.[5] The small difference in the number of dealerships is due to (i) dealerships that operated during the sample period but closed by the 2018 TADA count and (ii) TADA and I differing in how to treat two adjacent and related retail locations (TADA may recognize them as two dealerships while I count them as one or vice versa). Figure 2.2 plots the locations of all 1,314 auto dealerships across the state of Texas. As expected, dealership locations are highly correlated with the geographic distribution of the population as shown above in Figure 2.1.

Consistently identifying dealerships across data sets is essential for combining the search and purchase data. The manually compiled data on dealerships are matched (i) to the dealership street addresses in the Safegraph search data and (ii) to the previous owner names in the DMV registrations data. An overlapping subset of 1,197 dealerships are identified from both data sets. Not all dealerships could be merged either because the previous owner names in the Texas DMV data were not sufficiently descriptive to uniquely identify a dealership in the Safegraph data or because the Safegraph data did not contain visits to that dealership.

[5] https://www.tada.org/web/Online/About_TADA/Dealer_Search.aspx

Figure 2.2: Texas Dealership Locations



This latter case could occur if Safegraph did not have a geofence for that dealership or if no mobile device in the Safegraph data was observed to visit the dealership during the three-month sample period.

2.1.4 Vehicle Characteristics Data

VinAudit.com, Inc., a leading vehicle data and software solutions provider for the U.S. automotive market, provided data on vehicle characteristics. I obtained the data by "decoding" the registered and trade-in VINs using the company's API. Collected vehicle characteristics include model year, make, model, trim, "base" MSRP, vehicle type (car, SUV, truck, van), body type (e.g., sedan), number of doors, drive type (e.g., front-wheel drive), engine size, and transmission type.[6] I supplement these data with information on vehicle horsepower collected from Google.

I also collected data on vehicle "types" and rankings within each type from Edmunds.com. Edmunds classifies all vehicle models into one of 40 types. For example, the Honda Fit is categorized as an Extra-Small Hatchback, while the Toyota Highlander is categorized as a Midsize 3-Row SUV. Within a type, Edmunds ranks each vehicle. Their ranking process is proprietary and opaque.[7] The Edmunds data are used to make assumptions about (similar) searched or potentially searched, but not purchased vehicles when modeling consumer choices and behavior (see Section 2.2.3).

[6]The APIs used to gather car characteristics provided "base" MSRP. The definition of "base" is manufacturer-specific and may be at either the model-level or trim-level. These base prices do not include features that customize the vehicles beyond the set of included features in the most basic configuration of the manufacturer's model or trim. For example, if a sunroof is included as part of a specific 2017 Ford Fiesta SE but the sunroof is not a standard feature on all 2017 Ford Fiesta SEs, then the added cost of the sunroof is not included in the collected MSRP. This means that, for example, two vehicles that are largely similar (i.e., same model or same trim) but have different optional features will show the same MSRP.

[7]Edmunds describes its ranking process as: "Each vehicle is driven on a standardized road test loop and visits our test track for instrumented testing in controlled conditions. Our time behind the wheel is used to develop ratings that describe how a car stacks up against its direct rivals in a particular size and price class." (https://www.edmunds.com/new-car-ratings/).

2.1.5 Consumer Characteristics Data

I collected consumer demographic data from the U.S. Census Bureau's 2010 Census and American Community Survey. The data include information at the Census Blockgroup level on age, race, gender, educational attainment, income, employment status, and number of children.

2.2 Data Cleaning and Construction of Analysis Data Set

To empirically estimate a search model, I must (i) combine the search and purchase data, (ii) define the searchable set of dealerships, (iii) define the searchable set of vehicles, and (iv) create the final sample for the empirical analysis. I provide a detailed decription of these four steps below. Note that I limit the analysis sample to the five brands with the largest market shares in Texas during the sample period. This set of brands has a combined market share of 60%.

2.2.1 Combining Search and Purchase Data

There is no unique identifier in the Safegraph dealership visit data that corresponds to information in the Texas DMV registrations data to indicate which search sequence corresponds to which purchased vehicle. Therefore, to merge these two data sets, I employ the following algorithm: first, the algorithm finds a search sequence that includes a visit to the selling dealership prior to the vehicle registration date, and second, the algorithm requires that the home location of the searcher and of the buyer are approximately the same. In the following, I provide more details on the employed algorithm.

First, recall that the home location of a mobile device user in the search data is an estimated geohash while the purchase data from the Texas DMV provides the registrant's street address. I geocode both sets of location information as latitude and longitude coordinates.

For the search data, I use the center of the geohash; for the purchase data, I use coordinates from the GoogleMaps API, which is reputed to often report the centroid of the building at the specified address, but in some cases may report a streetfront location (i.e., effectively where a mailbox would be placed).

I then require that the two home locations are in close proximity. To do so, I calculate the distance between each of the 154,000 mobile device home locations in the dealership visits data and each of the 264,000 unique registrant's home addresses in the purchase data using the Haversine great-circle distance measure, which is the shortest distance between two points on a sphere. Due to computer memory constraints, for each mobile device home location, I retain up to a maximum of 50 "potential merges" (i.e., pairs of searcher and buyer vectors of information) where the distance between the home locations is one-fifth of a mile or closer. In short, I find up to 50 vehicle sales that could potentially be the result of each observed search sequence.

Next, I evaluate these potential merges in increasing order of distance. For each potential merge, I assess whether the mobile device had visited the selling dealership and whether that visit occurred on or before the vehicle registration date. If so, I merge that search sequence with that purchased vehicle, and I discard all other potential merges involving that search sequence or vehicle. If, instead, the potential merge involved home locations that were close in proximity but where the mobile device had not visited the selling dealership on or before the registration date, I discard that potential merge.[8] I then evaluate the next (in terms of shortest distance) potential merge. The process proceeds until all potential merges are evaluated and either accepted or discarded.

With this algorithm, I find the purchased vehicle for 12,065 search sequences. This repre-

[8]This procedure minimizes the probability that I incorrectly merge a search sequence and a purchased vehicle simply due to the proximity of home locations in the search and purchase data sets. For example, if two vehicles were purchased during the sample period by residents of the same apartment building and both have search sequences captured by Safegraph's dealership visits data, I am able to correctly merge the search sequences to the purchased vehicles unless both consumers visited both selling dealerships, which is rather unlikely given the relatively short search sequences observed by most consumers.

sents 9.5% of the mobile devices in the Safegraph data and 6.2% of the vehicle registrations in the Texas DMV data. These figures appear reasonable for the following reasons. First, the search data also include dealership visits for consumers who were shopping for used vehicles or who were taking a currently-owned vehicle to a dealership's service department for maintenance – both activities are likely to occur with greater frequency than new car shopping. Second, the search data also include dealership visits by consumers who searched, but did not purchase a new vehicle. And third, the search data purportedly represent 5%–10% of U.S. mobile devices during the sample period and thus it is expected that these data capture approximately the same percentage of new car purchases.

Figure 2.3 provides maps of the four metropolitan areas onto which dealership locations and consumer home locations are overlaid; the set of consumer home locations plotted includes only those for whom I could identify both the search sequence and registered vehicle following the algorithm outlined above. It shows that dealerships often occur spatially clustered and are generally located along major roadways. And, as expected, consumers are often clustered into residential areas (e.g., the southwest and suburban areas of Houston) and are sparse in industrial areas (e.g., the northeast area of Houston where the George Bush Intercontinental Airport is located).

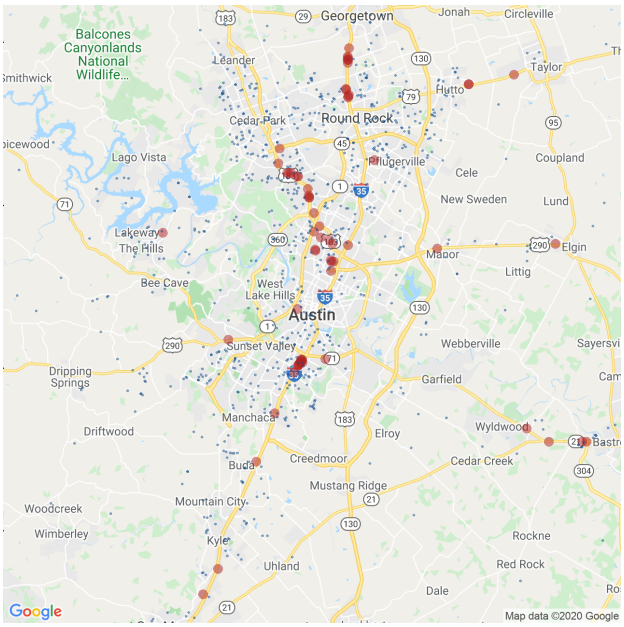
2.2.2 The Searchable Set of Dealerships

In much the same way that a multinomial discrete choice model requires information on the chosen product as well as the products not chosen, my structural search model requires information on the set of (potentially) searchable products, regardless of whether the consumer searched them or not.[9] Across the 35 brands, there are more than 1,100 dealerships in the state of Texas. To ease the computational burden imposed by such a large set of searchable

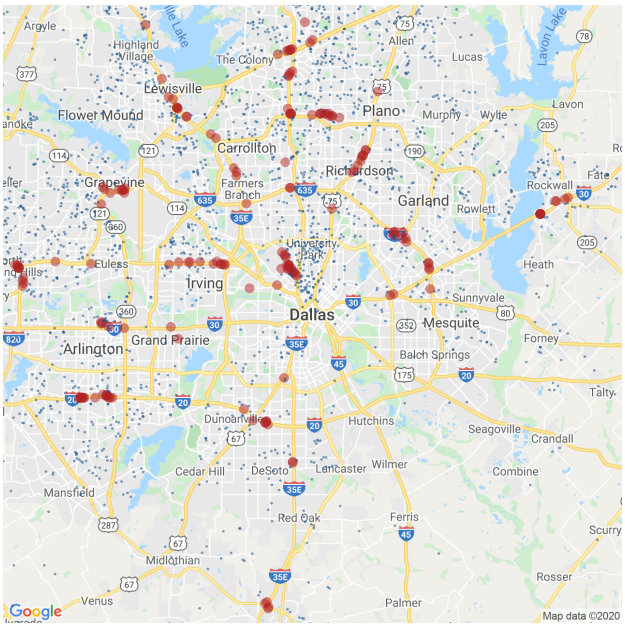
[9]As explained in Section 2.2.3, a "product" here is a vehicle at a dealership. I will assume a consumer searches her most-preferred vehicle at each dealership and thus it is equivalent to reference the searchable set of dealerships or searchable set of vehicles.

Figure 2.3: Dealership and Consumer Locations in Select Cities

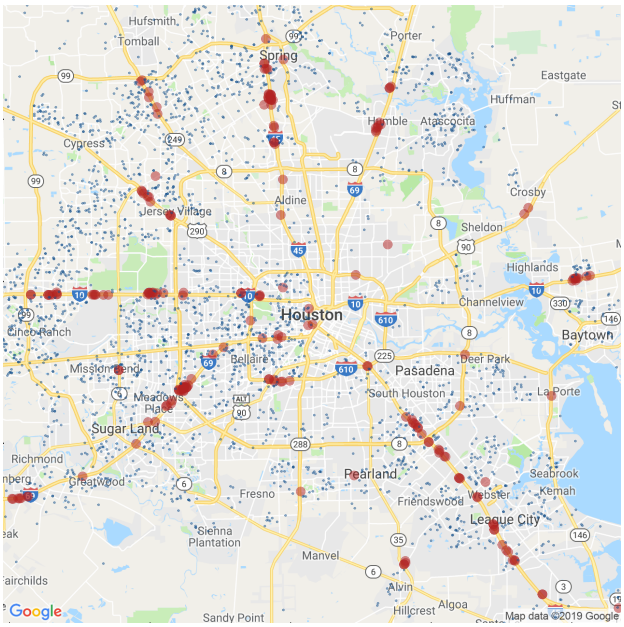
(a) Austin



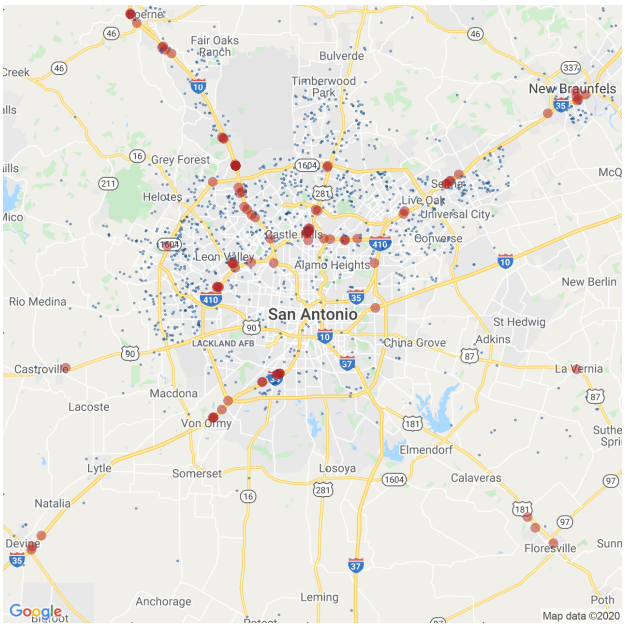
(b) Dallas



(c) Houston



(d) San Antonio



products, I take two steps.

First, I limit the analysis to the five brands with the largest market shares in Texas during the sample period, which are Chevrolet, Ford, Honda, Nissan, and Toyota. This set of brands has a combined market share of 60%. Focusing the analysis on these brands requires that I also limit the data to consumers who purchased these brands and whose search histories only include visits to dealerships that carry these brands. This reduces the sample size to 7,051 consumers.

Second, I limit the searchable set of dealerships to any dealership within a particular radius of the consumer. To define the radius for each consumer, I first partition consumers into urban and rural consumers. Urban consumers are those who live within the city limits of any major city (where major is defined as a city with a population exceeding 100,000 inhabitants) and rural consumers live outside of these major cities. For urban consumers in each major city and for rural consumers in each county (excluding inhabitants of the major cities in that county), I assign a radius as twice the 70th percentile of the distance between the home location of a car buyer and the selling dealership, where I calculate distance distributions from the original 16-month DMV data set containing 195,000 new vehicle DMV registrations.

For example, suppose a consumer lives in the rural part of Dallas County and my data contain 10,000 new vehicle registrations with the Texas DMV from rural residents of Dallas County. I calculate the 70th percentile of the distance between those 10,000 residents' home and the selling dealerships. Suppose this value is 30 miles. Then for each consumer in the analysis sample who lives in the rural part of Dallas County, I assign a radius of 60 miles. Therefore, the searchable set of dealerships for rural Dallas County residents includes any dealership within 60 miles of the consumer's home location. Note that this set of dealerships may differ across rural residents of Dallas County if some dealerships are within 60 miles of some consumers' homes, but others are not.

Across the state, the average radii for urban and rural consumers are 29.6 and 33.5 miles,

respectively. I find that 7.6% of consumers search or purchase a vehicle from a dealership outside of their assigned radius and exclude these consumers from the analysis sample. This step reduces the sample size to 6,511 consumers making 7,175 dealership visits.

2.2.3 The Searchable Set of Vehicles

Through the process of merging the search and purchase data, I know which vehicle each individual consumer purchased at which selling dealership. However, for all dealerships she visited but did not make a purchase, I do not observe which vehicle she searched. Further, for all dealerships the consumer decided not to visit, I do not observe which vehicle she would have searched if she had decided otherwise. To estimate the structural search model, I need both pieces of information. I make the following set of assumptions about searched and potentially searchable vehicles.

For each consumer and each searchable dealership identified by the process outlined in the previous section, I assume that there is one most-preferred vehicle that could be searched, i.e., if the consumer were to visit a particular dealership, she would have searched only her most preferred vehicle at that dealership. It is important to note that the most-preferred vehicle at a particular dealership is consumer-specific: if the same dealership is searchable by multiple consumers, each consumer has her own most-preferred vehicle at that dealership. To identify this consumer-specific, most-preferred vehicle at each dealership, I impose two criteria: (1) the potentially searched vehicle must be in that dealership's inventory and (2) it must be of the same Edmunds type as the vehicle ultimately purchased by the consumer.

For the first criterion, "inventory" is defined as any model sold by that dealership during the 16 months for which the Texas DMV registration data are available.[10] This criteria enforces that consumers can only search for a model at a particular dealership if that dealership has recently carried and sold that model.

[10]A model is a particular combination of model year, make, and model. For example, one model is a 2017 Toyota Corolla.

