Machine Learning for Demand Estimation in Long Tail Markets

Random coefficient multinomial logit models are widely used to estimate customer preferences from sales data. However, these estimation models can only allow for products with positive sales; this selection leads to highly biased estimates in long tail markets–i.e., markets where many products have zero or low sales. Such markets are increasingly common in areas such as online retail and other online marketplaces. In this paper, we propose a two-stage estimator that uses machine learning to correct for this bias. Our method first uses deep learning to predict the market shares of all products, where the neural network's structure mirrors the random coefficient multinomial logit model's data generating process. In the second stage, we use the predictions of the first stage to re-weight the observed shares in a way that corrects for the induced bias and maintains the causal interpretation of the structural model. We show that the estimated parameters are consistent in the number of markets. Our method performs well on simulated and real long tail data, producing accurate estimates of customer behavior. These improved estimates can subsequently be used to provide prescriptive policy recommendations on important managerial decisions such as pricing, assortment, etc.

1. Introduction

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

Given the importance of these decisions to managers and policy makers, it is key to empirically recover the parameter estimates without bias. To achieve this goal, many estimation techniques have been developed in the literature (Berry et al. 1995, Allenby and Rossi 1998, Jagabathula et al. 2020). The most commonly adopted approach is the estimation of random coefficient multinomial

logit (MNL) model introduced by Berry et al. (1995) (BLP). This method has two main advantages. First, it allows for endogenous product or service features–i.e., features that are correlated with unobserved product or service features, such as price. Second, it allows for heterogeneity in customer tastes captured by random coefficients in the utility function. The BLP approach has been widely applied to study customer choice behavior in many different industries across operations management, economics, and marketing (Nevo 2001, Sudhir 2001, Guajardo et al. 2015). This is the method we focus on in the current paper.

One disadvantage of BLP is that the method exhibits major limitations when the customer choice data has long tails. As discussed in detail in Section 2, BLP does not allow products or services with zero sales to be included in the estimation. Moreover, the method produces noisy estimates when products and services have very low sales or market shares. However, in practice, as shown in detail in Section 2, customer choice data often has very long tails–i.e., for a given period of observation, many products and services have zero or very low sales. The phenomenon has become increasingly common with big data, especially in settings with large customer choices such as online platforms and marketplaces. The common practice of dropping products and services with zero and very low sales introduces substantial bias in BLP estimates. Moreover, the bias in the estimates leads to biased policy recommendations and managerial decisions.

In this paper, we propose a two-stage estimator that corrects for the bias caused by long tail data in BLP estimates. In the first stage, we use a machine learning algorithm to predict the market shares of all products or services, including those with zero or low sales. In the second stage, we utilize the predicted shares to correct for the bias in the estimation. Intuitively, our method for correcting for the bias in BLP estimates is analogous to the bias correction method often used in the classic econometrics literature of estimating treatment effects. We use the predicted shares in the first stage to construct weights to correct for the bias in the second stage of the estimation. This is similar to using the estimated propensity score to construct weights to correct for bias in estimating treatment effects, applied to a nonlinear setting. Moreover, in combining the first stage prediction with the second stage estimation, we use recent tools at the intersection of machine learning and causal inference to de-bias our estimator, so that it maintains the causal interpretations of the parameters. We show that our proposed estimator is asymptotically consistent and easy to implement. To demonstrate the performance of our method, we conduct extensive simulation studies and test our method in real data. Both simulation results and estimation results using real data show that our method corrects for the bias in a wide variety of empirical settings and performs far better than BLP and alternative estimators.

Our paper contributes to the empirical literature in operations management, marketing, and economics, in which random coefficient multinomial logit models have been widely applied (Cachon

and Olivares 2010, Allon et al. 2011, Guajardo et al. 2015, Quan and Williams 2018). In particular, we provide an easy-to-implement estimation technique that is able to deal with the bias commonly observed in demand estimation of large platforms and marketplaces (Kabra et al. 2019, He et al. 2018). One recent paper in the econometrics literature studies the same problem and proposes an inequality-based estimator to correct for the bias (Gandhi et al. 2020). The limitation of this approach is that it can only allow for very few covariates with continuous values as it requires clustering products with similar values in the covariate space. We provide detailed comparisons of the performance of the two approaches in our simulations, which show that our approach performs substantially better. A concurrent paper provides an alternative solution to this problem (Dubé et al. 2021). However, the estimator proposed by Dubé et al. (2021) relies on additional data on the supply side and detailed market equilibrium conditions which are often unavailable in many empirical settings. By contrast, our estimator only relies on demand-side information. There is also a recent literature studying large scale demand estimation problems (Bajari et al. 2015, Amano et al. 2019, Jiang et al. 2020). While the objective of Bajari et al. (2015) is to improve out-of-sample prediction, our method focuses on recovering the causal preference parameters without bias. Amano et al. (2019), Jiang et al. (2020) leverages customer search data to analyze product substitution patterns. By contrast, our method, like BLP, uses sales or market shares data only. Our paper also complements the broader literature in operations management where the random coefficient multinomial logit models have been used to develop extensive tools to study pricing and assortment decisions in revenue management (Aksoy-Pierson et al. 2013, Rusmevichientong et al. 2014, Besbes and Sauré 2016). The asymptotic properties of our proposed estimator are derived based on Freyberger (2015) and Chernozhukov et al. (2018).

The rest of the paper is organized as follows. Section 2 describes the BLP estimator and explains the source of the bias when applied to long tail data. We describe the detailed construction of the two-stage estimator and its asymptotic properties in Section 3. We discuss the simulation results in Section 4 and test our estimator using real data in Section 5.

2. BLP and Selection Bias

In this section, we first describe the BLP estimator and explain the source of the selection bias when the data has a long tail. We then describe the intuition of the solution we propose to correct for the bias.