In addition to being in inventory, the second criterion requires that a searched vehicle must be of the same Edmunds type as the vehicle ultimately purchased by the consumer. This assumption is equivalent to assuming that consumers search for a specific vehicle conditional on having decided the type of vehicle they would like to buy. This assumption is consistent with prior literature and industry reports of consumer search behavior. For example, Honka (2014) and Honka and Chintagunta (2017) also do not observe searched-but-not-purchased alternatives and make the assumption that the most-similar product offered by other sellers is the product searched. Moreover, industry reports indicate that the new car purchase process usually starts with online search, which is broader and where consumers may consider different types of cars. But as consumers move to offline search (e.g., visiting dealerships), they usually have a specific car in mind.[11] This view is reflected in prior academic literature on modeling vehicle choice. For example, Albuquerque and Bronnenberg (2012) estimate a nested logit model and choose the first level of nests defined by car type and the second level of nests composed of alternatives of the same type (but different brands).

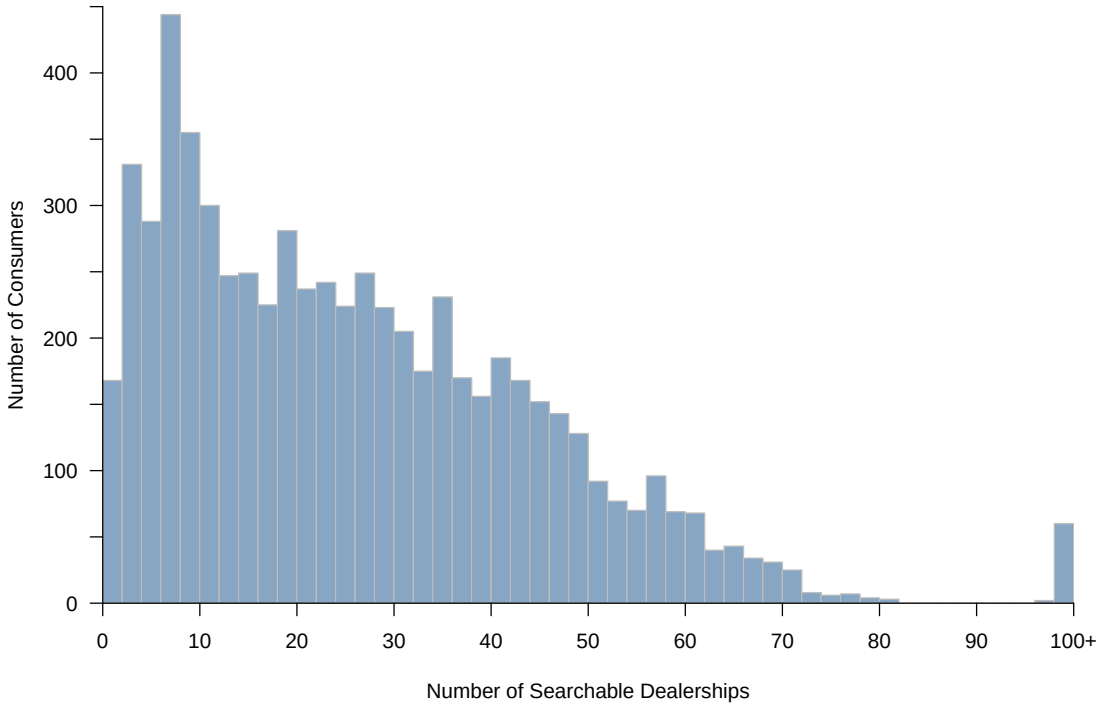
To continue the earlier example in order to demonstrate this process: suppose the rural Dallas County consumer is observed to purchase a 2017 Nissan Versa Note and that her searchable set of dealerships includes XYZ Chevrolet. According to Edmunds, two Chevrolet vehicles are of the same type as the Nissan Versa Note (type: "Extra-Small Hatchback"). Those two vehicles are the Chevrolet Spark and the Chevrolet Sonic. Suppose that the Texas DMV registrations data show that the XYZ Chevrolet dealership sold zero 2017 Chevrolet Sparks and nineteen 2017 Chevrolet Sonics during the 16 months for which I have Texas DMV registration data. Then the searchable vehicle for my example consumer at XYZ

[11] See, for example:
https://v12data.com/blog/automotive-marketing-overview-current-marketing-trends-statistics-and-strategies/;
https://www.acaresearch.com.au/australian-market-research-blog/the-automotive-vehicle-purchase-journey-in-2015;
https://www.edmunds.com/car-buying/10-steps-to-finding-the-right-car-for-you.html;
https://santanderconsumerusa.com/learning-center/i-need-another-car-now-what;
https://growwithcars.com/wp-content/uploads/2014/05/2014.05-Consumer-Journey-Report Final.pdf.

Chevrolet is assumed to be the 2017 Chevrolet Sonic because (i) it has the same model year and the same Edmunds type as the purchased vehicle and because (ii) it was in inventory at the searched dealership.[12]

In total, my analysis sample contains 175,840 potentially searched vehicles across the 6,511 consumers. The number of potentially searchable vehicles is plotted in Figure 2.4, which range from 1 to 177, with a mean of 27 and a median of 24 vehicles.

Figure 2.4: Searchable Set Sizes



Finally, I must also infer the set of vehicle characteristics for each searchable vehicle. To do so, I compute the median values (for numeric characteristics) and modal values (for categorical characteristics) by model and dealership. Thus if the rural Dallas County consumer could search a 2017 Chevrolet Sonic at two different Chevrolet dealers (say ABC Chevrolet

[12]In the event that multiple vehicles meet these criteria, the vehicle that is ranked highest according to Edmunds.com is selected as the most-preferred vehicle. This occurs in the analysis dataset for less than 1% of searchable vehicles and occurs only once for one consumer at a dealership that she chose to search.

and XYZ Chevrolet), the city mileage of the 2017 Sonic at ABC Chevrolet depends on the city mileage of the other 2017 Sonics sold by ABC Chevrolet and differs from the city mileage calculated for the 2017 Sonic at XYZ Chevrolet. This approach is consistent with that taken by Moreno and Terwiesch (2017).[13]

2.2.4 A Detailed Example

To summarize the information available following the aforementioned assumptions made during the data cleaning and construction of the analysis dataset, I provide a detailed look at one consumer. This consumer is identified with a unique, anonymized Safegraph identification code ending in "2c76". This consumer is observed to have visited 3 dealerships during the sample period: First Texas Honda at 11:32am on 1/20/17, Round Rock Toyota at 2:34pm on 1/20/17, and Round Rock Honda at 3:38pm on 1/24/17. This consumers' home location is the closest remaining home location (during the merging algorithm) to the registration address for a Honda Odyssey EX-L sold by Round Rock Honda (VIN ending in "4939") with paperwork processed by the Texas DMV on 2/6/17. Because the consumer visited the selling dealership prior to the date the DMV processed the registration paperwork and because the consumer's home location is the closest remaining homelocation (and not greater than 0.2 miles from the registrant's address) during the merging algorithm, I identify this vehicle as the one purchased by this consumer. For the rest of this example, I will arbitrarily choose refer to the consumer as Mary for simplicity.

Mary lives in Round Rock, Texas (William County), a city of more than 100,000 inhabitants and thus Mary is classified as an urban consumer. Out of the approximately 195,000

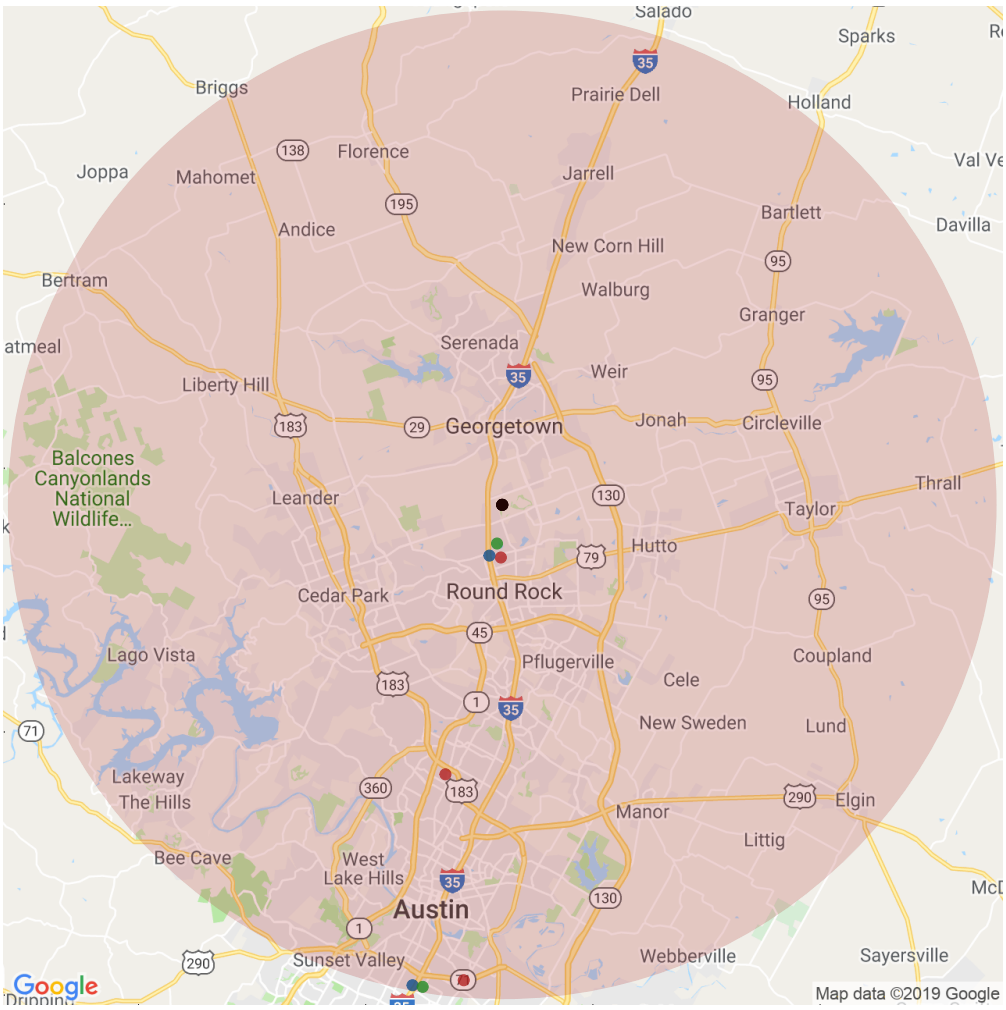
[13]Note that I use median and modal values for the purchased vehicle as well as for each potentially searchable, as consistency is required in order to model the consumer's choice between vehicles. Also note that it would be possible to obtain the exact characteristics of each vehicle sold during my sample period by obtaining and codifying the information on the Monroney Label (also known as the "window sticker"). However, while some private services exist to view a single Monroney Label, I have not found any services that provide the information for all brands in the market, that offer an API, or that price reasonably to facility large-scale statistical analyses. Given that I would require such data for every potentially searched vehicle during my sample period, the time and monetary costs exceed what is feasible.

new vehicle registrations of passenger cars and light duty trucks in the Texas DMV data made available to me, 3,012 were registered to an address in William County and, of those, 858 were registered to an address in Round Rock, Texas. The 70th percentile of the distances between these 858 registrant addresses and the selling dealerships is 14.38 miles. Thus I assume that Mary could search any dealership within 28.76 miles of her home (twice the 70th percentile distance). There are 18 dealerships across the five focal brands within 28.76 miles of Mary's home. However, Mary purchased a Honda Odyssey EX-L, a vehicle in the Edmunds "minivan" class. Chevrolet and Ford do not offer a vehicle in this class, and thus I assume Mary will not search any Chevrolet or Ford dealerships. There are 7 remaining dealerships for which I identify a most-preferred vehicle for Mary (3 Honda dealerships, 2 Nissan, and 2 Toyota) and which she could potentially search. According to the Edmunds classification system, the other minivan vehicles include the Toyota Sienna and Nissan Quest, both of which were found to have been sold by the nearby dealerships and thus assumed to be in inventory and potentially searchable by Mary. I then compute the distance to each of these dealerships from Mary's approximate home location and the median/modal vehicle values as the vehicle characteristics for the potentially searchable Honda Odysseys, Toyota Siennas, and Nissan Quests. Note that the characteristics computed may differ by dealership, even for the same make and model. Figure 2.5 plots Mary's approximate home location in black as well as the Honda (red), Toyota (blue), and Nissan (green) dealership locations, and the radius of 28.76 miles around Mary's home.

2.3 Descriptive Statistics

The final analysis sample contains $6,511$ consumers who make $7,175$ visits to one of 544 unique dealerships. 91.1% of consumers search once, 7.8% of consumers search twice, and 1.1% of consumers search three or more times. Figure 2.6 shows a histogram of the number of searches. The average number of searches per consumer is 1.1. This average number of

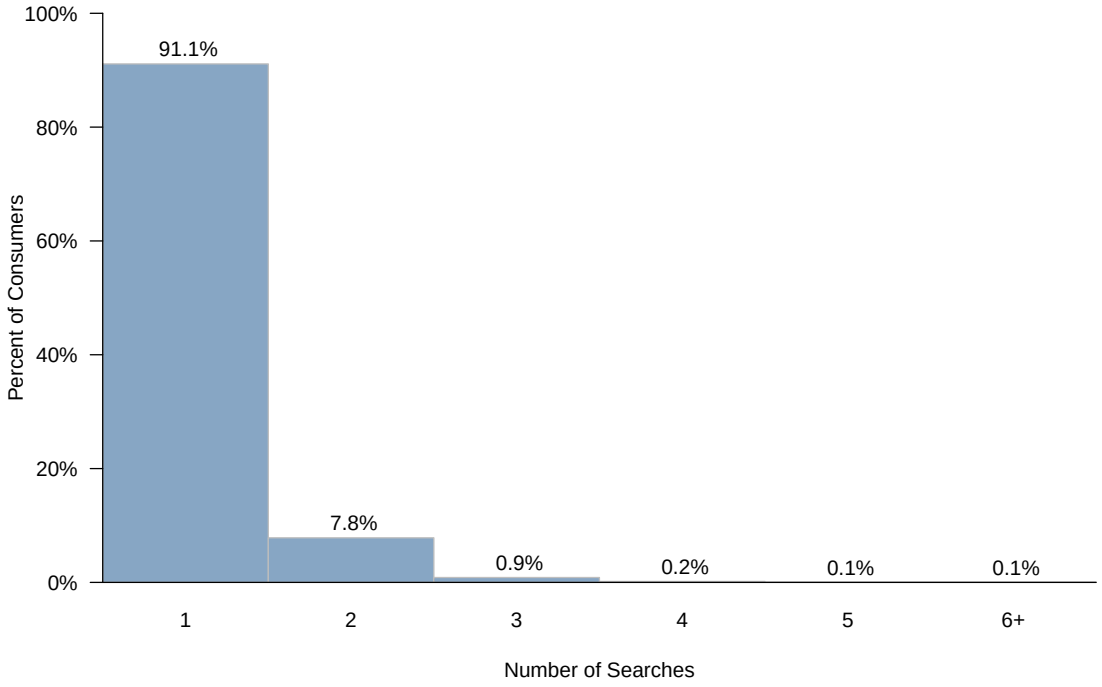
Figure 2.5: Example Consumer's Home Location and Nearby Searchable Dealerships



searches in my data is consistent with the trend of search behavior reported in academic literature (see Table 2.2) and with the most recent reports, albeit slightly smaller. For example, during an FTC Workshop on auto distribution in the U.S., Peter Welch, President of the National Automotive Dealers Association (NADA), reported that "the average number of dealerships that a consumer visits before they make a purchase, an actual purchase, has gone down from 4.1 dealerships in 2005 to today, it's 1.3.[14] Much of this decline is driven by

[14]https://www.ftc.gov/system/files/documents/public events/895193/auto distribution transcript.pdf.

Figure 2.6: Distribution of Number of Searches



online search (i.e., online browsing or online research), which is how the new car shopping process commonly starts.[15]

The distance a consumer must travel to visit a dealership is my key variable. Consumers live in close proximity to many dealerships. On average, consumers in the analysis sample live within 10 miles of 6 dealerships. As I extend the radius to 20 and 30 miles, that number increases 16 and 24 dealerships, respectively. The observation that consumers live in close proximity to many dealerships, but only search a limited number of them suggests that it is important to take their search behavior into consideration when modeling demand.

Not only are many dealerships located near consumers, but consumers also tend to purchase from nearby dealerships. The median distance from a consumer's home to the selling

[15]See, e.g., Ratchford, Talukdar, and Lee (2007) or https://geomarketing.com/86-percent-of-car-shoppers-do-research-online-before-visiting-a-dealership.