2.1. Random Coefficient MNL Model

Let $x_{jt} \in \mathbb{R}^{d_x}$ be a set of observed characteristics of product j in market t, including its price $p_{jt} \in \mathbb{R}$ and other features. Let $\xi_{jt} \in \mathbb{R}$ denote the product's unobserved feature which can be correlated with x_{jt}; this characteristic affects consumer choice as it is known to the consumer and the seller, but it

is not observed by the researcher. ξ_{jt} are independently distributed across markets. Now consider consumer i in market t, with preferences $\beta_i \in \mathbb{R}^{d_x}$. These consumer preferences are heterogeneous: β_i is a random draw from a distribution with mean $\bar{\beta} \in \mathbb{R}^{d_x}$ and variance $\Sigma_0 \in \mathbb{R}^{d_x \times d_x}$. Note that we can write $\beta_i = \bar{\beta} + \Sigma_0 \nu_i$, where $\nu_i \in \mathbb{R}^{d_x}$ are i.i.d. Consumer i's utility, u_{ijt}, of purchasing product j in market t is

$$u_{ijt} = x'_{jt}\bar{\beta} + x'_{jt}\Sigma_0\nu_i + \xi_{jt} + \epsilon_{ijt}, \tag{1}$$

where ϵ_{ijt} captures consumer- and product-specific idiosyncratic taste and is assumed to be i.i.d. following type I extreme value distribution. The utility of the 'no purchase' option u_{i0} is normalized to ϵ_{i0}, where ϵ_{i0} is also type I extreme value distributed. Consumer i selects the product $j \in \{0, 1, ..., J_t\}$ in market t that maximizes their utility u_{ijt}; these choices yield product j's (true) market share

$$\pi_{jt} = \int \frac{e^{x'_{jt}\bar{\beta} + x'_{jt}\Sigma_0\nu + \xi_{jt}}}{1 + \sum_{k=1}^{J_t} e^{x'_{kt}\bar{\beta} + x'_{kt}\Sigma_0\nu + \xi_{kt}}} dP_0(\nu) \tag{2}$$

where market t has J_t products.

π_{jt} conveys the true market share of product j in market t, and by definition, must be strictly positive. However, this true market share is not observed in practice; instead, in each market, we observe the aggregate results of N consumers making purchasing decisions with choice probabilities $\pi_t = (\pi_{1t}, ..., \pi_{J_t t})$. Thus, in market t, we observe the aggregate share vector $s_t = (s_{1t}, ..., s_{J_t t})$, a realization of the random share vector $S_t \sim \frac{1}{N} Multinomial(N, \pi_t)$. Note that for sufficiently large N, $s_{jt} \approx \pi_{jt}$, $\forall j \in \{1, .., J_t\}$. However, there is one key difference: while π_{jt} is strictly positive, the sampled s_{jt} can be zero, if a product has low utility. This fact can create a long tail of products with zero observed sales, which is the setting we consider in this paper.

2.2. BLP Estimation

In the data generating process described in Section 2.1, $\theta_0 = (\bar{\beta}, vec(\Sigma_0))$ is the (causal) parameter of interest. For any $\theta = (\beta, vec(\Sigma))$, a product's market share can be written using the share function σ,

$$\sigma_{jt}(\theta, x_t, \xi_t) = \int \frac{e^{x'_{jt}\beta + x'_{jt}\Sigma\nu + \xi_{jt}}}{1 + \sum_{k=1}^{J_t} e^{x'_{kt}\beta + x'_{kt}\Sigma\nu + \xi_{kt}}} dP_0(\nu) \tag{3}$$

where $x_t = (x_{1t}....x_{J_t t})$ and $\xi_t = (\xi_{1t}....\xi_{J_t t})$.

For now, assume all products have positive observed shares, that is, $s_{jt} > 0 \; \forall j, t$. Berry et al. (1995) show that for any value of θ, there exists a set of $\tilde{\xi}_t = (\tilde{\xi}_{1t}....\tilde{\xi}_{J_t t})$, such that

$$\sigma_{jt}(\theta, x_t, \tilde{\xi}_t) = s_{jt}$$

Note that $\tilde{\xi}_t$ is not necessarily the true value of the unobservables. Of course, at the true value of the parameters, $\theta = \theta_0$, plugging in the true unobservables ξ_t will yield $\sigma_{jt}(\theta_0, x_t, \xi_t) \approx s_{jt}$ (in large samples), which is a key insight for the consistency result of the BLP estimator. Crucially, these share equations are invertible; we can thus define the inverse share function

$$\xi_{jt}(\theta, x_t, s_t) = \sigma_{jt}^{-1}(\theta, x_t, s_t) \tag{4}$$

It is important here to note the difference between ξ_{jt}, the true value of the unobserved product feature, and $\xi_{jt}(\theta, x_t, s_t)$, the value inferred from the inverse share function for a given value of θ. As per convention (e.g. Freyberger 2015), we overload ξ_{jt} to denote both these concepts; however, for the sake of clarity, we always write the latter in its full form as $\xi_{jt}(\theta, x_t, s_t)$ (i.e. as the inverse share function) and the true value of the unobservable simply as ξ_{jt}.

Since some of the product features are endogenous–i.e., ξ_{jt} is correlated with some x_{jt}, to estimate θ_0, we need a set of valid instrumental variables, $z_{jt} \in \mathbb{R}^{d_z}$. Note that for those exogenous product features, the instruments are just functions of themselves. Let $z_t \in \mathbb{R}^{J_t \times d_z}$ be the matrix of instruments for all products in this market. At the true value of $\theta = \theta_0$,

$$\mathbb{E}_T[\xi_t(\theta_0, x_t, s_t)|z_t] \to \mathbb{E}[\xi_t|z_t] = 0 \tag{5}$$

as $T \to \infty$, where $\mathbb{E}_T$ denotes the average in the observed sample of T markets. This moment condition is the key to estimating θ_0 given the observed covariates x_t, instruments z_t, and market shares s_t in the sample. BLP proposes an estimator that uses Generalized Method of Moments (GMM) to solve for θ by setting $\mathbb{E}_T[z_t'\xi_t(\theta, x_t, s_t)] = 0$,

$$\hat{\theta}_{BLP} \equiv \arg\min_{\theta} \Big(\frac{1}{T}\sum_{t=1}^{T} z_t'\xi_t(\theta, x_t, s_t)\Big)' W \Big(\frac{1}{T}\sum_{t=1}^{T} z_t'\xi_t(\theta, x_t, s_t)\Big) \tag{6}$$

where W is a positive semi-definite weighting matrix. Berry et al. (1995) and Dubé et al. (2012) discuss different approaches to solving Equation (6). Freyberger (2015) establishes that $\hat{\theta}_{BLP}$ is a $\sqrt{T}$ consistent estimator for θ_0 and derives its asymptotic distribution.

2.3. Long Tail Markets

The BLP estimator discussed above assumes that all products have positive sales in the sample–i.e., $s_{jt} > 0$ for all products $j \in \{1, ..., J_t\}$ in markets $t \in \{1, .., T\}$. In this paper, we deal with the estimation problem in long tail markets, that is, in markets where a large number of products have zero sales in the sample. As we discuss later, the BLP procedure cannot handle such markets and excludes the zero-sale products from estimation. This approach leaves out vital information and can lead to biased estimates of the parameters of interest.

Long tail markets are becoming increasingly common as more and more granular data becomes available on consumer decisions. Several examples can be found in online retail and other marketplaces. Here we present one such example from London's Bike Share System (He et al. 2018). This system consists of a network of bikes and docking stations set up around London to facilitate public transport. Customers can pick up a bike at any station with available bikes, ride it to their destination, then drop it off at a nearby station with empty docks. The dataset tracks these rides and measures the usage of any route, where a route is defined by a pair of starting and ending stations and usage is measured by the number of times customers used that route in 2017. Each route corresponds to a product, and its usage corresponds to the observed sales.

This dataset displays the long tail pattern, as a large percentage of routes have zero usage. Consider Table 1, which sorts the routes into buckets based on their distance and displays the percentage of unused routes in each bucket. As one may expect, as the route distance increases, the percentage of routes with zero usage in the sample increases. However, even around the median route length, about 40% of all routes were never used. This pattern poses a challenge in the estimation. For example, if we wanted to use BLP to determine which route features are most valued by customers, we would need to exclude all the unridden routes from our analysis. As we discuss next, this limitation can lead to highly biased estimates of customer preferences. These inaccurate results may in turn lead to poor policy decisions. For example, if the bike share system's managers are looking to expand the network, they may want to build new stations in places that will maximise usage. However, if they rely on inaccurate estimates customer demand at different locations, the network expansion is unlikely to focus on the service features that customers really value. The build will thus be far from optimal, wasting time, capital, and other resources.

Distance Percentile	Percentage of Unused Routes
0-10$^{\text{th}}$	4.7%
10-20$^{\text{th}}$	5.3%
20-30$^{\text{th}}$	9.7%
30-40$^{\text{th}}$	17.2%
40-50$^{\text{th}}$	27.0%
50-60$^{\text{th}}$	40.6%
60-70$^{\text{th}}$	55.2%
70-80$^{\text{th}}$	68.2%
80-90$^{\text{th}}$	81.9%
90-100$^{\text{th}}$	94.5%

Table 1 Zero usage routes in the London bike share system (He et al. 2018)

2.4. Selection

The BLP estimation procedure described in Section 2.2 fits the share function, $\sigma_{jt}(\theta, x_t, \xi_t)$ to the observed market shares in the sample, s_{jt}. By definition (Equation 3), σ is strictly positive; as a result, $\sigma_{jt}^{-1}(\theta, x_t, s_t)$ is undefined if $s_{jt} = 0$. This fact makes BLP unable to deal with products that have zero sales. The most common solution to this problem in practice is to exclude any zero-share products from estimation. However, the moment conditions used in BLP reflect expectations over all products, not just those with positive shares. Excluding zero-share products is akin to selecting on the products with higher values of ξ_{jt}; this selection violates the moment conditions and leads to biased estimation results. To see why, consider the following lemma.

LEMMA 1. *Let $\mathbb{E}_T[\cdot]$ denote the sample average of a variable or function in the observed sample. Markets $t = 1, 2, ..., T$ are independent. ξ_{jt} are independent across markets and satisfy Equation (5). When T is large,*

$$\mathbb{E}_T[\xi_{jt}(\theta_0, x_t, s_t) \,|\, z_t, s_{jt} > 0] \to \mathbb{E}[\xi_{jt} | z_t, S_{jt} > 0] > 0.$$

This lemma states that although each ξ_{jt} is mean independent of the instruments, excluding products with zero sales skews their expectation, and, thus, their conditional sample average in the data we observe. The proof of Lemma 1 is provided in Appendix B, but the following simple example provides the key intuition. Without loss of generalization, assume x_{jt} is a scalar and exogenous–that is, ξ_{jt} is mean independent of x_t–and that all J_t products have similar values of x_{jt}. In this case, the utility and market share of each product j will be driven entirely by its value of ξ_{jt}. The law of large numbers ensures that if T is large, the sample average of ξ_{jt} over all products will be close to zero. Now consider only the subset of products that have positive sales. As the observed market shares are driven by ξ_{jt}, it must be the case that on average, products with positive sales have higher values of ξ_{jt} than products with zero sales. As the overall average is near zero, the sample average of ξ_{jt} in these positive share products is positive, which yields the lemma.