Table 2.2: Consumer Search Behavior in the U.S. Auto Industry Reported in Academic Literature

Study	Findings
Kiel and Layton (1981)	35% of carbuyers made 2 or fewer visits, 20% made 6+
Ratchford and Srinivasan (1993)	4.6 dealer visits on average
Lapersonne, Laurent, and Le Goff (1995)	22% only considered one brand
Klein and Ford (2003)	39% visited 2 or fewer dealers
Ratchford, Lee, and Talukdar (2003)	"seriously considered" 2.5 dealers on average
Zettelmeyer, Morton, and Silva-Risso (2006)	online buyer collected info for 2–3 cars; offline buyer 1 car
Ratchford, Talukdar, and Lee (2007)	2.19–2.53 median dealer visits between 1990–2002
Kim and Ratchford (2012)	1.7 median number of manufacturers considered
Singh, Ratchford, and Prasad (2014)	3.24 dealers visited on average
Jang, Prasad, and Ratchford (2017)	0.7–1.43 dealers visited on average between 2002–2012
Palazzolo and Feinberg (2015)	45.8% considered 1 vehicle, while 32.5%, 15.2%, and 6.5% considered 2, 3, and 4+ vehicles
Murry and Zhou (2019)	estimate 46.7% search 1 dealer cluster, 34.4% 2 clusters, 18.7% 3 clusters, 0.2% 4+ clusters
Moraga-Gonzalez, Sandor, and Wildenbeest (2018)	histogram: (1 dealer visit) 47%, (2) 20%, (3) 16%, (4) 7%, (5) 3%, (6) 3%, (7) 1%, and (8+) 2%

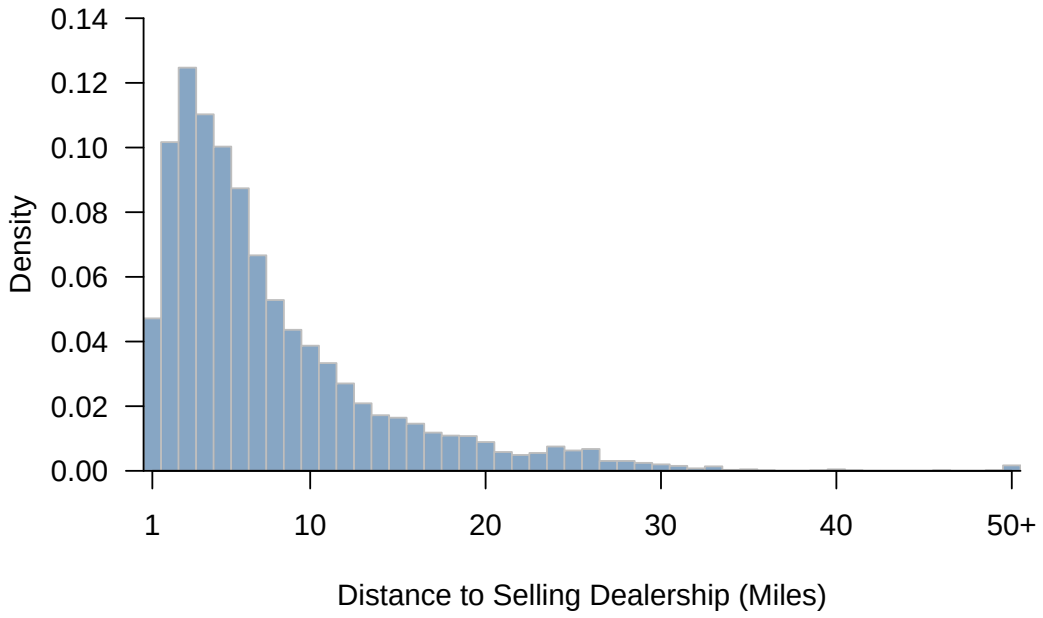
dealership is 5.2 miles. Figure 2.7a shows the distance distribution from home to the selling dealership: most consumers purchase from dealerships within 30 miles of their home.[16] As a point of comparison, in Figure 2.7b, I display the distribution of percentiles for the distance to the selling dealership in each consumer's set of searchable dealerships. For example, suppose a consumer bought from the closest dealership and had five searchable dealerships. Then the consumer purchased from a dealership in the 0.2 percentile. The distribution in Figure 2.7b shows that consumers tend to purchase from dealerships below the 0.5 percentile. The few consumers who purchase from a high percentile dealership are mostly consumers with small sets of searchable dealerships (fewer than five dealerships).

Consumers can purchase a vehicle from the closest dealership carrying the purchased

[16] Albuquerque and Bronnenberg (2012) find the mean and median distance between consumers' homes and *selling* dealerships to be 10 and 7.3 miles, respectively, based on zip code centroids. These distances are also consistent with not observed, but *estimated* distance distributions reported by prior literature (e.g., Nurski and Verboven 2016, Moraga-Gonzalez, Sandor, and Wildenbeest 2018, Murry and Zhou 2019).

Figure 2.7: Consumers' Distances to Dealerships

(a) In Miles



(b) As Percentiles of Their Searchable Sets of Dealerships

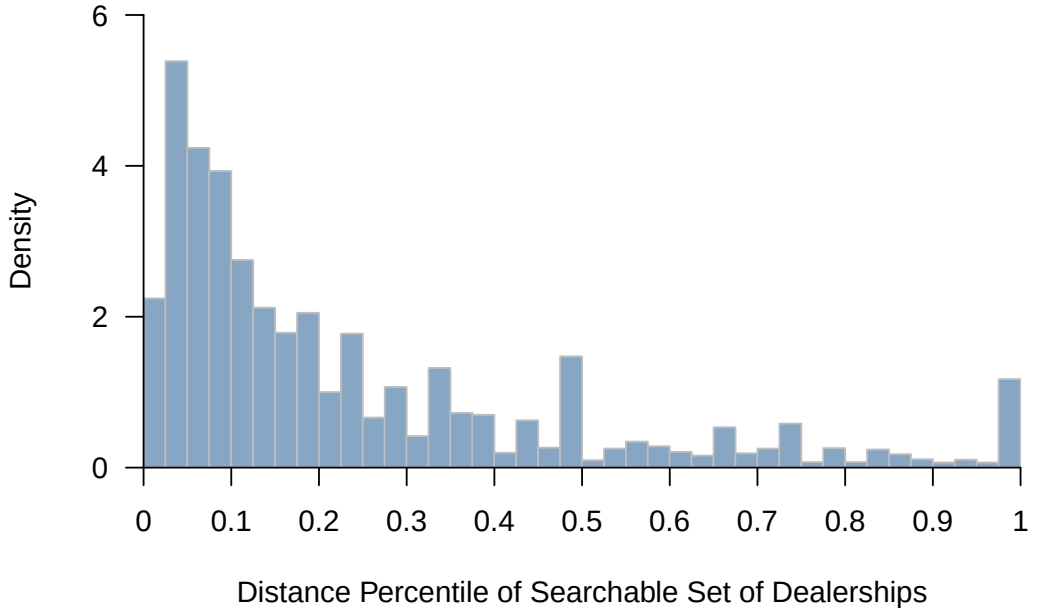
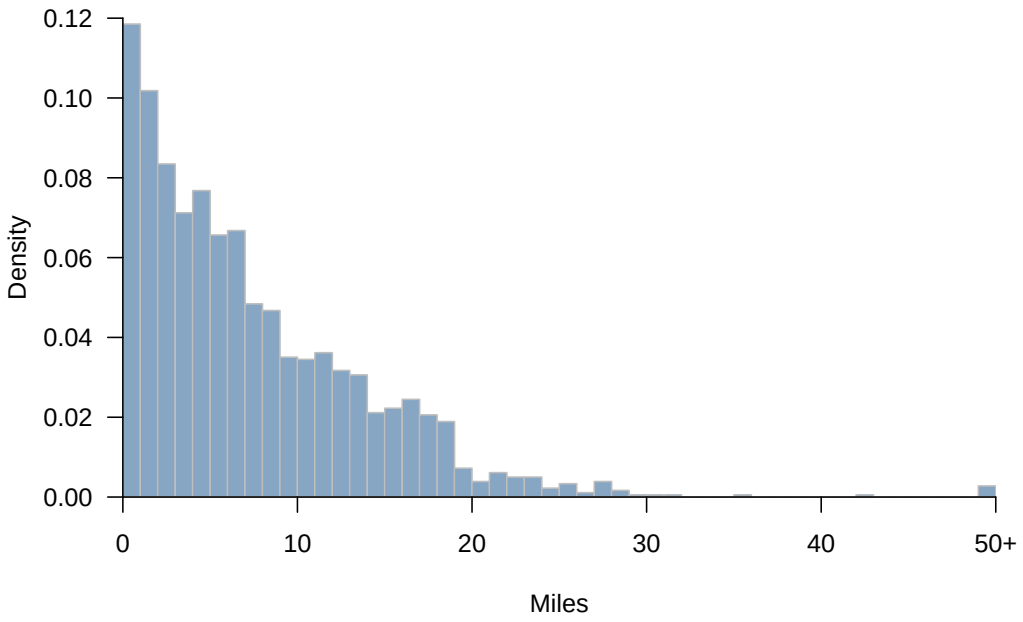


Figure 2.8: Miles Beyond Closest Same-Brand Dealership to Selling Dealership



brand or they can purchase from a more distant dealership carrying the same brand. 72% of consumers in the analysis sample buy from the closest dealer offering the purchased brand, the remaining 28% do not. Figure 2.8 shows the distance (beyond the closest dealership) to the selling dealership for those consumers who did not purchase from the closest dealership offering the ultimately purchased brand. The mean and median "extra" traveled distance are 7.6 and 5.7 miles, respectively. Buying from a more distant dealership suggests that that dealership offers an observed (to the consumer) prior-to-search benefit that exceeds the additional time, travel, and mental costs incurred from making the longer trip.

And lastly, descriptive statistics on dealerships and purchased vehicles are displayed in Table 2.3. Chevrolet and Ford have many more dealerships than Honda, Nissan, or Toyota.[17]

[17]As documented in Murry and Schneider (2016), dealership entry and location result from a complex combination of timing, firm strategy, and the political economy of the U.S. auto market. To briefly summa-

The average vehicle characteristics reported in Table 2.3 are influenced by pickup truck sales. Chevrolet and Ford sell a higher percentage of trucks in the Texas market than the other brands and, as a result, tend to have higher average horsepower and engine sizes, while offering lower gas mileage.

Table 2.3: Descriptive Statistics

			Mean Values of Purchased Vehicles			
	Purchases	Dealerships	MSRP	Horsepower	Engine Size	City MPG
Chevrolet	1,222	212	$34,912	241	3.95	19.9
Ford	1,860	213	$36,670	259	3.31	18.7
Honda	981	56	$26,879	192	2.32	26.9
Nissan	652	64	$29,944	208	2.97	23.8
Toyota	1,796	79	$31,277	212	3.19	22.5

Note: The number of purchases are those in my final analysis sample. MSRP abbreviates manufacturer's suggested retail price; MPG abbreviates miles per gallon.

2.4 Reduced-Form Evidence

In this section, I investigate the set of variables that predict consumer purchase and show that including distance to dealerships is important when modeling demand for cars. I do so by estimating three standard multinomial logit models. Because consumers are assumed

rize, the number of new-car dealerships in the U.S. peaked in 1927 at over 50,000 dealerships (today there are approximately 20,000). These were almost entirely small dealerships selling domestic-branded vehicles (Ford, General Motors, and Chrysler) and they blanketed virtually all urban and rural corners of the country. In the decades that followed, U.S. auto manufacturers became an oligopoly and dealerships responded by growing in size (usually through merging or acquiring nearly dealerships) and through legal protection. Shortly thereafter in the 1950s, the U.S. relaxed import barriers on automobiles and foreign-branded dealerships entered the U.S. market. The relatively-late entry of foreign-branded dealerships and territorial protections imposed by dealership franchise laws led to distinct patterns of dealership sizes and locations. As reported in Lafontaine and Scott Morton (2010), there are many more domestic-branded dealerships than foreign-branded dealership in metropolitan areas (in part a legacy of residential organization prior to suburbanization), and domestic-branded dealerships have many rural locations, whereas foreign brands have few.

to choose a vehicle from a set of vehicles that are of the same Edmund's vehicle type, it is not necessary (or possible) to include vehicle characteristics such as the number of doors or engine type (gas vs electric) as these characteristics do not vary across vehicles within an Edmund's type. Instead, I focus on three major characteristics likely to impact consumer decision-making conditional on type: horsepower, engine size (i.e., displacement measured in liters), and estimated city mileage (MPG). In addition, because my data are from Texas, I include interactions with a dummy variable ("large vehicle") to indicate if the vehicle is a truck or a large SUV (most of which are built on a truck chassis). And lastly, I use MSRP as a measure of price.[18]

In Table 2.4, I show coefficient estimates and implied price elasticities for two full information models (a) and (b), in which the consumer is assumed to have knowledge of all searchable alternatives, and a limited information model (c), in which the consumer only chooses among the products she has searched. The two full information models differ in the included covariates with model (b) additionally including the distance to each dealership.

Across all three multinomial logit models, the coefficient estimates are similar in terms of signs and magnitudes. Coefficient significance is also generally consistent with the exception of MPG in the limited information model (c). Price negatively affects utility, as expected. On average, consumers prefer Toyota and Honda to Nissan, Ford, and Chevrolet for small and medium-sized vehicles. For large vehicles including pickup trucks, consumers prefer Ford. Among the vehicle attributes, engine size and city mileage have the expected positive sign. Horsepower is estimated to have a negative effect on utility. A potential explanation is that, conditional on vehicle type, engine size, and city mileage, vehicles configured to provide extra horsepower may require additional maintenance and are therefore not preferred by consumers.

Recall that model (b) additionally includes distances to dealerships as a covariate. The

[18]While the data include the actual price paid for each purchased vehicle, MSRP is observed for both purchased and not-purchased vehicles. It also better reflects consumers' knowledge of prices prior to search.

Table 2.4: Multinomial Logit Models

	Model (a) Full Information		Model (b) Full Information		Model (c) Limited Information	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
MSRP (in $10,000)	-0.388^{**}	0.034	-0.415^{**}	0.037	-0.441^{**}	0.169
Chevrolet	-1.786^{**}	0.069	-1.662^{**}	0.071	-2.658^{**}	0.450
Ford	-1.283^{**}	0.073	-1.104^{**}	0.075	-2.190^{**}	0.446
Nissan	-0.143	0.073	-0.118	0.076	-0.027	0.378
Toyota	0.069	0.055	0.188^{**}	0.056	-0.332	0.285
Large Vehicle x Chevrolet	1.349^{**}	0.115	1.375^{**}	0.119	2.017^{**}	0.601
Large Vehicle x Ford	1.824^{**}	0.110	1.922^{**}	0.114	2.478^{**}	0.611
Large Vehicle x Toyota	-0.010	0.098	-0.064	0.101	-0.538	0.494
Horsepower (in 100)	-1.013^{**}	0.059	-1.050^{**}	0.061	-0.989^{**}	0.303
Engine Size	0.284^{**}	0.034	0.324^{**}	0.036	0.388^{**}	0.186
MPG	0.085^{**}	0.010	0.095^{**}	0.011	0.027	0.053
MPG x Large Vehicle	0.016^{**}	0.006	0.015^{**}	0.006	0.034	0.031
Distance (in Miles)			-0.219^{**}	0.003		
Number of Consumers	$6,511$		$6,511$		$6,511$	
Number of Products	$175,840$		$175,840$		$7,175$	
Log-Likelihood	$-18,157$		$-13,071$		-381	
BIC	$36,455$		$26,247$		867	
Own-Price Elasticities						
Chevrolet	-1.18		-1.26		-0.88	
Ford	-1.21		-1.30		-0.92	
Honda	-1.00		-1.08		-0.54	
Nissan	-1.40		-1.50		-0.68	
Toyota	-1.20		-1.29		-0.81	

Note: Asterisks indicate statistical significance at the 95% confidence level. The base brand is Honda. BIC $= \log(N) \times k - 2 \times \text{LL}$ where $N = 6,511$ is the number of consumers and k is the estimated number of parameters. Own-price elasticities calculated as $\epsilon_b = (\sum_{i=1}^{N} J_i^b)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{J_i^b} \beta_{\text{MSRP}} \times \text{MSRP}_{ij} \times (1 - p_{ij})$ where i indexes consumers, j indicates the vehicles of brand b with consideration set size J_i^b for consumer i, and p_{ij} is the probability consumer i purchases vehicle j.

distance coefficient is (as expected) negative, large (in absolute terms), and precisely estimated. Moreover, its addition provides substantial improvements in the log-likelihood and Bayesian Information Criterion (BIC) indicating that distance adds to the explanatory power of the model. In addition, price elasticities are more elastic when distance is included.