Lemma 1 establishes that excluding zero-sale products is equivalent to selecting products with higher values of the unobservable ξ_{jt}. This selection violates the moment conditions in Equation (5), and renders the estimator in Equation (6) biased. Consider again the simple example we discussed above. Assume that the single covariate x_{jt} is non-negative and has a positive effect on market share, that is, $x_{jt} \geq 0$, $\forall j$, and its coefficient $\beta > 0$. For simplicity, assume that there are no random coefficients, and that $\theta = \beta$. In this simple model, the BLP estimator is equivalent to running the following ordinary linear regression, $\log s_{jt} - \log s_{0t} = \beta x_{jt} + \xi_{jt}$. Since $x_{jt} \geq 0$ and $\beta > 0$, products with high values of x_{jt} are more likely to generate higher utility and have positive sales, whereas products with low values of x_{jt} are more likely to generate lower utility and have zero sales. As

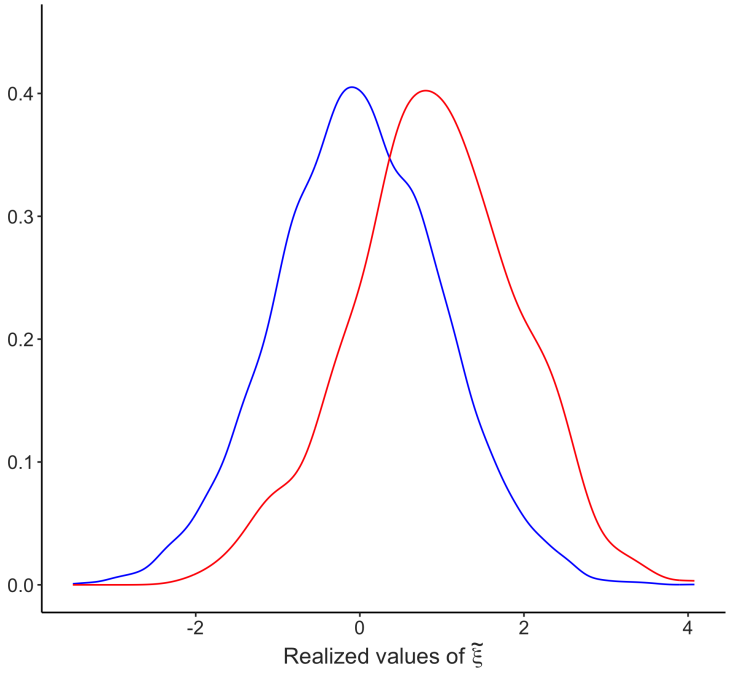


Figure 1 **Simulated distributions of ξ_{jt} for products with similar values of x_{jt}. The blue line represents the distribution of ξ_{jt} in all products, while the red line represents the distribution of ξ_{jt} only in products with positive sales. This plot clearly demonstrates that though the overall sample average of ξ_{jt} is near zero, it is positive if we exclude products with zero sales**

a result, when excluding zero-sale products, products with high values of x_{jt} are less likely to be exposed to the selection, while products with low values of x_{jt} are more likely to be exposed to the selection–i.e., those products with positive sales are more likely to have higher-valued ξ_{jt}. In other words, x_{jt} and ξ_{jt} are negatively correlated. Given the omitted variable bias formula, the estimated coefficient will have a negative bias. In this paper, we propose a two stage estimator that corrects for this bias, allowing us to obtain consistent, unbiased inference in the presence of selection. Section 3 discusses this estimator in detail; in the following section, we present its key intuition.

2.5. Overcoming Selection

We have now established why excluding the zero-share products from estimation can lead to biased estimates. In this paper, we propose a two stage estimator that corrects for this bias. Before sketching out the details of this method in the following section, it is worth briefly discussing the intuition behind this process. Broadly, there are two types of products with positive shares. The first category consists of those products which would have had non-zero sales regardless of their value of ξ_{jt}. We call these "safe" products, as they do not suffer from selection and consequently do not contribute to estimation bias. The second category consists of those products which likely

would have had zero sales without their high values of ξ_{jt}. These products resemble unsold products in their observable features, and are only included in the estimation set because of the high realized value of ξ_{jt}. If we were to include all products in estimation, these positive values of ξ_{jt} would be balanced by negative values in the unsold products. Excluding unsold products from estimation skews the sample expectation of ξ_{jt}: these products are thus "risky", and contribute heavily to the bias.

Our method first identifies risky products, then re-weights their observed market shares in a way that corrects for the bias in BLP estimates. Products that suffer from high degrees of selection are assigned low weights; this procedure reduces their importance in the estimation, minimizing the amount their high values of ξ_{jt} skew the sample expectation of the moment conditions. We discuss our weighting procedure in detail in the following section, and show that under weak assumptions, it produces consistent estimates in the number of markets.

3. Two Stage Estimator

The construction of our proposed estimator has two key stages. First, we use a neural network to predict each product's market share using only its product features. These predictions estimate the utility a product derives from its observed features; comparing them to the observed shares helps us identify how much each product suffers from selection. Second, we match products with similar predicted shares and re-weight the observed shares in a way that down-weights risky products. We then run standard BLP estimation using these weighted shares to obtain more accurate estimates of customer preferences.

3.1. Stage I: Market Share Prediction

In Stage I, we construct a model that predicts a product's market share π_{jt} given its features x_{jt}. We use this model to predict the market share of every product in our dataset, including those with zero sales. These predictions form the inputs to the weighting scheme in Stage II that corrects for the long-tail bias.