Comparing models (a) and (b) to model (c), price elasticities are smaller than one (in absolute terms) in the limited information model. The price elasticities are inelastic because of the large number of consideration sets of size one: consumers with such a consideration set would not select a different alternative even if prices increase because they only have one alternative available to them. This is a particular limitation of limited-information discrete choice models, one that can be addressed through a structural search model as described next.

CHAPTER 3

Model and Estimates

3.1 Utility and Search

I model consumer search and purchase decisions using a sequential search model for match value. A product is defined as a combination of a specific vehicle and a dealership (e.g., a Honda Civic at the John Eagle Honda Dealership of Dallas). The match value captures a mix of hard-to-quantify product characteristics that provide a unique, idiosynchratic match (or mismatch) with the consumer. The consumer learns about the match value by visiting a dealership. The match value might, for example, include how much a consumer likes the layout of the dashboard, how well a consumer can see in the car, how much a consumer enjoys driving the car, how courteous and helpful the dealership staff is, etc. Given that I use MSRP as the measure of price in the utility function (prior and post search), the match value also includes any deviation from MSRP due to, e.g., bargaining or a trade-in vehicle.

Consumer $i = 1, \ldots, N$ derives utility from product $j = 1, \ldots, J_i$ with (indirect) utility u_{ij} given by

$$u_{ij} = \delta_{ij} + \varepsilon_{ij} \tag{3.1}$$

with

$$\begin{aligned} \delta_{ij} &= \boldsymbol{x}_j \boldsymbol{\beta} + \eta_{ij}, \\ \eta_{ij} &\sim \mathrm{N}(0,1), \text{ and} \\ \varepsilon_{ij} &\sim \mathrm{N}(0,\sigma). \end{aligned}$$

This framework partitions what the consumer knows and does not know prior to searching. Prior to search, δ_{ij} is known by the consumer. It is composed of a vector of observable vehicle characteristics $\boldsymbol{x}_j$, a vector of consumer preferences for those characteristics $\boldsymbol{\beta}$, and consumer i's product-specific idiosyncratic preferences η_{ij}, which are unobserved by the researcher but known by the consumer prior to search. The second component of utility, ε_{ij}, is the product fit or match value. The consumer knows the distributions of match values, $N(0, \sigma)$, but she is uncertain about her specific match value with each product ε_{ij} and must search to discover it.

Search is performed sequentially and at a cost. Searching a product completely reveals the consumer's match value with that product, but does not reveal information about any other product. The cost of searching a product is parameterized as $c_{ij} = \exp\{\boldsymbol{d}_{ij}'\boldsymbol{\gamma}\}$ to ensure search costs are positive. $\boldsymbol{d}$ is a vector of an intercept, the distance between the consumer's home location and the dealership of the searched product, an urban/rural indicator for the geographic location of the consumer, and an interaction between the latter two variables. $\boldsymbol{\gamma}$ is a parameter vector for the cost covariates. Consumers are assumed to have perfect recall and there is no cost for a consumer to revisit an already-searched product.

My data are conditional on search and purchase and I therefore do not model the outside option of not searching and/or not making a purchase.

3.2 Optimal Consumer Behavior

A consumer searches a product if the marginal benefit of doing so exceeds her marginal search cost. Since the match values ε_{ij} follow $N(0, \sigma)$, prior to search, the utilities u_{ij} follow $N(\delta_{ij}, \sigma)$. Hereafter, I refer to the dispersion parameter of this distribution (σ) as the match-value standard deviation, or MVSD for short. Define u_i^* as the highest utility among the searched products thus far.[1] Conditional on u_i^*, a consumer's expected marginal benefit

[1]For the first search, I set $u_i^* = -\infty$ since my data are conditional on search.

from searching product j is given by

$$\begin{aligned} B_{ij} &= \int_{u_i^*}^{\infty} (u_{ij} - u_i^*)\ f_{u_{ij}}(u_{ij})\ du_{ij} \\ &= \Pr\left[\varepsilon_{ij} > u_i^* - \delta_{ij}\right] \times \mathbb{E}\left[\varepsilon_{ij} - (u_i^* - \delta_{ij}) \,\middle|\, \varepsilon_{ij} > u_i^* - \delta_{ij}\right]. \end{aligned} \tag{3.2}$$

This marginal benefit is the probability that the realized utility for j, u_{ij}, exceeds the best realized utility among the already searched products, u_i^*, multiplied by the expected value of u_{ij} given $u_{ij} > u_i^*$. As shown in Equation (3.2), the marginal benefit depends on the MVSD through the integration over the utility distribution $f_{u_{ij}}(u_{ij})$. Holding everything else constant, a (symmetric) distribution with larger variance has more mass in the tails of the distribution and thus has both a higher probability that the next realized utility will exceed the currently best realized utility and a larger conditional expected value. Thus quantifying the MVSD informs me about the magnitude of the marginal benefit from searching.

Weitzman (1979) derived the rules for optimal behavior under sequential search. The rules involve "reservation utilities" z_{ij}, which are the values that equate the marginal cost and expected marginal benefit of search – that is, z_{ij} is the value for u_{ij}^* that solves $B_{ij} = c_{ij}$ as in Equation (3.2). Kim, Albuquerque, and Bronnenberg (2010) show that there is a semi-closed-form expression for calculating reservation utilities under the assumption of normally distributed match-values:

$$\frac{c_{ij}}{\sigma} = \phi(\zeta_{ij}) - \zeta_{ij} \times (1 - \Phi(\zeta_{ij}))\,. \tag{3.3}$$

I follow the approach of Kim, Albuquerque, and Bronnenberg (2010) by solving ζ_{ij} from the implicit function expressed in Equation (3.3), and then calculating reservation utilities as:

$$z_{ij} = \delta_{ij} + \zeta_{ij} \times \sigma. \tag{3.4}$$

Next, I formally state the optimal sequential search rules developed in Weitzman (1979).

Because the rank of the reservation utilities is a one-to-one mapping with the product index j, I cast the model using j as the order of the reservation utilities such that $j = 1$ is the product with the highest reservation utility for the consumer and $j = J_i$ for the product with the lowest reservation utility. Let me denote the number of searches made by a consumer as K_i. For notational simplicity, I drop the consumer-specific subscript i for the remainder of this section.

Three rules govern consumer search and purchase behavior:

1. *Selection Rule:* A consumer searches products in a decreasing order of reservation utilities, i.e.,

$$z_1 \geq z_2 \geq \ldots \geq z_K \geq \max_{l>K} \{z_l\} . \tag{3.5}$$

2. *Stopping Rule:* A consumer stops searching when the maximum realized utility among the searched products is larger than the maximum reservation utility among the unsearched products, i.e.,

$$\max_{h \leq K} \{u_h\} \geq \max_{l>K} \{u_l\} . \tag{3.6}$$

   Equivalently, at each step during the search process, when a consumer decides to continue searching, the opposite of the Stopping Rule must hold, i.e.,

$$\max_{h<k} \{u_h\} < z_k \qquad \forall k = 2, \ldots, K. \tag{3.7}$$

3. *Choice Rule:* A consumer purchases the product with the highest realized utility among those searched, i.e.,

$$u_{j^*} = \arg \max_{h \leq K} \{u_h\} . \tag{3.8}$$

To understand the implementation of the Weitzman (1979) rules, it can be instructive to walk through an example. Consider again the consumer Mary from Round Rock, Texas. There are 7 dealerships within the Mary's searchable radius carrying in inventory vehicles

of the class for which she is searching. Mary begins by sorting these dealerships in descending order according to their reservation utilities z_{ij}, which are determined by equating the expected benefit of searching each dealership with the cost of visiting it. (See Table 3.1; the values in this example are invented.) Mary then searches the top-rated dealership by reservation utility, First Texas Honda. She receives a (poor) match value of -2.7, which when combined with the pre-search observable value of utility of 9.9, yields a realized utility value of 6.2. Because this realized utility value is lower than the next-best reservation utility value, she searches that next-best dealership, Round Rock Toyota. From this second search, she receives a match value of -3.3 and a realized utility value of 6.7. On the same logic, she continues to search, making her third search at Round Rock Honda. She receives a match value of -2.5 which yields a realized utility value of 7.5. Now, unlike the prior searches, Mary's largest realized utility value (7.5) exceeds the next-best reservation utility value (7.1) and so she stops searching. She purchases the vehicle with the highest realized utility value, the 2016 Honda Odyssey from Round Rock Honda.

Table 3.1: Example Consumer Search Process

Rank	Year-Make-Model	Dealership	z_{ij}	$x'_{ij}\beta$	ε_{ij}	u_{ij}
1	2016 Honda Odyssey	First Texas Honda	11.1	9.9	-2.7	6.2
2	2016 Toyota Sienna	Round Rock Toyota	10.7	10.0	-3.3	6.7
3	2016 Honda Odyssey	Round Rock Honda	9.4	10.0	-2.5	7.5
4	2016 Honda Odyssey	Howdy Honda	7.1	8.2	–	–
5	2016 Nissan Quest	Clay Cooley	6.9	7.1	–	–
6	2016 Nissan Quest	Round Rock Nissan	5.0	5.0	–	–
7	Toyota Sienna	Toyota South Austin	4.2	3.3	–	–

It is worth noting that Mary was not required to purchase the vehicle from the last-searched dealership. If her match value draws were (-2, -2, -5) instead of (-2.7, -3.3, -2.5), then she would still have made three searches, but now she would have purchased from Round

Rock Toyota, the dealership she searched second. It is also worth noting that negative match value draws are not necessary for a consumer to continue their search. If Mary's first match value draw was 0.5, her realized utility (10.4) would not exceed the next-best reservation utility (10.7) and so she would continue her search.

3.3 Likelihood Function

The probability that a specific sequence of searches and an ultimate purchase are made by a consumer is the probability that each of the Weitzman (1979) rules holds at their respective steps in the consumer's shopping and purchase process, i.e.,

$$
\begin{aligned}
&L_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma; \boldsymbol{x}, \boldsymbol{d}) \\
&\quad = \text{Prob}\big[\text{Consumer } i \text{ selects } j_i = 1 \text{ on the first search,} \\
&\qquad\qquad \text{Continues to search and selects } j_i = k \text{ on the } k^{\text{th}} \text{ search for } k = 2, \ldots, K_i, \\
&\qquad\qquad \text{Stops searching after the } K_i^{\text{th}} \text{ search, and purchases } j_i^*\big] \\
&\quad = \int \mathbb{1}\big[z_{ij} \geq \max_{h<j}\{u_h\} \text{ for } j = 2, \ldots, K_i \bigcap z_{ij} = \arg\max_{k>j}\{z_{ik}\} \text{ for } j = 1, \ldots, K_i \bigcap \\
&\qquad\qquad \max_{h \leq K_i}\{u_{ih}\} \geq \max_{k>K_i}\{z_{ik}\} \bigcap u_{ij_i^*} = \arg\max_{h \leq K_i}\{u_{ih}\}\big]\, dF(\eta, \varepsilon).
\end{aligned}
\tag{3.9}
$$

The model likelihood is the product of the N individual likelihoods, i.e.,

$$
L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma; \boldsymbol{x}, \boldsymbol{d}) = \prod_{i=1}^{N} L_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma; \boldsymbol{x}, \boldsymbol{d}). \tag{3.10}
$$

Note that I parametrize the MVSD as $\theta = \log(\sigma)$ during estimation.

Neither the search nor purchase probabilities can be expressed in closed form. I approximate the integrals in the likelihood function with averages using logit-smoothed accept-reject simulation. This simulated maximum likelihood estimation algorithm follows Train (2009).

Details and R code for estimation are provided in Appendix A.

3.4 Identification

The parameters to be estimated include the preference parameters $\boldsymbol{\beta}$, search cost parameters $\boldsymbol{\gamma}$, and the MVSD σ.

The preference parameters $\boldsymbol{\beta}$ are identified by purchase frequency, search order, and the number of searches. In much the same way that the purchase decision among a set of products identifies preference parameters in a traditional discrete choice model, the purchase decision among searched products identifies preference parameters in a search model. In addition, because consumers search in order of decreasing reservation utilities (recall, $z_{ij} = \delta_{ij} + \zeta_{ij} \times \sigma$), holding everything else constant, products with higher δ_{ij} have higher reservation utility values and thus are ranked higher by search order and searched more frequently.

Identification of the search cost parameters $\boldsymbol{\gamma}$ depends on the specification of the search cost function and the parametric assumptions of the model. With *no* exogenous search cost shifter (e.g., with fixed cost of search) and no parametric assumption for the distribution of ε, only the ratio of c/σ can be identified. The reason is that only the number of searches informs c and σ, but the search order does not. To see this, recall that consumers search in decreasing order of reservation utilities which are a function of search cost and the MVSD σ (see Equation (3.4) reproduced in the paragraph above). If there is a common search cost across products and a common MVSD, then ζ_{ij} is the same for all products and the search order is entirely driven by δ_{ij}.

Conversely, without an exogenous search cost shifter but *with* a parametric assumption on the match value distribution, search cost c and MVSD σ are separately identified by parametric form. This can be seen in the implicit function of Equation (3.3): $c_{ij}/\sigma = \phi(\zeta_{ij}) - \zeta_{ij} \times (1 - \Phi(\zeta_{ij}))$. If σ were only part of the left-hand side, then c and σ would not be separately identified, i.e., only their ratio would be parametrically identified, as discussed

above. However, σ is also part of the right-hand side of Equation (3.3) because $\zeta_{ij} = (z_{ij} - \delta_{ij})/\sigma$ and this ensures parametric identification. In practice, despite making parametric assumptions on the match value distribution, previous empirical research has commonly fixed σ to 1 and only estimated c (e.g., Kim, Albuquerque, and Bronnenberg 2010, Chen and Yao 2017, Honka and Chintagunta 2017, Kim, Albuquerque, and Bronnenberg 2017, Ursu 2018). The results of my simulation study, presented next, suggest that substantial data would be required to estimate c and σ when no cost shifter is present, and thus potentially rationalize the choice by prior researchers to fix σ to 1.

By contrast, *with* an exogenous search cost shifter and a parametric assumption on the match value distribution, (i.e., distance to dealerships in my empirical application), both the search cost c_{ij} (and its governing parameters $\boldsymbol{\gamma}$) and the MVSD σ can be separately identified. The reason is that both the number of searches and the search order inform c and σ which, in turn, influence reservation utilities (see Equation (3.4)). For example, higher search costs yield lower ζ_{ij} values. Thus a consumer who must incur higher cost to search a product will assign it a lower reservation utility, rank it lower in the search order, and stop her search earlier (on average) than an otherwise identical consumer facing an identical searchable set of products but with lower search cost. Thus the observable differences in search cost due to the exogenous search cost shifter (coupled with the patterns of search order and search length) identify the search cost parameters $\boldsymbol{\gamma}$.

Lastly, a higher MVSD (σ) assigns a larger value to the second component of the reservation utility (see Equation (3.4)) by increasing the mass in the upper tail of the utility distribution. With a larger MVSD, search terminates later because, holding everything else constant, a larger MVSD results in a higher marginal benefit from an additional search (see Equation (3.2) and Figure 3.1 below) and thus higher reservation utilities increase the probability that consumers continue searching. As a result, the MVSD is identified by the extent to which the search order is driven by variance (rather than the mean) of the distribution of utility.

3.5 Simulation Study

In this section, I describe the results from a simulation study in which I demonstrate that my estimation approach recovers preference and cost parameters as well as the MVSD parameter if an exogenous search cost shifter is included in the estimation.

For the simulation study, I generate 5,000 consumers each searching up to five brands. The number of searchable dealerships is drawn from a combination of a chi-squared distribution and an exponential distribution so as to generally mimic the observed searchable set sizes in the empirical application. These simulated searchable set sizes range from 1 to 40 with a median of 11 and a mean of 12. Observable characteristics include the brand as well as price and city mileage; the latter two are drawn from uniform distributions on the interval -2 to 2. Distances to dealerships are drawn from the absolute value of the sum of a chi-squared distribution with 8 degrees of freedom and a normal distribution with mean zero and standard deviation 16. See Appendix B for the R code used for simulation.

Simulated distances to the searchable alternatives range from 0.01 miles to 102.4 miles with a median of 15.5 and a mean of 18.2 miles. For estimation, I use 1,000 draws from the distributions of the consumers' idiosyncratic preferences and match-value terms and smoothing parameters of (15, 15, 15, 5). I replicate each estimation 50 times and report mean coefficient estimates, their standard deviation, and average standard errors. The results are shown in Table 3.3a and show that, for a search model, the parameters are recovered well.

For a point of comparison, I repeat the exercise without a search cost shifter. The results are presented in Table 3.3b. Although the cost and benefit parameters are parametrically identified, they are estimated imprecisely here. For example, the search cost intercept has an asymptotic standard error greater than 10. Significantly more data would be required to obtain accurate estimates. A comparison of these two simulation results highlights the value that a search cost shifter can bring to the estimation of a sequential search model.

Table 3.2: Simulation Study

(a) With Exogenous Cost Shifter

	True Values	Average Estimate	Standard Deviation of Estimates	Average Asymptotic Standard Error
Brand 1	0.2	0.178	0.088	0.038
Brand 2	0.4	0.374	0.064	0.045
Brand 3	0.6	0.540	0.066	0.042
Brand 4	0.8	0.714	0.080	0.038
MSRP	-0.5	-0.443	0.023	0.012
MPG	0.5	0.445	0.023	0.011
Cost Intercept	-2.0	-2.623	0.167	0.078
Distance	0.3	0.347	0.015	0.007
Log Sigma	0.69	0.439	0.059	0.024

(b) Without Exogenous Cost Shifter

	True Values	Average Estimate	Standard Deviation of Estimates	Average Asymptotic Standard Error
Brand 1	0.2	0.177	0.055	0.027
Brand 2	0.4	0.344	0.057	0.030
Brand 3	0.6	0.522	0.055	0.025
Brand 4	0.8	0.689	0.056	0.027
MSRP	-0.5	-0.429	0.016	0.008
MPG	0.5	0.428	0.014	0.008
Cost Intercept	0	-20.900	6.499	10.983
Log Sigma	0.69	-3.976	0.309	0.386

Note for both tables: Simulations based on 50 repetitions of 5,000 consumers searching 5 brands. Estimations uses 1,000 draws to approximate the integral over idiosyncratic preferences and match-value terms. Smoothing parameter vector set to $\lambda = (15, 15, 15, 5)$.

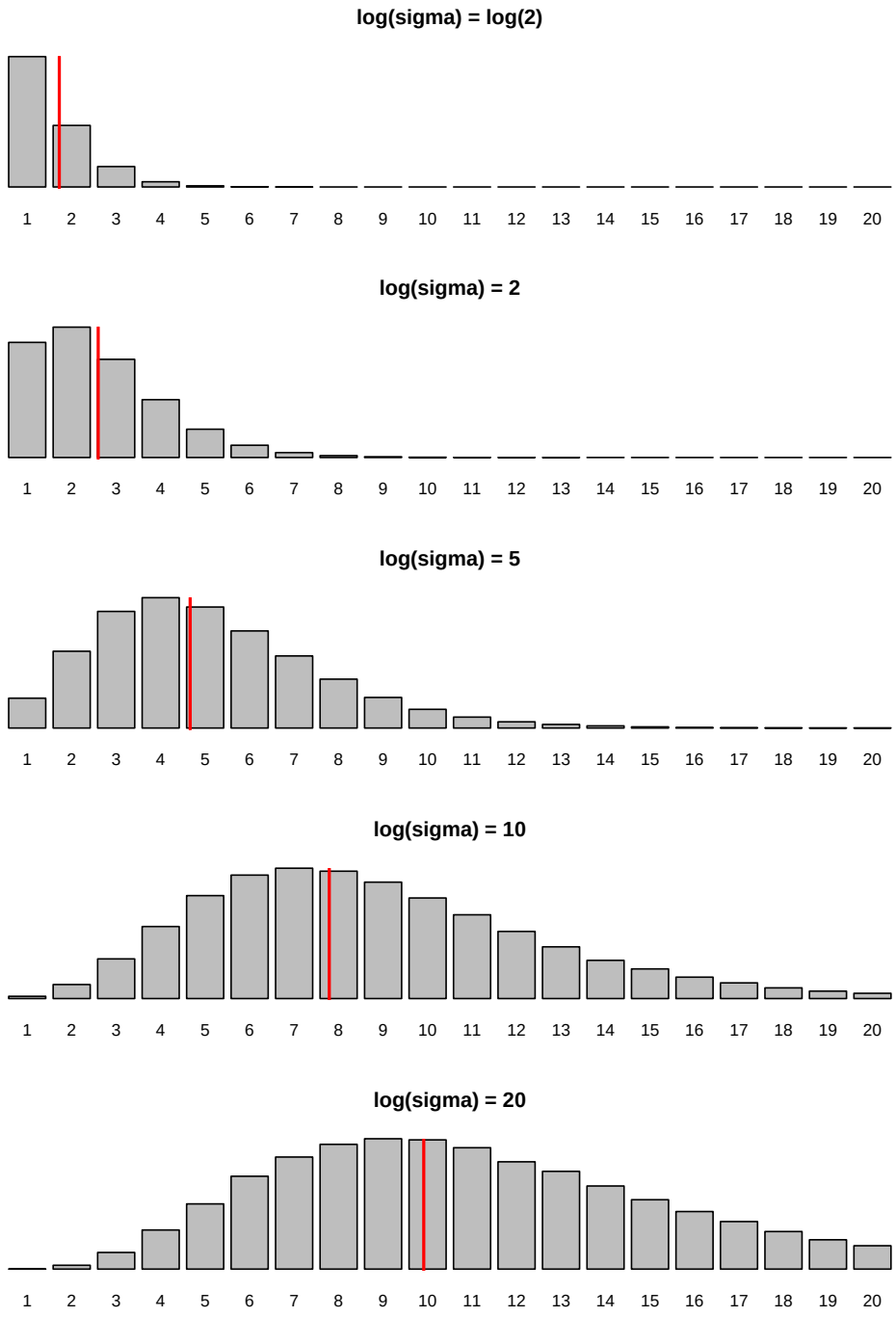
In addition, to illustrate the relationship between the MVSD and optimal search behavior, I simulate five additional groups of data sets. Each group contains 50 replications and the data generation process follows what is outlined above. However, across all generated data sets, I hold all characteristics constant except for the MVSD. The MVSD is parameterized as $\log(\sigma)$ and set to the following values for each group of data sets, respectively: $\log(2)$, 2, 5, 10, and 25. Using these values, I simulate the number of searches a consumer makes when optimally searching sequentially using the Weitzman (1979) rules and then I average across the 50 replications. The five average search-length distributions (where I average across sets of 50 replications) are shown in Figure 3.1 with red bars to indicate mean values. The figure shows that an increase in the MVSD yields an increase in the average number of searches and results in a search-length distribution with a longer tail. These results support the argument above that the number of searches helps identify the MVSD, and will be a useful reference when discussing model fit in Section 4.1.

3.6 Empirical Results

I show the results from three search model specifications in Table 3.4. In model (i), I fix the MVSD to 1 and estimate a fixed cost of search. In model (ii), I continue to fix the MVSD to 1, but allow search cost to vary with distance, an urban dummy, and an interaction between both variables. In model (iii), I allow the MVSD σ to enter the model as a parameter to be estimated. In all three model specifications, I include the same set of covariates as in the reduced-form choice models in Section 2.4. Across the three model specifications, the utility parameter estimates are similar, generally sharing the same sign, similar magnitude, and significance with the exception of the brand intercepts for Nissan and Toyota. The utility parameter estimates also generally similar to the results from the reduced-form choice models presented in Table 2.4.

Comparing models (i) and (ii) in Table 3.4, the addition of distance, an urban dummy,

Figure 3.1: Simulated Search Patterns for Various Values of the Match Value Standard Deviation



Note: For each value of $\log(\sigma)$, 50 data sets are simulated. The search-length distribution is tabulated for each data set, and then averaged across the 50 data sets. That average is then plotted. The procedure is repeated for each value of $\log(\sigma)$. Red vertical bars indicate the average search length.

Table 3.4: Search Model Results

	Model (i)		Model (ii)		Model (iii)	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Preference Parameters						
MSRP (in $10,000)	-0.182^{**}	0.018	-0.186^{**}	0.019	-0.202^{**}	0.020
Chevrolet	-0.807^{**}	0.035	-0.787^{**}	0.034	-0.803^{**}	0.038
Ford	-0.602^{**}	0.037	-0.579^{**}	0.039	-0.578^{**}	0.040
Nissan	-0.068^{**}	0.038	-0.063	0.040	-0.073	0.039
Toyota	0.047	0.031	0.072^{**}	0.030	0.097^{**}	0.033
Large Vehicle x Chevrolet	0.611^{**}	0.059	0.611^{**}	0.060	0.670^{**}	0.066
Large Vehicle x Ford	0.886^{**}	0.056	0.823^{**}	0.060	1.020^{**}	0.064
Large Vehicle x Toyota	0.028	0.053	0.011	0.054	0.046	0.058
Horsepower (in 100)	-0.473^{**}	0.030	-0.501^{**}	0.031	-0.518^{**}	0.034
Engine Size	0.132^{**}	0.019	0.146^{**}	0.018	0.151^{**}	0.021
MPG	0.042^{**}	0.005	0.042^{**}	0.006	0.044^{**}	0.006
MPG x Large Vehicle	0.005	0.003	0.006	0.003	0.004	0.004
Log Search Cost Parameters						
Intercept	0.361^{**}	0.018	0.273^{**}	0.043	2.250^{**}	0.040
Urban			0.203^{**}	0.048	-0.055^{**}	0.015
Distance (in Miles)			0.017^{**}	0.001	0.006^{**}	0.000
Urban x Distance			0.002^{**}	0.001	0.003^{**}	0.000
Match-Value Standard Deviation						
Sigma	1.00		1.00		8.16	0.310
Number of Consumers	6,511		6,511		6,511	
Number of Products	175,840		175,840		175,840	
Log-Likelihood	$-20,765$		$-18,362$		$-17,652$	
BIC	41,645		36,872		35,453	

Note: Asterisks indicate statistical significance at the 95% confidence level. The base brand is Honda. Optimization via BFGS with relative tolerance convergence criterion set to 1e-6. Simulated likelihood using Q=1,000 independent random draws of the random utility error and the match-value distribution. Search costs are parameterized as $c_{ij} = \exp\{\gamma_0 + \gamma_1 \text{Urban} + \gamma_2 \text{Distance} + \gamma_3 \text{Urban} \times \text{Distance}\}$. Delta Method used to calculate standard error of σ. BIC $= \log(6511) \times k - 2 \times \text{LL}$ where k is the estimated number of parameters.

and an interaction between both variables to the search cost function leads to a much larger log-likelihood value (a change of over 2,400). This improvement is driven by a better ability of the model to fit the search order during the estimation process because consumers tend to visit dealerships close to their homes. When distance is not included in the model, two dealerships are equivalent from a modeling perspective if they offer a vehicle with the same characteristics even if one dealership is located next to the consumer's home and the other dealership is located 100 miles away. All search cost parameter estimates are positive and statistically significant. Comparing urban and rural consumers, urban consumers face higher cost than rural consumers, and that cost difference increases with distance. For example, the cost (in utils) to a rural consumer to visit a dealership 1 mile away is 1.34, while the cost to an urban consumer to do the same is 1.64 (23% higher). At a distance of 20 miles, the rural consumer's cost is 1.43 and the urban consumer's cost is 1.77 (28% higher).

In model (iii), my main model, I additionally estimate the MVSD. The improvement of over 700 in the log-likelihood is large and shows that model (iii) fits the data better than model (ii). In addition, the correlation between $\log(\sigma)$ and the search cost intercept is 0.95. If a correlation between two parameters equals ± 1.0, then those two parameters are not separate identified. While this value is high, it is not high enough to suggest concern for identification. The correlations between all other estimated coefficients lie within ± 0.8 with most being smaller than ± 0.4. See Table 3.5.

Moreover, lack of identification between parameters would be observable as a ridge in the log-likelihood – that is, a region with equivalent likelihood values. In Figures 3.2a and 3.2b, I show bivariate plots of the log-likelihood in the space of the search cost intercept (γ_0) and the MVSD (σ) at two scales of perspective. (The white area in the contour plots contain log-likelihood values that are worse than the dark blue area.) The plots show no ridge. Rather, there is a unimodal mountain with a unique maximum, indicating the separate identification of these parameters.[2]

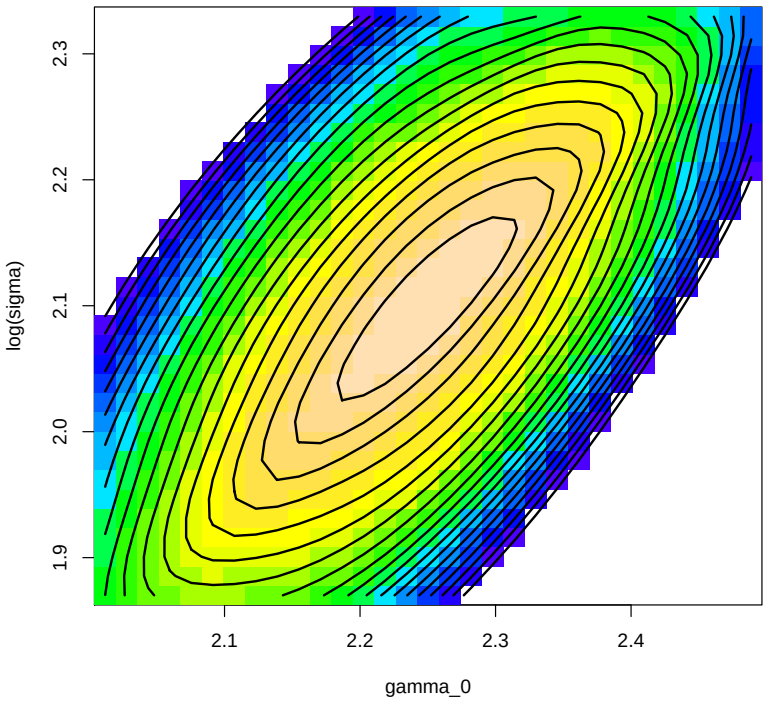
[2]Simulation error and smoothing are two potential concerns with the evidence shown for parametric

Table 3.5: Correlation Matrix of Coefficient Estimates for Search Model (iii)

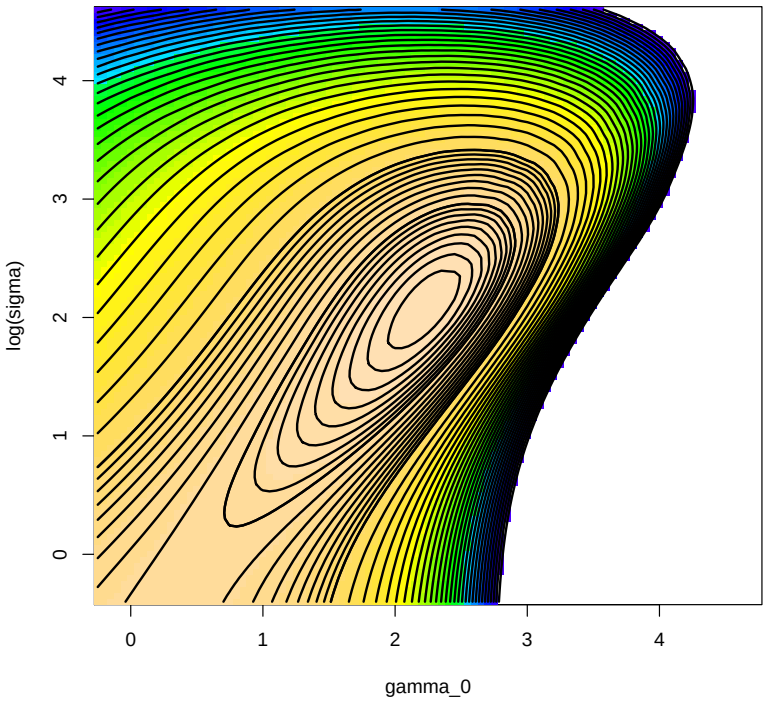
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	MSRP	1.00	0.17	0.25	-0.13	0.07	0.08	0.05	0.19	-0.09	-0.21	-0.02	-0.08	0.08	-0.02	-0.07	-0.02	0.09
2	Chevrolet	–	1.00	0.53	0.15	0.53	0.32	0.08	0.36	-0.44	-0.20	-0.14	0.00	0.16	0.02	-0.13	-0.05	0.17
3	Ford	–	–	1.00	0.20	0.58	0.11	0.29	0.67	-0.21	-0.45	-0.11	0.08	-0.05	0.03	0.05	-0.01	-0.05
4	Nissan	–	–	–	1.00	0.23	0.06	-0.15	0.06	0.39	0.31	0.37	-0.09	0.02	0.00	0.00	-0.03	0.02
5	Toyota	–	–	–	–	1.00	0.15	0.06	0.40	-0.15	-0.19	-0.39	0.03	0.15	0.00	-0.11	-0.04	0.16
6	LV x Chevy	–	–	–	–	–	1.00	-0.40	0.21	0.04	-0.05	0.12	-0.08	0.38	0.03	-0.30	-0.14	0.41
7	LV x Ford	–	–	–	–	–	–	1.00	0.37	-0.23	0.02	-0.22	0.05	-0.01	0.05	0.01	-0.02	0.00
8	LV x Toyota	–	–	–	–	–	–	–	1.00	-0.14	-0.24	-0.01	0.05	0.04	-0.01	-0.02	-0.03	0.05
9	Horsepower	–	–	–	–	–	–	–	–	1.00	0.69	0.64	-0.38	0.08	0.02	-0.05	-0.04	0.09
10	Engine Size	–	–	–	–	–	–	–	–	–	1.00	0.57	-0.42	0.21	0.00	-0.16	-0.05	0.23
11	MPG	–	–	–	–	–	–	–	–	–	–	1.00	-0.42	-0.15	0.01	0.11	0.04	-0.14
12	MPG x LV	–	–	–	–	–	–	–	–	–	–	–	1.00	-0.10	-0.01	0.08	0.02	-0.11
13	Cost Int.	–	–	–	–	–	–	–	–	–	–	–	–	1.00	-0.29	-0.78	-0.15	**0.95**
14	Urban	–	–	–	–	–	–	–	–	–	–	–	–	–	1.00	0.25	-0.39	-0.05
15	Distance	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.00	-0.35	-0.73
16	Urban x Dist.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.00	-0.25
17	Log-Sigma	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.00