Our only goal in Stage I is prediction: we do not make any causal claims or impose any interpretations on the estimated model. As a result, we are free to use any supervised machine learning algorithm, parametric or non-parametric. While this exercise seems straightforward, it is complicated by the fact that a product's market share depends not only on its own features, but also on the features of every other product in the same market. Expanding each product's covariate set to include the features of all other products in its market is an ineffective workaround, as it is cumbersome and cannot handle different markets having different numbers of products. In this section, we detail an approach that uses deep learning to overcome this challenge. The structure of a multi-layer neural network facilitates modeling these intra-market dependencies in the final

output layer. We also set up our neural network to mirror the data generating process of the random coefficient MNL model, further facilitating accurate predictions. Note that using deep learning is not necessary; any machine learning or other statistical model that produces high quality predictions can be used for this stage of our method. However, the intra-market dependencies and long tail of market shares make this prediction task challenging, and our proposed method greatly outperforms other common approaches (see Appendix D for details).

Our proposed model consists of F individual predictors $f_i : x_t \to s_t$, each of which predicts a product's market share given its observed features. We average the predictions from all F models to obtain our final output. Later, we discuss the merits of this structure; for now, we focus on how these individual predictors are constructed. Let $f_i(x_{jt})$ denote the share f_i predicts for product j in market t. We can think of f_i as a composite of two functions: $\tilde{f}_i : x_{jt} \to u_{jt}$, which estimates the utility of each product, and deterministic mapping $g : u_t \to s_t$, which translates these estimated utilities into market shares using the MNL transform

$$f_i(x_{jt}) = g(\tilde{f}_i(x_{jt})) = \frac{e^{\tilde{f}_i(x_{jt})}}{1 + \sum_{k=1}^{J_t} e^{\tilde{f}_i(x_{kt})}} \tag{7}$$

While we could use any function to define $\tilde{f}_i$, this structure lends itself neatly to a multi-layer neural network. We define $\tilde{f}_i$ to be a neural network with trainable weights w_i that predicts utility. This network can be as simple or complex, deep or shallow as necessary; note that if we use a one layer network with no hidden layers or nonlinear activation, we recover the standard logit functional form $f_i(x) = w_i' x$. As we observe shares not utilities, we define g as the final layer of this neural network to map the utilities output by $\tilde{f}_i$ to market shares.

We have now defined a function f_i with free parameters w_i that predicts the market share of a product using its own covariates as well as the covariates of other products in the same market. We can solve for the value of w_i that minimizes the squared error between the observed and predicted shares (with gradients from backpropagation), then use these predictions in Stage II. However, no matter how complex we make f_i, this approach still has the independence of irrelevant alternatives (IIA) property. We can eliminate the reliance on this restrictive property by defining F functions f_i, and computing our final output as the average of each of these individual models

$$f(x_t) = \frac{1}{F} \sum_{i=1}^{F} \frac{e^{\tilde{f}_i(x_{jt})}}{1 + \sum_{k=1}^{J} e^{\tilde{f}_i(x_{kt})}} \tag{8}$$

We train this aggregate network by minimizing the following loss function

$$\mathcal{L}(w_1, ..., w_F) = \sum_{t=1}^{T} \sum_{j=1}^{J_t} (f(x_t) - s_{jt})^2 \tag{9}$$

The trained neural network can then be used to predict the market shares for each product without relying on the IIA property. In this setup, each individual neural network f_i represents a customer, while its free parameters w_i represent the customer's preferences. As the optimized w_i can vary from network to network, different "customers" are allowed to have different preferences, exactly as in the random coefficients MNL model. The structure of our neural network thus captures the data generating process assumed by Berry et al. (1995) and many other similar studies, and can produce accurate predictions without the IIA restriction. In fact, it is even more flexible than the standard random coefficients model: each $\tilde{f}_i$ can be defined to capture nonlinear relationships, and the network as a whole can allow for non-Gaussian heterogeneity in the customer base. These facets allow our approach to capture a wide variety of data generating processes and make accurate predictions.

Notice that our neural network models the relationship between the product's observed market share and its covariates for all products, including those with zero sales. By including all products, we ensure that our neural network effectively estimates $\mathbb{E}[S_{jt}|x_t]$, a product's expected market share conditional on its observed features. This expectation is taken over the unobservable ξ_{jt} and is conditional on the observed covariates; it thus provides a way to assess how risky or safe a product is. If $\mathbb{E}[S_{jt}|x_t]$ is large, then product j in market t is likely to have non-zero sales regardless of its value of ξ_{jt}. These products are safe and do not contribute to the bias caused by the selection. However, if $\mathbb{E}[S_{jt}|x_t]$ is low, then it is possible for product (j,t) to have zero share, depending on its value of ξ_{jt}. These products are risky; using some of them for estimation but not others (i.e. excluding those with zero share) violates BLP's moment condition (Equation (5)) and leads to biased results.

The next section discusses how to use these estimates of $\mathbb{E}[S_{jt}|x_t]$ to correct for BLP's estimation bias. Before moving on, note that when we fit the Stage I neural network in practice, we use k-fold cross fitting to obtain the final predictions of market share. This approach is analogous to that described in Chernozhukov et al. (2018). We split the dataset into k folds, where each fold contains T/k markets. We then hold out the first fold, train the neural network on the remaining $(k-1)$ folds, and predict the shares of products in the held out markets. Repeating this process for every fold, we obtain market share predictions for all products in the dataset. This procedure avoids overfitting, yielding high quality predictions of market shares that improve the performance of the estimation routine in Stage II. This approach also provides an easy way to specify the number of individual predictors, F, which can be chosen to maximise out-of-sample predictive accuracy (see Appendix C for further details).

3.2. Stage II: Structural Estimation

Stage I predicts the market share of every product in the dataset given their observed covariates. In Stage II, we use these predictions to re-weight the observed market shares and estimate the structural parameters. From now on, we only include products that have non-zero sales. The set of products we run the Stage II estimation on is thus exactly the same as in BLP; our weighting scheme, however, corrects for the bias induced by excluding zero-share products.

3.2.1. Goal Before we get into the procedural details of this stage, we first establish our overall goal. Recall that the reason BLP estimation is biased for long tail markets is that its moment conditions are violated when we drop products with zero sales. Specifically, Lemma 1 implies that $\mathbb{E}_T[\xi_{jt}(\theta_0, x_t, s_t) \,|\, z_t, s_{jt} > 0] > 0$ at the true $\theta = \theta_0$. To correct for the bias, we must instead use moment conditions that do hold (i.e. equal 0) at the true parameters. Our method accomplishes this goal by re-weighting the observed market shares s_t so that the products that contribute more to the bias contribute less to the moment conditions. If w_t is a vector that represents the weights assigned to products in market t, the weighting scheme should satisfy the following property

$$\mathbb{E}_T[\xi_{jt}(\theta_0, x_t, w_t s_t) \,|\, s_{jt} > 0, z_t] \to 0, \tag{10}$$

as $T \to \infty$.

3.2.2. Weighting If we have access to weights w_t that satisfy Equation (10), then the estimation is easy. All we need to do is to run BLP using the re-weighted shares $w_t s_t$ instead of the observed shares s_t. We use our Stage I predictions to construct such a weighting scheme. First, sort all products (with positive sales) across all markets by their predicted shares, $\hat{\pi}_{jt}$, and partition them into B bins, grouping products that have similar values of $\hat{\pi}_{jt}$. We choose the number of bins B to be the smallest number that sufficiently reduces the intra-bin variation in $\hat{\pi}_{jt}$. The details of this process are discussed in in Appendix C. Now, consider the k^{th} bin B_k; to every product $b \in B_k$ with $s_b > 0$, we denote their predicted share $\hat{\pi}_b$ and assign the following weight

$$\hat{w}_b = \frac{\frac{1}{|B_k|} \sum_{b' \in B_k, s_{b'} > 0} \hat{\pi}_{b'}}{\frac{1}{|B_k|} \sum_{b' \in B_k, s_{b'} > 0} s_{b'}}. \tag{11}$$

A product's weight is thus the ratio between the average predicted market share and the average observed market share over all products with positive sales in the same bin. As the predicted shares $\hat{\pi}_{jt}$ estimate $\mathbb{E}[S_{jt}|x_t]$, products in the same bin derive similar amounts of utility from their observed features. A bin's weight thus reflects the average degree of selection it exhibits. To see this, first consider bins with high predicted shares. In these bins, few to no products are dropped. As these products would have had positive sales no matter their value of ξ_{jt}, they do not exhibit

selection on unobservables. For these products, the predicted share is as likely to overestimate the observed share as to underestimate it. The predicted shares will thus be close to the observed shares on average, and the products will be assigned $\hat{w}_{jt} \approx 1$. On the other hand, consider bins with low predicted shares. Products in these bins would likely have gone unselected without their high values of ξ_{jt}. Much of these products' utility is derived from the unobserved feature; their average observed share will hence be larger than their average predicted share, and they will be assigned $\hat{w}_{jt} < 1$. A product's weight thus reflects how safe or risky it is: the more the product suffers from selection on the unobservable, the more it contributes to the bias, and the smaller the assigned weight. Downweighting the risky products reduces their contribution to the moment conditions used in BLP estimation and shrinks the expected value of ξ_{jt} at $\theta = \theta_0$ to zero, even in the presence of selection.

3.2.3. Orthogonalization Our proposed weighting scheme satisfies condition 10 (see Appendix E for full proof); we can thus replace the observed shares s_t with our re-weighted shares $\hat{w}_t s_t$ and run standard BLP to obtain improved estimates of θ_0. However, since we do not observe the true weight w_t, and $\hat{w}_t$ is estimated using neural network in Stage I, the resulting $\hat{\theta}$ in Stage II is biased. To make the estimates $\sqrt{T}$ consistent, we need to "orthogonalize" the BLP moment condition using the double machine learning approach (Chernozhukov et al. 2018). This modification makes BLP robust to small biases in the neural network estimates of w from regularization or overfitting. To do so, we replace the moment conditions $z_t'\xi_t$ with $\hat{\mu} z_t'\xi_t$, where $\hat{\mu}$ is a $d_\theta \times d_z$ matrix that estimates the true orthogonalization parameter μ_0. We then define our two stage (TS) estimator as follows

$$\hat{\theta}_{TS} \equiv \arg\min_{\theta} \Big(\frac{1}{T} \sum_{t=1}^{T} \hat{\mu}\, z_t \xi_t(\theta, x_t, \hat{w}_t s_t) \Big)' W \Big(\frac{1}{T} \sum_{t=1}^{T} \hat{\mu} z_t \xi_t(\theta, x_t, \hat{w}_t s_t) \Big). \tag{12}$$

Computing $\hat{\mu}$ is straightforward, and relies mostly on values needed to compute the standard BLP estimator (see Appendix F for details). Moreover, it is easy to incorporate $\hat{\mu}$ into the estimation as we can interpret the orthogonalization as a re-definition of the GMM weighting matrix. Inspecting equation 12 reveals that orthogonalizing the moment condition is equivalent to replacing weighting matrix W with $\bar{W} = \hat{\mu}' W \hat{\mu}$. [1] In other words, we can use an existing BLP routine to solve for $\hat{\theta}_{TS}$. Examples of BLP packages include PyBLP in Python (Conlon and Gortmaker 2019) and the MATLAB implementation of the MPEC method (Dubé et al. 2012). The debiased estimation thus requires minimal additional work compared to standard BLP and is easy to implement.

[1] Note that in practice, the original weighting matrix W is often set to be the identity, though practitioners may also use an estimate of the optimal GMM weighting matrix.