Figure 3.2: Log-Likelihood Bivariate Contour Plots

(a) Zoom In Near Maximum



(b) Zoom Out



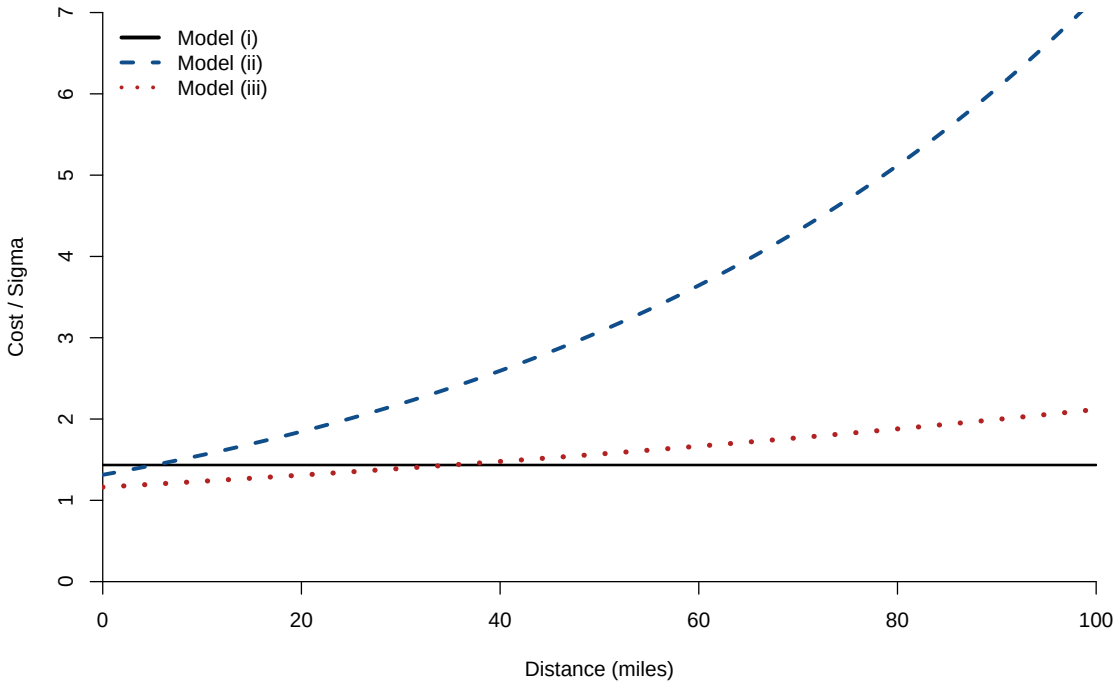
The MVSD estimate is 8.16. This estimate is large compared to the common practice of setting σ to 1. This large MVSD estimate indicates that consumers gain substantial benefits when visiting car dealerships. The MVSD has an estimated standard error of 0.31 and a 95% confidence interval (3.88, 17.19).[3] Note that the interval does not include 1.

The parameter estimates for model (iii) also indicate that fixing the MVSD to 1 imposes a bias on the other parameters, in particular, on the cost parameters. More specifically, I find that the cost intercept is smaller, the urban intercept has a different sign, and the coefficient on distance is estimated to be almost three times as large in model (ii) as in model (iii). To demonstrate the differences visually, in Figure 3.3, I plot c_{ij}/σ for a rural consumer over distances ranging from 0 to 100 miles for all three search models. The c_{ij}/σ value is substantially different between model (ii) and (iii) for almost all positive distances. Thus I conclude that it is important to estimate the MVSD to correctly recover search cost parameters.

identification. Regarding simulation error, I use 1,000 draws in the estimation, a much larger number of draws that in previous literature. In addition, I have re-estimated the model using 10,000 draws and found the coefficient estimates and log-likelihood values to barely change. This leads me to conclude that simulation error is not a concern. Regarding smoothing, I have also estimated the model using smoothing factor λ set to 6, 7, and 8, and confirmed that the bivariate (cost intercept and MVSD) log-likelihood contour plots show only a unimodal mountain, leading me to conclude that smoothing is not a concern for identification.

[3]I use the Delta Method to calculate the standard error for $\hat{\sigma}$. Specifically, I estimate $\theta = \log(\sigma)$. Thus, $\text{StdErr}(\hat{\sigma}) = \exp(\hat{\theta}) \times \text{StdErr}(\hat{\theta})$.

Figure 3.3: Comparison of Cost-Sigma Ratio vs Distance Across Fitted Search Models



CHAPTER 4

Post-Estimation Analyses

4.1 Model Fit

I evaluate the in-sample predictive performance of the estimated models using several measures. First, in Table 4.1a, I show the product and brand hit rates for four models: full information multinomial logit model (model (b) in Table 2.4) and the three search models (models (i) to (iii) in Table 3.4). The "hit rate" is the percent of time that simulated behavior matches observed behavior. For example, the product hit rate is the average percent of time that the simulated-to-be-purchased product is the same as the actually purchased product. I calculate the hit rates as follows: for each consumer, I simulate 500 vectors of the random utility errors $\boldsymbol{\eta}_i$. For each simulation, I calculate the consumer's reservation utilities for each searchable product and allow search and choice to progress following Weitzman (1979) rules. For each consumer and each simulation, I record the predicted search length, the searched product(s), and the chosen product as well as its brand.

The results show that all three search models considerably outperform the multinomial logit model (model (b) in Table 2.4). Among the three search models, added flexibility yields better hit rates. The improvement is the largest when search costs are modeled as a function of distance and the urban dummy (model (ii) in Table 3.4); the improvement in brand hit rate is marginal when the MVSD is estimated (model (iii) in Table 3.4).

Additionally, Table 4.1b displays the distribution of the number of searches consumers make (observed in my data) and the distributions of the number of searches that are predicted

Table 4.1: In-Sample Predictive Performance

(a) Hit Rates

Hit Rates in %	Model (b)	Model (i)	Model (ii)	Model (iii)
Product Hit Rate	8.4	10.3	14.9	19.4
Brand Hit Rate	30.0	39.4	39.8	39.9

(b) Search Distributions

Number of Searches	Observed (Data)	Model (i) (Predicted)	Model (ii) (Predicted)	Model (iii) (Predicted)
1	91.1%	96.2%	98.7%	88.5%
2	7.8%	3.7%	1.3%	10.3%
3	0.9%	0.1%	0.1%	1.1%
4	0.2%	-	-	0.1%
5	0.1%	-	-	0.1%
6+	0.1%	-	-	-

by the three search models from Table 3.4. While allowing search cost to vary with search cost shifters (model (ii)) enables the model to more precisely predict which products (dealerships) consumers search, model (ii) does not predict the distribution of the number of searches well: it overpredicts the proportion of consumers who search once and underpredicts the proportion of consumers who make two or more searches, i.e., the tail of the distribution of the number of searches. While the maximum number of searches consumers conduct is 6 in my data, model (ii) predicts that the maximum number of searches is 3. Model (iii) – in which the MVSD is estimated – provides by far the best fit to the observed distribution of number of searches. In other words, permitting the MVSD to be freely estimated allows the model to fit the data better – in particular, the tail of the distribution of the number of searches. Model (iii) predicts the maximum number of searches to be 5, while the maximum number of search in my data is 6.

To summarize, my results show that estimating the MVSD is beneficial to correctly

predict product choice and crucial to correctly predict the distribution of the number of searches, especially the tail of the search distribution.

4.2 Price Elasticities

I show the implied own-price elasticities for all three search models in Table 4.2. Elasticities are calculated by first simulating search and choice behavior from the fitted model. Then, separately for each brand, I increase the prices of available alternatives by 10% and re-simulate search and choice behavior. I repeat this exercise 500 times for each consumer. Elasticities are computed as the average difference in percent brand choice between the simulated outcomes with and without a price increase, scaled to reflect a 1% price increase.

Table 4.2: Own-Price Elasticities

	Model (i)	Model (ii)	Model (iii)
Chevrolet	−0.79	−0.78	−0.75
Ford	−0.69	−0.68	−0.67
Honda	−0.52	−0.58	−0.56
Nissan	−0.72	−0.72	−0.71
Toyota	−0.62	−0.63	−0.60

The own-price elasticity estimates range from -0.6 to -0.8. The elasticity estimates are influenced by the assumption that consumers make search and purchase decisions only among 5 brands and conditional on having already selected the type of vehicle they will purchase. While these assumptions were useful in easing the computational burden of fitting the model, they restrict the interpretability of the elasticity estimates because, for example, consumers may substitute to other brands of the same Edmunds type and, at some level of price increase, consumers will likely not simply substitute to other vehicles in the same Edmunds type, but will substitute to other vehicle types.

To the extent that managerial pricing decisions are driven by elasticity estimates, the consistency of the estimates across models suggests that, for pricing decisions, there is little benefit of including the MVSD in the model. However, as discussed and explored in Section 4.4, there are other non-pricing benefits.

4.3 Consumer Surplus

While the MVSD is a direct measure of the magnitude of potential benefits achievable through search, I also calculate expected consumer surplus as an alternative measure of achieved benefits through search. For an individual consumer, the consumer surplus is defined as the expected utility from the chosen product net of all costs (price and search costs), i.e.,

$$\mathbb{E}\left[CS_i\right] = \int_{\varepsilon}\int_{\eta}\left(u_{ij^*} - \sum_{j=1}^{K_i} c_{ij}\right) dF(\eta)\, dF(\varepsilon). \tag{4.1}$$

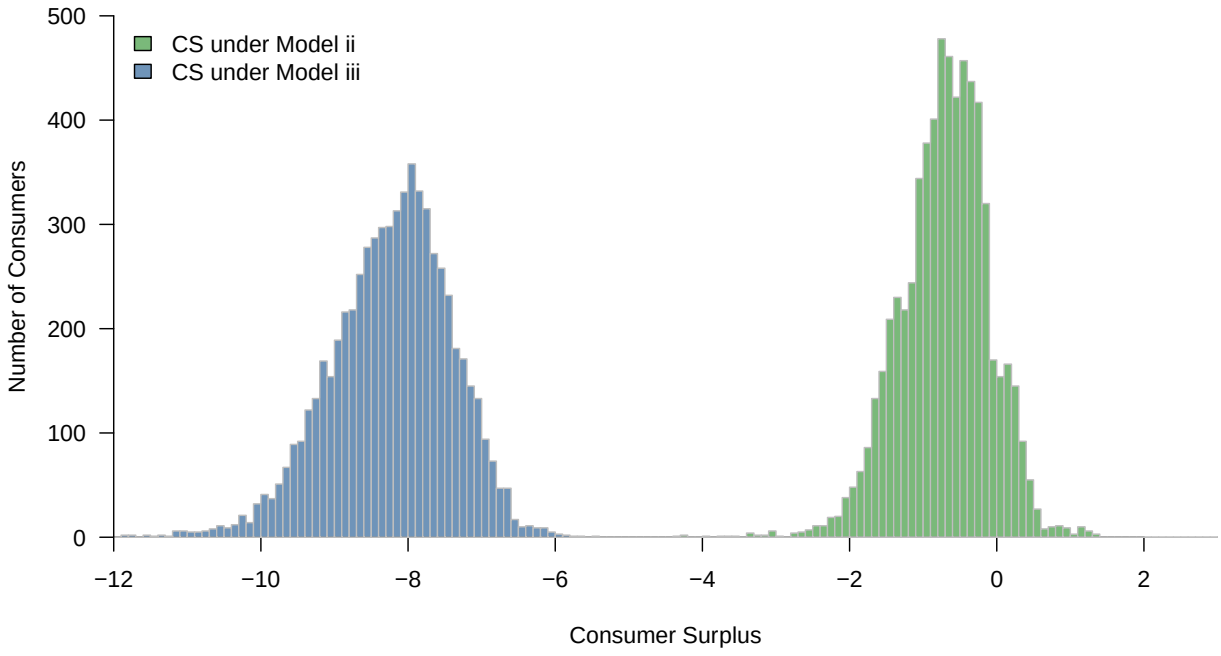
Note that, because I do not model the no-search and no-purchase decisions as a result of the fact that my data are conditional on search and on purchase, my consumer surplus estimates taking both prices and search costs into account can be negative (see, e.g., Moraga-Gonzalez, Sandor, and Wildenbeest 2017). In particular, because I do not model an outside option, utility values from my model are interval (not ratio) values – that is, utility values are relative to each other, but do not have a fixed point (e.g., a zero utility value) and thus may be "shifted" when compared to a similar model that includes the outside options of no-search and/or no-purchase. To calculate consumer surplus for a consumer, I simulate 500 realizations from a model and average the resulting consumer surplus estimates for each consumer:

$$\widehat{CS_i} = \frac{1}{Q}\left(\sum_{q=1}^{Q}\left(u^q_{ij^*} - \sum_{j=1}^{K_i^q} c^q_{ij}\right)\right). \tag{4.2}$$

Figure 4.1 shows the distributions of consumer surplus from models (ii) and (iii). The

mean consumer surplus estimates for models (ii) and (iii) are -0.7 and -8.0, respectively. Thus normalizing the MVSD to 1 severely *overstates* consumer surplus.

Figure 4.1: Consumer Surplus Based on Models (ii) and (iii)



To assess the relationship between consumer surplus and consumer characteristics, the consumer surplus from model (iii) is regressed on a set of consumer demographics. The results are displayed in Table 4.3. The included demographic variables explain a large proportion of the variation in the consumer surplus, i.e. 38.5%. I find significantly larger consumer surplus for urban and non-white consumers and significantly smaller consumer surplus for older consumers and consumers with kids.

4.4 Counterfactual

In 2017, Hyundai initiated a program called "Hyundai Drive" in which a consumer can schedule a test drive of a Hyundai vehicle at a location that is convenient to her and a dealer

Table 4.3: Consumer Surplus and Demographics

	(1)	(2)
Urban Indicator	0.836**	0.717**
	(0.022)	(0.029)
Median Age 30 - 40	−0.011	−0.063**
	(0.024)	(0.024)
Median Age 40 +	−0.103**	−0.135**
	(0.030)	(0.030)
Percent Male	−0.053	−0.048
	(0.181)	(0.172)
log(Number of Kids)	−0.069**	−0.081**
	(0.010)	(0.010)
Percent with College Degree	0.236**	−0.031
	(0.052)	(0.052)
log(Income in $1,000)	0.015**	0.001
	(0.008)	(0.007)
Unemployment Rate	0.233	0.167
	(0.210)	(0.204)
Percent Black	0.670**	0.329**
	(0.070)	(0.073)
Percent Other Race	0.960**	0.684**
	(0.081)	(0.082)
Intercept	−8.862**	−8.137**
	(0.130)	(0.136)
Month Fixed Effects	No	Yes
County Fixed Effects	No	Yes
Number of Observations	6,511	6,511
R^2	0.276	0.385

Note: Standard errors reported in parentheses. Asterisks indicate statistical significance at the 95% confidence level.

will bring the car to that location. This program reduces travel costs to a dealer, but does not eliminate all search costs as it takes time to perform a test drive and there continue to be mental costs of considering an additional alternative. Here, I assess the potential market share changes to each brand from adopting at-home test drives as well as the search and purchase-conditional-on-search decisions that lead to market share changes.

To do so, I first assume that brands adopt at-home test drives unilaterally. The brand offering the program permits the consumer to select a vehicle from that brand's closest dealership to be brought to the consumer for inspection and a test drive. Thus, the consumer incurs no travel costs to "visit" that dealership. To calculate the impact of adopting at-home test drives, I first simulate search and choice behavior for each consumer 500 times. Then, for each brand, I set the distance of the closest dealership for each consumer to zero and re-simulate search and choice behavior. I then calculate the average differences in search, choice, and choice conditional on search between the base simulation to the counterfactual simulation. Results are presented in Table 4.4 where the top half of the table report results under Model (iii) estimates while the bottom half reports results under Model (ii) estimates.

Unilateral implementation of at-home tests drives permits brands to capture an additional 3.9–7.7 percentage points of the analyzed market. Because the five brands under study account for 60.1% of the Texas auto market, this corresponds to overall market share changes of 2.3–4.6 percentage points (holding constant the choices made by consumers who purchase other brands). This program yields the smallest market share increases for Chevrolet and Nissan and the largest market share increase for Toyota.

As a point of comparison, I conduct the same exercise with the estimates from model (ii) in which the MVSD is fixed to 1. I find the results to be roughly half as large as those from model (iii) in which the MVSD is estimated. Specifically, estimated increases in market shares under model (ii) range from 0.8 to 1.6 percentage points. Thus, failing to estimate MVSD yields underestimated market share increases from the unilateral implementation of at-home test drives.

I then assess three additional counterfactuals. In the first two, two brands implement at home test drives simultaneously. In the first, Chevrolet and Ford implement at-home test drives, while in the second, Honda and Toyota implement them. In the final counterfactual, all five brands implement at-home test drives simultaneously. Results are presented in Tables 4.5, 4.6, and 4.7.

There are consistent patterns in the results across the counterfactuals. First, the average number of searches increases as more brands implement at-home test drives. Under the baseline simulation, consumers make 1.129 searches on average. This increases to 1.131–1.133 when one brand unilaterally implements; to 1.134–1.135 when there is bilateral implementation; and finally to 1.140 when all brands simultaneously implement. The implementing brand(s) increase their probability of being searched, but realize a decrease in the probability of purchase given search. However, increased search more than offsets the decreased conditional purchase probability, yielding an increase in market share in all but the last counterfactual. In the last counterfactual, all brands simultaneously implement at-home test drives and since not all brands can simultaneously increase market share, I find instead that foreign brands (Honda, Nissan, and Toyota) are able to gain market share at the expense of domestic brands (Chevrolet and Ford).

In addition, all counterfactual results indicate that model (ii) underestimates the effects of at-home test drives relative to market share predictions from model (iii). The disparity in predictions highlights the importance of including the MVSD as a parameter to be estimated in the sequential search model, as the failure to do so leads to incorrect estimates of search cost parameters, which translate into incorrect results of counterfactual analyses that affect consumers' search costs. In the counterfactuals presented here, manipulation of the distance a consumer must incur to inspect a vehicle changes her search cost, which impact how she will search and ultimately the what she will purchase. Thus, accurately assessing her sensitivity to the distance component of search cost is vital to predicting her behavior.

Table 4.4: Unilateral Adoption of At-Home Test Drives

Brand Adopting At-Home Test Drives	No At-Home Test Drives	With At-Home Test Drives	Change in Percent
Model (iii)			
Market Share			
Honda	0.145	0.194	34.2%
Chevrolet	0.200	0.247	23.4%
Ford	0.286	0.338	18.0%
Nissan	0.093	0.132	41.3%
Toyota	0.276	0.353	28.0%
Probability of Search			
Honda	0.163	0.221	35.4%
Chevrolet	0.225	0.281	24.4%
Ford	0.323	0.384	18.8%
Nissan	0.106	0.151	42.7%
Toyota	0.311	0.402	29.0%
Probability of Purchase Conditional on Search			
Honda	0.904	0.896	-0.9%
Chevrolet	0.903	0.898	-0.6%
Ford	0.911	0.907	-0.5%
Nissan	0.897	0.886	-1.2%
Toyota	0.909	0.903	-0.7%
Model (ii)			
Market Share			
Honda	0.148	0.164	11.3%
Chevrolet	0.198	0.211	6.4%
Ford	0.283	0.298	5.1%
Nissan	0.096	0.109	13.8%
Toyota	0.275	0.302	9.8%
Probability of Search			
Honda	0.149	0.167	11.9%
Chevrolet	0.201	0.214	6.8%
Ford	0.287	0.303	5.5%
Nissan	0.097	0.111	14.4%
Toyota	0.279	0.307	10.3%
Probability of Purchase Conditional on Search			
Honda	0.990	0.985	-0.5%
Chevrolet	0.988	0.985	-0.3%
Ford	0.989	0.987	-0.3%
Nissan	0.988	0.982	-0.6%
Toyota	0.990	0.986	-0.4%

Table 4.5: Chevrolet and Ford Adopt At-Home Test Drives

Brand	No At-Home Test Drives	With At-Home Test Drives	Change in Percent
Model (iii)			
Market Share			
Honda	0.145	0.131	-9.1%
Chevrolet	0.200	0.224	11.7%
Ford	0.286	0.313	9.3%
Nissan	0.093	0.084	-9.8%
Toyota	0.276	0.248	-10.1%
Probability of Search			
Honda	0.163	0.149	-9.1%
Chevrolet	0.225	0.254	12.6%
Ford	0.323	0.356	10.2%
Nissan	0.106	0.095	-9.7%
Toyota	0.311	0.280	-10.0%
Probability of Purchase Conditional on Search			
Honda	0.904	0.903	-0.2%
Chevrolet	0.903	0.896	-0.8%
Ford	0.911	0.904	-0.8%
Nissan	0.897	0.895	-0.2%
Toyota	0.909	0.907	-0.2%
Model (ii)			
Market Share			
Honda	0.148	0.144	-2.3%
Chevrolet	0.198	0.204	3.0%
Ford	0.283	0.291	2.6%
Nissan	0.096	0.093	-2.7%
Toyota	0.275	0.268	-2.7%
Probability of Search			
Honda	0.149	0.146	-2.3%
Chevrolet	0.201	0.208	3.4%
Ford	0.287	0.296	3.0%
Nissan	0.097	0.095	-2.7%
Toyota	0.279	0.271	-2.7%
Probability of Purchase Conditional on Search			
Honda	0.990	0.990	0.0%
Chevrolet	0.988	0.985	-0.3%
Ford	0.990	0.987	-0.3%
Nissan	0.988	0.987	0.0%
Toyota	0.990	0.990	0.0%

Table 4.6: Honda and Toyota Adopt At-Home Test Drives

Brand	No At-Home Test Drives	With At-Home Test Drives	Change in Percent
Model (iii)			
Market Share			
Honda	0.145	0.172	19.2%
Chevrolet	0.200	0.172	-14.0%
Ford	0.286	0.247	-13.7%
Nissan	0.093	0.081	-13.8%
Toyota	0.276	0.328	19.0%
Probability of Search			
Honda	0.163	0.197	20.4%
Chevrolet	0.225	0.194	-14.1%
Ford	0.323	0.279	-13.6%
Nissan	0.106	0.091	-13.7%
Toyota	0.311	0.374	20.0%
Probability of Purchase Conditional on Search			
Honda	0.904	0.894	-1.1%
Chevrolet	0.903	0.901	-0.3%
Ford	0.911	0.908	-0.3%
Nissan	0.897	0.895	-0.3%
Toyota	0.909	0.901	-0.9%
Model (ii)			
Market Share			
Honda	0.148	0.157	6.6%
Chevrolet	0.198	0.188	-5.2%
Ford	0.283	0.269	-5.0%
Nissan	0.096	0.092	-4.5%
Toyota	0.275	0.294	6.9%
Probability of Search			
Honda	0.149	0.160	7.2%
Chevrolet	0.201	0.190	-5.2%
Ford	0.287	0.273	-5.0%
Nissan	0.097	0.093	-4.5%
Toyota	0.279	0.299	7.4%
Probability of Purchase Conditional on Search			
Honda	0.990	0.985	-0.5%
Chevrolet	0.998	0.988	0.0%
Ford	0.990	0.989	0.0%
Nissan	0.988	0.988	0.0%
Toyota	0.990	0.986	-0.4%

Table 4.7: All Brands Simultaneously Adopt At-Home Test Drives

Brand	No At-Home Test Drives	With At-Home Test Drives	Change in Percent
Model (iii)			
Market Share			
Honda	0.145	0.156	8.1%
Chevrolet	0.200	0.188	-5.9%
Ford	0.286	0.263	-8.1%
Nissan	0.093	0.104	11.5%
Toyota	0.276	0.288	4.6%
Probability of Search			
Honda	0.163	0.179	9.3%
Chevrolet	0.225	0.214	-5.1%
Ford	0.323	0.299	-7.3%
Nissan	0.106	0.119	12.8%
Toyota	0.311	0.329	5.6%
Probability of Purchase Conditional on Search			
Honda	0.904	0.892	-1.4%
Chevrolet	0.903	0.893	-1.1%
Ford	0.911	0.900	-1.2%
Nissan	0.897	0.883	-1.4%
Toyota	0.909	0.897	-1.1%
Model (ii)			
Market Share			
Honda	0.148	0.153	3.4%
Chevrolet	0.198	0.191	-3.3%
Ford	0.283	0.272	-3.9%
Nissan	0.096	0.101	5.3%
Toyota	0.275	0.283	2.7%
Probability of Search			
Honda	0.149	0.155	3.9%
Chevrolet	0.201	0.195	-2.9%
Ford	0.287	0.277	-3.5%
Nissan	0.097	0.103	5.9%
Toyota	0.279	0.288	3.2%
Probability of Purchase Conditional on Search			
Honda	0.990	0.985	0.5%
Chevrolet	0.988	0.984	0.4%
Ford	0.990	0.986	0.3%
Nissan	0.988	0.981	0.6%
Toyota	0.990	0.985	0.5%

CHAPTER 5

Conclusions and Future Research

Prior to the Internet, store visits and the physical inspection of products played a prominent role during a consumer's purchase process – especially for high-ticket durable goods such as automobiles. Information on prices and vehicle features was difficult to obtain except via a dealership visit. During the last 15 years, a wealth of online information sources have provided a low cost alternative to time-consuming dealership visits. Therefore it is an important empirical question whether dealerships continue to provide (substantial) value to consumers.

To answer this question, it is necessary to separately quantify both the cost *and* the benefit of search. I show that – with an exogenous search cost shifter – both the cost and the benefit of search can be identified. I estimate a sequential search model for product fit. My empirical results show that the benefit provided by dealerships to consumers remains substantial. This finding points to a continuously important role of physical stores in the age of the Internet. The actions of online retailers in other product categories also corroborate this conclusion. Initially online-only retailers, such as Warby Parker and Bonobos, have recently opened brick and mortar stores to provide consumers with information that is difficult to communicate electronically.

My research is not without limitations and offers opportunities for future work. First, for vehicles not purchased, I observe that a consumer visited a dealership, but not which specific car the consumer was interested in. This is a limitation of my data. I therefore make the assumption that the consumer was searching for similar cars across different dealerships.

The data also do not indicate the set of available vehicles at each dealership, and therefore I do not model vehicle choice, but only brand-dealer choice (conditional on the type of vehicle the consumer is interested in). And lastly, the benefit provided by dealers might be heterogenous, varying with brand, dealership size, or other characteristics. I leave it for future research to explore this type of heterogeneity.

APPENDIX A

Estimation Code

Neither the search nor purchase probabilities can be expressed in closed form. I approximate the integrals in the likelihood function with averages using logit-smoothed accept-reject simulation. This simulated maximum likelihood estimation algorithm follows Train (2009) and is outlined in the following steps:

1. Do steps 2–4 for each consumer $i = 1, \ldots, N$
2. Draw Q values from the density $f_\varepsilon(\varepsilon)$ for each of the K_i searches for a total of $Q \cdot K_i$ draws and draw Q values from the density $f_\eta(\eta)$ for a total of $Q \cdot J_i$ draws
3. For each set of random draws:
   (a) Compute $\nu_{1,j} = z_{ij} - \max_{h \leq j} \{u_{ih}\}$ for $j = 2, \ldots, K_i$
   (b) Compute $\nu_{2,j} = z_{ij} - \max_{k > j} \{z_{ik}\}$ for $j = 1, \ldots, K_i$
   (c) Compute $\nu_3 = \max_{h \leq K_i} \{u_{ih}\} - \max_{k > j} \{z_{ik}\}$
   (d) Compute $\nu_4 = u_{ij_i^*} - \max_{h \leq K_i} \{u_{ih}\}$ for the chosen j_i^*
   (e) Compute the simulated individual likelihood given one set of draws:

$$\tilde{L}_i^q = \left(1 + \sum_{j=2}^{K_i} \mathrm{e}^{-\lambda_1 \nu_{1,j}} + \sum_{j=1}^{K_i} \mathrm{e}^{-\lambda_2 \nu_{2,j}} + \mathrm{e}^{-\lambda_3 \nu_3} + \mathrm{e}^{-\lambda_4 \nu_4}\right)^{-1}$$

4. Average the draw-specific simulated individual likelihoods over the draws:

$$\tilde{L}_i = \frac{1}{Q} \sum_{q=1}^{Q} \tilde{L}_i^q$$

5. Take logs and aggregate the individual-specific likelihoods over individuals to form the total simulated log-likelihood function:

$$\log\left(\tilde{L}\right) = \sum_{i=1}^{N} \log\left(\tilde{L}_i\right)$$

Computing time when fitting the sequential search model is directly proportional to the number of draws used to approximate the integrals in the likelihood function. However, a large number of draws is necessary to achieve a good approximation. When calculating each individual likelihood, I therefore use $1,000$ draws for η_{ij} and $1,000$ draws for ε_{ij}. By comparison, Kim, Albuquerque, and Bronnenberg (2017) use 40 draws when testing the kernel-smoothed AR estimation approach; Honka (2014) and Ursu (2018) use 50 draws to estimate their search models with the same approach.