We have now described the second stage of our estimation process. In Appendix A, we provide an algorithm that succinctly summarizes our method, as well as more detail on computing the orthogonalization parameter $\hat{\mu}$. Next, we briefly discuss our estimators asymptotic properties, before moving on to its performance on simulated and real data.

3.3. Inference of the Two Stage Estimator

Our estimator $\hat{\theta}_{TS}$ is $\sqrt{T}$ consistent and asymptotically normal as the number of markets T goes to infinity. As long as the Stage I estimates are sufficiently predictive and the chosen number of draws R grows faster than $\sqrt{T}$, we obtain consistent estimates of customer preferences even in the presence of the long tail. These properties follow directly from the results provided in Freyberger (2015) and Chernozhukov et al. (2018). As our paper's focus is more on the empirical contributions of our method, we do not discuss the asymptotic theory in detail; however, we provide a full statement of our estimator's properties, along with the underlying assumptions and proofs, in Appendix G. The following section also demonstrates this theory in practice; our simulation experiments clearly show that the re-weighting method does indeed correct for the bias induced by excluding the zero-share products.

4. Monte Carlo Simulations

Since the objective of our method is to obtain the parameter estimates without bias, before applying our method in real data where we do not observe the true parameter values, we conduct extensive simulation studies to demonstrate the performance of our method. Broadly, we conduct two sets of simulations: first with all exogenous features (i.e. where the unobservable is mean-independent of all observed product features), then with one endogeneous feature like most demand estimation settings where price is endogenous. Within each set, we consider a variety of challenging settings, and compare our method's results with BLP (Berry et al. 1995) and the bound estimator (Gandhi et al. 2020). We demonstrate that our estimator outperforms both other methods, producing far less biased estimates of all parameters.

4.1. Exogenous Features

4.1.1. Simulation Setup Our setup is similar to that used in Gandhi et al. (2020). We simulate $T = 25$ markets, each with $J_t = 50$ products and $N = 10,000$ customers. The utility of customer i purchasing product j in market t is given by

$$u_{ijt} = \beta_0 + \sum_{m=1}^{5} \beta_i^{(m)} x_{jt}^{(m)} + \xi_{jt} + \epsilon_{ijt}$$

where ϵ_{ijt} is the standard type I extreme value error term and $x_{jt}^{(m)}$ is the m^{th} covariate of of product j in market t. Unobserved characteristics are randomly generated from a normal distribution $\xi_{jt} \sim$

$\mathcal{N}(0, \sigma_\xi^2)$. Each $\beta_i^{(m)}$ is a random variable with mean $\bar{\beta}^{(m)}$ and standard deviation $\lambda^{(m)}$, $\beta_i^{(m)} \sim \mathcal{N}(\bar{\beta}^{(m)}, (\lambda^{(m)})^2)$. In our simulations, $\bar{\beta}^{(m)} = 1$ and $\lambda^{(m)} = 0.5$ for all $m \in \{1, ..., 5\}$. Along with β_0, these are the causal parameters to be estimated.

Of the five covariates, the first $x_{jt}^{(1)}$ is a vertical feature, and is generated as

$$x_{jt}^{(1)} = \frac{j}{10} + \mathcal{N}(0, 1).$$

The same product offered in different markets will have correlated values of the vertical feature. This feature thus resembles price, which is similar but not identical across markets. The other four covariates are non-vertical and are randomly drawn from a standard normal distribution. In this section, we assume that all these features are exogenous: $x_{jt} \perp \xi_{jt} \, \forall \, j, t$. This assumption will be relaxed in the next section. We further generate instruments $z_{jt} = (x_{jt}, x_{jt}^2, x_{jt}^3 - 3x_{jt})$ for each product j in market t; this follows the choice in Gandhi et al. (2020).

We have now specified all the variables required to generate market shares except β_0 and σ_ξ. β_0 reflects the popularity of the outside option: the more negative its value, the more products have zero sales and the longer the tail. σ_ξ, on the other hand, controls the variation of the unobserved feature. The higher its value, the less the observed covariates impact utility, and the more difficult estimation is. We allow β_0 and σ_ξ to take on one of three values, $\beta_0 = -10, -12, -14$ and $\sigma_\xi = 0.5, 1.0, 1.5$. For each combination of these values, we use Monte Carlo simulations to generate 1,000 different replications of market shares. We use our two stage estimator to compute the mean estimates of the popularity of the outside option β_0, the mean consumer preference $\bar{\beta}$, and the customer heterogeneity λ, as well as the standard error of these estimates. We use $F = 100$ one-layer neural networks in Stage I, $B = 50$ equally sized product bins to calculate the product weights, and $R = 200$ random draws to estimate the market shares. Finally, we calculate the corresponding estimates using BLP and the bound estimator (Gandhi et al. 2020), and compare the results in the following section. For the sake of brevity, though we model five covariates, we only display results for the intercept, the vertical feature, and the first non-vertical feature; the complete set of results is presented in Appendix H.

4.1.2. Results Table 2 demonstrates the simplest case we consider, where a modest proportion of products (around 26%) are unsold and the unobservable feature has low variance. Even in this conservative scenario, our two stage estimator outperforms both other methods. Our estimate of the mean preference $\bar{\beta}$ and customer heterogeneity λ differ from the true value by 1-2%, while the bound estimator (Gandhi et al. 2020) and BLP both show biases of 10% and greater. We observe a similar pattern in the intercept (β_0) estimates, where our method again shows 3-4% bias compared to 7-8% for the other two methods.