R code to implement the estimation follows:

```
# description of data required
        # N         - scalar number of consumers
        # P         - scalar number of utility (X) variables
        # R         - scalar number of cost (d) variables
        # params    - vector of P beta parameters, R gamma parameters, and the MVSD (sigma_eps)
        # datlist   - a list of N sublists; each sublist has elements:
                       # y        - length-J_i vector of 0/1 to indicate choice
                       # x        - J_i x P matrix of utility variables
                       # searched - length-J_i vector of 0/1 to indicate not/searched
                       # distmat  - J_i x R matrix of cost variables
        # ndraws    - number of draws to approximate integrals
        # sigma_eta - eta distributed N(0, sigma_eta^2)
        # lambda    - length-1 or length-4 vector of logit-smoothing parameters
        # seed      - permits reproducibility
```

```
# packages required for parallel processing
        library(doParallel)
        library(foreach)

# function to combine individual likelihoods
LL_ksf <- function(params, datlist, ndraws, sigma_eta, lambda, seed) {

        N <- length(datlist)

        # calc likelihood by looping over consumers i=1,...,N
        ll_vec <- foreach(i=1:N, .combine=c) %dopar% {
                L_ksf_i(params, i, datlist[[i]], ndraws, sigma_eta, lambda, seed+i)
        }

        return(sum(ll_vec))
}

# function to calculate one consumer's likelihood
L_ksf_i <- function(params, i, dat, ndraws, sigma_eta, lambda, seed) {

        ## x'beta, cost, etc.
        set.seed(seed)

        J <- length(dat$searched)  # num searchable alts
        K <- sum(dat$searched)     # num searches made in data
        P <- ncol(dat$x)           # num x-variables
        R <- ncol(dat$distmat)     # num cost-variables

        beta      <- params[1:P]
        gamma     <- params[P+1:R]
        sigma_eps <- exp(params[P+R+1])

        xb   <- as.vector(dat$x %*% beta)
        cost <- as.vector(exp(dat$distmat %*% gamma))

        purchased <- which(dat$y == 1)

        eta <- matrix(rnorm(J*ndraws, mean=0, sd=sigma_eta), nrow=J, ncol=ndraws)
        eps <- matrix(rnorm(J*ndraws, mean=0, sd=sigma_eps), nrow=J, ncol=ndraws)

        Vij <- xb + eta
```

```
Uij <- Vij + eps

## zeta & reservation utility

zetafun <- function(zeta, cost, sigma_eps) {
        exp(dnorm(zeta, log=T)) -
                exp(pnorm(zeta, lower.tail=F, log.p=T))*zeta -
                cost/sigma_eps
}

zeta <- rep(NA_real_, length(cost))
zlb <- -1e10; zub <- 6.070461

for(cc in 1:length(cost)) {
        zeta[cc] <- uniroot(zetafun, cost=cost[cc], sigma_eps=sigma_eps,
                        lower=zlb, upper=zub, extendInt="yes",
                        tol=.Machine$double.eps^0.5,
                        maxiter=1e6, check.conv=T)$root
}

# reservation utility (dim J x ndraws)
Zij <- Vij + zeta*sigma_eps

## KSF components

        # nu1: decision to cont. searching -- z searched > max(u realized)
        # nu2: selection of what to search -- z_{h-1} > z_{h} for h=1..K
        # nu3: decision to stop searching  -- max(u realized) > max(z notsearched)
        # nu4: choice of what to purchase  -- choose max(u realized)

# if only 1 search
if(K == 1) {
        max_z_notsearched <- Rfast::colMaxs(Zij[2:J,], value=T)
        nu2 <- Zij[1,] - max_z_notsearched
        nu3 <- Uij[1,] - max_z_notsearched

        res <- 1/(1 + exp(-lambda[2]*nu2) + exp(-lambda[3]*nu3))
}

# if search everything
if(K > 1 && K == J) {
        nu1 <- Zij[-1,,drop=F] - apply(Uij, 2, cummax)[-J,,drop=F]
```

```
        nu2 <- Zij[-J,,drop=F] - Zij[-1,,drop=F]
        nu4 <- Uij[purchased,,drop=F] - Rfast::colMaxs(Uij[-purchased,,drop=F], value=T)

        res <- 1/(1 + colSums(exp(-lambda[1]*nu1)) +
                colSums(exp(-lambda[2]*nu2)) +
                exp(-lambda[4]*nu4) )
    }

    # if multiple searches but not complete search
    if(K > 1 && K < J) {
        nu1  <- Zij[2:K,,drop=F] - apply(Uij, 2, cummax)[1:(K-1),,drop=F]
        nu2a <- Zij[1:(K-1),,drop=F] - Zij[2:K,,drop=F]
        max_z_notsearched <- apply(Zij[J:1,,drop=F], 2, cummax)[J:1,,drop=F]
        nu2b <- Zij[K,] - max_z_notsearched[K+1,]
        nu3  <- Rfast::colMaxs(Uij[1:K,,drop=F], value=TRUE) - max_z_notsearched[K+1,]
        idx  <- c(1:K)[-purchased]
        nu4  <- Uij[purchased,] - Rfast::colMaxs(Uij[idx,,drop=F], value=T)

        res <- 1/(1 + colSums(exp(-lambda[1]*nu1 )) +
                colSums(exp(-lambda[2]*nu2a)) +
                exp(-lambda[2]*nu2b) +
                exp(-lambda[3]*nu3 ) +
                exp(-lambda[4]*nu4 ) )
    }

    ## individual log-likelihood value
    return(log(mean(res)))
}
```

APPENDIX B

Simulation Code

R code to simulate optimal consumer sequential search behavior according to Weitzman (1979) is provided below.

```
# description of elements
    # N       - Number of consumers
    # J_df    - scalar number of alternatives, drawn from chisq(df=J_df)+3*exp(1)
    # P       - scalar number of non-brand X variables
    # alpha   - vector of alternative-specific intercepts (the first should be zero)
    # beta    - vector of P coefficients for non-brand X variables
    # gamma   - vector of 2 search-cost parameters
    # x       - X var distributed unif[x_min, x_max] where x_m** are length-P vectors
    # dist_df - distances are drawn from chisq(df=dist_df)
    # eta     - component of utility known to consumer, assumed N(0, sigma_eta^2)
    # eps     - component of utility to search for, assumed N(0, sigma_eps^2)
    # seed    - permits reproducibility


# function to simulate many consumers
    sim_dat_fun <- function(N, J_df, P, alpha, beta, gamma, x_min,
                            x_max, dist_df, sigma_eta, sigma_eps, seed) {

        set.seed(seed)
        simdat <- vector(mode="list", length=N)

        for(i in 1:N) {
            simdat[[i]] <- sim_one_fun(J_df, P, alpha, beta, gamma,
                                       x_min, x_max, dist_df,
                                       sigma_eta, sigma_eps, seed+i)
        }
        return(simdat)
    }
```

```
# function to simulate one consumer

    sim_one_fun <- function(J_df, P, alpha, beta, gamma, x_min,
                            x_max, dist_df, sigma_eta, sigma_eps, seed) {

        set.seed(seed)

        # draw number of alternatives
        J <- ceiling( rchisq(1, df=J_df) + rexp(1)*3 )

        # draw brands for each vehicle
        brandvec <- sample(1:length(alpha), size=J, replace=T)
        x1mat <- matrix(0, nrow=J, ncol=length(alpha))
        x1mat[cbind(1:J, brandvec)] <- 1

        # draw X vars for each of J vehicles;
        x2mat <- t(sapply(rep(P,J), function(x) runif(x, x_min, x_max)))

        # calc X'beta
        xmat <- cbind(x1mat, x2mat)
        if(J==1) xmat <- matrix(xmat, nrow=1)
        colnames(xmat) <- c(paste0("brand", 1:length(alpha)), paste0("xvar", 1:P))
        xb <- as.vector(xmat %*% c(alpha, beta))

        # draw eta_j value for each of J vehicles
        eta <- rnorm(J, mean=0, sd=sigma_eta)

        # all known utility values to consumer i
        delta <- xb + eta

        # draw distances
        dist <- abs(rchisq(J, df=dist_df) + rnorm(J, mean=0, sd=dist_df*2))
        cost <- as.vector( exp( cbind(1, dist) %*% gamma ) )

        # calc zeta
        zetafun <- function(zeta, cost, sigma_eps) {
            exp(dnorm(zeta, log=T)) -
            exp(pnorm(zeta, lower.tail=F, log.p=T))*zeta -
            cost/sigma_eps
        }
```

```
zeta <- rep(NA, length(cost))
zlb <- -1e10; zub <- 6.070461

for(cc in 1:length(cost)) {
    zeta[cc] <- uniroot(zetafun, cost=cost[cc], sigma_eps=sigma_eps,
                        lower=zlb, upper=zub, extendInt="yes",
                        tol=.Machine$double.eps^0.5,
                        maxiter=1e6, check.conv=T)$root
}

# reservation utilities
Zij <- delta + zeta*sigma_eps

# get search order (AISO = alts in search order)
ord <- rank(-Zij)   #  ord=c(3,1,2) means 2nd searched first : B >> C >> A
aiso <- order(ord)  # aiso=c(3,1,2) means 3rd searched first : C >> A >> B

# prep for searches
eps <- rep(NA_real_, J)
Uij <- rep(NA_real_, J)
searched <- rep(F,J)
searchnum <- 1

# first search is free
altj <- aiso[searchnum]
eps[altj] <- rnorm(1, mean=0, sd=sigma_eps)
Uij[altj] <- delta[altj] + eps[altj]
searched[altj] <- T

# act optimally according to weitzmann
if(J>1) {
    while(max(Uij[searched]) <= max(Zij[!searched])) {
        searchnum <- searchnum + 1
        altj <- aiso[searchnum]
        eps[altj] <- rnorm(1, mean=0, sd=sigma_eps)
        Uij[altj] <- delta[altj] + eps[altj]
        searched[altj] <- T
        if(searchnum == J) break
    }
}

return( list(x         = xmat[aiso,,drop=F],
```

```
                    purchased = which(aiso == which.max(Uij)),
                    searched  = searched[aiso],
                    distmat   = cbind(1, dist)[aiso,])  )
}
```

Bibliography

Albuquerque, Paulo and Bart J Bronnenberg (2012), "Measuring the Impact of Negative Demand Shocks on Car Dealer Networks," *Marketing Science*, 31 (1), 4–23.

Anderson, Simon P, Regis Renault et al. (2018), "Firm pricing with consumer search," volume 2, chapter 8, 177–224.

Armstrong, Mark (2017), "Ordered Consumer Search," *Journal of the European Economic Association*, 15 (5), 989–1024.

Baye, Michael R, John Morgan, Patrick Scholten et al. (2006), "Information, Search, and Price Dispersion," *Handbook on Economics and Information Systems*, 1, 323–375.

Berry, Steven, James Levinsohn, and Ariel Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, 63 (4), 841–890.

Bronnenberg, Bart J, Jun B Kim, and Carl F Mela (2016), "Zooming in on Choice: How do Consumers Search for Cameras Online?" *Marketing Science*, 35 (5), 693–712.

Chade, Hector and Lones Smith (2006), "Simultaneous Search," *Econometrica*, 74 (5), 1293–1307.

Chen, Yuxin and Song Yao (2017), "Sequential Search with Refinement: Model and Application with Click-Stream Data," *Management Science*, 63 (12), 4345–4365.

Chick, Stephen E and Peter Frazier (2012), "Sequential Sampling with Economics of Selection Procedures," *Management Science*, 58 (3), 550–569.

Choi, Michael, Anovia Yifan Dai, and Kyungmin Kim (2018), "Consumer Search and Price Competition," *Econometrica*, 86 (4), 1257–1281.

Chung, Jae, Pradeep Chintagunta, and Sanjog Misra (2019), "Estimation of Sequential Search Models," *Working paper*.

De los Santos, Babur, Ali Hortacsu, and Matthijs R Wildenbeest (2012), "Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior," *American Economic Review*, 102 (6), 2955–80.

Diamond, Peter A (1971), "A Model of Price Adjustment," *Journal of Economic Theory*, 3 (2), 156–168.

Dong, Xiaojing, Ilya Morozov, Stephan Seiler, and Liwen Hou (2020), "Estimation of Preference Heterogeneity in Markets with Costly Search," *Working paper*.

Geweke, John, Michael Keane, and David Runkle (1994), "Alternative Computational Approaches to Inference in the Multinomial Probit Model," *The review of economics and statistics*, 609–632.

Hauser, John R and Birger Wernerfelt (1990), "An Evaluation Cost Model of Consideration Sets," *Journal of Consumer Research*, 16 (4), 393–408.

Honka, Elisabeth (2014), "Quantifying Search and Switching Costs in the U.S. Auto Insurance Industry," *The RAND Journal of Economics*, 45 (4), 847–884.

Honka, Elisabeth and Pradeep Chintagunta (2017), "Simultaneous or Sequential? Search Strategies in the U.S. Auto Insurance Industry," *Marketing Science*, 36 (1), 21–42.

Honka, Elisabeth, Ali Hortacsu, and Maria Ana Vitorino (2017), "Advertising, Consumer Awareness, and Choice: Evidence from the US Banking Industry," *The RAND Journal of Economics*, 48 (3), 611–646.

Honka, Elisabeth, Ali Hortacsu, and Matthijs Wildenbeest (2019), "Empirical Search and Consideration sets," in "Handbook of the Economics of Marketing," Elsevier, volume 1, 193–257.

Jang, Sungha, Ashutosh Prasad, and Brian T Ratchford (2017), "Consumer Search of Multiple Information Sources and Its Impact on Consumer Price Satisfaction," *Journal of Interactive Marketing*, (40), 24–40.

Jiang, Zhenling, Tat Chan, Che Hai, and Youwei Wang (2019), "Consumer Search and Purchase: An Empirical Investigation of Retargeting Based on Consumer Online Behaviors," *Working paper*.

Kiel, Geoffrey C and Roger A Layton (1981), "Dimensions of Consumer Information Seeking Behavior," *Journal of Marketing Research*, 18 (2), 233–239.

Kim, Jun B, Paulo Albuquerque, and Bart J Bronnenberg (2010), "Online Demand Under Limited Consumer Search," *Marketing Science*, 29 (6), 1001–1023.

— (2017), "The Probit Choice Model under Sequential Search with an Application to Online Retailing," *Management Science*, 63 (11), 3911–3929.

Kim, Jung Seek and Brian T Ratchford (2012), "Consideration Set of Automobiles: Purchase Feedback and Exclusivity in Formation," *Journal of Management and Marketing Research*, 9, 1.

Klein, Lisa R and Gary T Ford (2003), "Consumer Search for Information in the Digital Age: An Empirical Study of Prepurchase Search for Automobiles," *Journal of Interactive Marketing*, 17 (3), 29–49.

Lafontaine, Francine and Fiona Scott Morton (2010), "Markets: State Franchise Laws, Dealer Terminations, and the Auto Crisis," *Journal of Economic Perspectives*, 24 (3), 233–50.

Lapersonne, Eric, Gilles Laurent, and Jean-Jacques Le Goff (1995), "Consideration Sets of Size One: An Empirical Investigation of Automobile Purchases," *International Journal of Research in Marketing*, 12 (1), 55–66.

McCall, John Joseph (1970), "Economics of Information and Job Search," *The Quarterly Journal of Economics*, 113–126.

Mehta, Nitin, Surendra Rajiv, and Kannan Srinivasan (2003), "Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation," *Marketing Science*, 22 (1), 58–84.

Moraga-Gonzalez, Jose, Zsolt Sandor, and Matthijs Wildenbeest (2018), "Consumer Search and Prices in the Automobile Market," *Working paper*.

Moraga-Gonzalez, Jose Luis, Zsolt Sandor, and Matthijs R Wildenbeest (2017), "Nonsequential Search Equilibrium with Search Cost Heterogeneity," *International Journal of Industrial Organization*, 50, 392–414.

Moreno, Antonio and Christian Terwiesch (2017), "The Effects of Product Line Breadth: Evidence from the Automotive Industry," *Marketing Science*, 36 (2), 254–271.

Morgan, Peter and Richard Manning (1985), "Optimal Search," *Econometrica*, 923–944.

Morozov, Ilya (2019), "Measuring Benefits from New Products in Markets with Information Frictions," *Working paper*.

Murry, Charles and Henry S Schneider (2016), "The Economics of Retail Markets for New and Used Cars," in "Handbook on the Economics of Retailing and Distribution," Edward Elgar Publishing.

Murry, Charles and Yiyi Zhou (2019), "Consumer Search and Automobile Dealer Co-Location," *Management Science*, forthcoming.

Nurski, Laura and Frank Verboven (2016), "Exclusive Dealing as a Barrier to Entry? Evidence from Automobiles," *The Review of Economic Studies*, 83 (3), 1156–1188.

Palazzolo, Mike and Fred Feinberg (2015), "Modeling Consideration Set Substitution," *Working paper*.

Ratchford, Brian T, Myung-Soo Lee, and Debabrata Talukdar (2003), "The Impact of the Internet on Information Search for Automobiles," *Journal of Marketing Research*, 40 (2), 193–209.

Ratchford, Brian T and Narasimhan Srinivasan (1993), "An Empirical Investigation of Returns to Search," *Marketing Science*, 12 (1), 73–87.

Ratchford, Brian T, Debabrata Talukdar, and Myung-Soo Lee (2007), "The Impact of the Internet on Consumers' Use of Information Sources for Automobiles: A Re-Inquiry," *Journal of Consumer Research*, 34 (1), 111–119.

Ratchford, Brian T et al. (2008), "Consumer Search Behavior and Its Effect on Markets," *Foundations and Trends in Marketing*, 3 (1), 1–74.

Roberts, John H and James M Lattin (1991), "Development and Testing of a Model of Consideration Set Composition," *Journal of Marketing Research*, 429–440.

Seiler, Stephan and Fabio Pinna (2017), "Estimating Search Benefits from Path-Tracking Data: Measurement and Determinants," *Marketing Science*, 36 (4), 565–589.

Singh, Sonika, Brian T Ratchford, and Ashutosh Prasad (2014), "Offline and Online Search in Used Durables Markets," *Journal of Retailing*, 90 (3), 301–320.

Stigler, George J (1961), "The Economics of Information," *Journal of Political Economy*, 69 (3), 213–225.

Train, Kenneth E (2009), *Discrete Choice Methods with Simulation*, Cambridge University Press, 2nd edition.

Ursu, Raluca (2018), "The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions," *Marketing Science*, 37 (4), 530–552.

Ursu, Raluca, Qingliang Wang, and Pradeep Chintagunta (2018), "Search Duration," *Working paper*.

Vishwanath, Tara (1992), "Parallel Search for the Best Alternative," *Economic Theory*, 2 (4), 495–507.

Weitzman, Martin L (1979), "Optimal Search for the Best Alternative," *Econometrica*, 47 (3), 641–654.

Wolinsky, Asher (1986), "True Monopolistic Competition as a Result of Imperfect Information," *The Quarterly Journal of Economics*, 101 (3), 493–511.

Zettelmeyer, Florian, Fiona Scott Morton, and Jorge Silva-Risso (2006), "How the Internet Lowers Prices: Evidence from Matched Survey and Automobile Transaction Data," *Journal of Marketing Research*, 43 (2), 168–181.