We now make this estimation problem more challenging in two ways. First, we make the outside option more attractive by reducing the true value of β_0 (see Table 3). This change increases the proportion of zero-share products from 26% to 59%. This is a challenging problem, as over half of the product set must be excluded from estimation. As a result, BLP yields inaccurate estimates, particularly for the coefficients of the vertical feature ($\bar{\beta}^{(1)}$, $\lambda^{(1)}$). While Gandhi et al. (2020) performs better, it still shows biases of around 20% in its estimates of the intercept, mean preference, and customer heterogeneity. On the other hand, our re-weighting procedure is less affected: even using less than half of all products, our estimates of the vertical and non-vertical coefficients fall within 5% of the true values.

Next, we increase the variance in the unobserved feature ξ. This is a fundamentally difficult problem, as the observed features contribute less to the product's overall market share. The estimators proposed by BLP and Gandhi et al. (2020) reflect this difficulty (see Table 4); the latter in particular vastly underestimates the customer heterogeneity λ. Our two stage estimator, however, still performs well. Though the neural network in Stage I is less predictive of overall share, the re-weighting allows us to correct for most of the bias induced by excluding the long tail. Our method is thus more robust not only to a more popular outside option, but also to a more influential unobservable.

The discussed improvement in accuracy holds across other combinations of β_0 and σ_ξ; these results are provided in Appendix H, and further demonstrate our method's potential. To conclude this set of simulations, consider an extreme case with an extremely popular outside option ($\beta_0 = -20$). This setting corresponds to an extremely long tail, with the percentage of zero-sale product close to 95%. To account for the much smaller number of products used in the second stage of our estimation, we reduce the number of bins to 20 (more details on bin selection are included in Appendix C). Table 5 summarizes the results. This is a very challenging problem, as the estimation relies on a very small number of observations; it is thus unsurprising that all three methods show more biased results compared with the other simulation settings. However, the two stage estimator is the most accurate in its estimate of thec intercept, vertical coefficients, and non-vertical coefficients. While this extreme example shows our method's limitations, it is worth noting that the two stage estimator still offers the best solution to a difficult estimation problem.

4.2. Endogenous Feature

4.2.1. Simulation Setup In the previous section, we assumed that all product features were exogenous, and $x_{jt}^{(m)} \perp \xi_{jt} \, \forall \, m, j, t$. We now relax this assumption, and allow for correlation between the vertical feature $x_{jt}^{(1)}$ and the unobserved feature ξ_{jt}. The vertical feature and its first instrument $z_{jt}^{(1)}$ are now generated as follows:

$$z_{jt}^{(1)} = \frac{j}{10} + \mathcal{N}(0, 0.5^2)$$

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-10.000	-10.331	-9.284	-9.296
		(0.016)	(0.004)	(0.003)
$\bar{\beta}^{(1)}$	1.000	1.016	0.830	0.890
		(0.002)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	1.012	0.898	0.886
		(0.001)	(0.001)	(0.001)
$\lambda^{(1)}$	0.500	0.497	0.664	0.450
		(0.003)	(0.002)	(0.001)
$\lambda^{(2)}$	0.500	0.504	0.496	0.449
		(0.001)	(0.002)	(0.001)

Table 2 **Baseline case,** 26% **of products unsold (**$\beta_0 = -10$**,** $\sigma_\xi = 0.5$**)**

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-14.000	-14.300	-11.565	-12.136
		(0.026)	(0.014)	(0.006)
$\bar{\beta}^{(1)}$	1.000	0.984	0.123	0.818
		(0.012)	(0.010)	(0.002)
$\bar{\beta}^{(2)}$	1.000	1.012	0.827	0.792
		(0.002)	(0.002)	(0.001)
$\lambda^{(1)}$	0.500	0.518	0.848	0.398
		(0.007)	(0.006)	(0.001)
$\lambda^{(2)}$	0.500	0.493	0.474	0.418
		(0.002)	(0.004)	(0.002)

Table 3 **Popular Outside Option,** 59% **of products unsold (**$\beta_0 = -14$**,** $\sigma_\xi = 0.5$**)**

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-10.000	-11.170	-8.429	-8.697
		(0.050)	(0.007)	(0.005)
$\bar{\beta}^{(1)}$	1.000	0.992	0.718	0.805
		(0.007)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	1.052	0.770	0.781
		(0.006)	(0.002)	(0.002)
$\lambda^{(1)}$	0.500	0.571	0.637	0.345
		(0.006)	(0.004)	(0.002)
$\lambda^{(2)}$	0.500	0.535	0.456	0.348
		(0.004)	(0.004)	(0.004)

Table 4 **Influential Unobservable,** 33% **of products unsold (**$\beta_0 = -10$**,** $\sigma_\xi = 1.5$**)**

$$x_{jt}^{(1)} = z_{jt}^{(1)} + e_{jt}$$

where e_{jt} is a new error term that is correlated with unobservable ξ_{jt}. Specifically, (e_{jt}, ξ_{jt}) are drawn from a bivariate normal distribution with covariance term $\sigma_{e\xi}$. $\sigma_{e\xi}$ represents the strength of confounding: the higher its value, the stronger the dependence between $x_{jt}^{(1)}$ and ξ_{jt}. The other

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-20.000	-21.420	-13.386	-13.159
		(0.107)	(0.039)	(0.034)
$\bar{\beta}^{(1)}$	1.000	1.368	0.346	0.420
		(0.022)	(0.014)	(0.007)
$\bar{\beta}^{(2)}$	1.000	0.951	0.475	0.488
		(0.009)	(0.005)	(0.004)
$\lambda^{(1)}$	0.500	0.342	0.348	0.345
		(0.008)	(0.006)	(0.002)
$\lambda^{(2)}$	0.500	0.411	0.352	0.309
		(0.008)	(0.008)	(0.006)

Table 5 Extremely Popular Outside Option, 95% of products unsold ($\beta_0 = -20$, $\sigma_\xi = 0.5$)

features, instruments, utilities, and market shares are generated as in the exogenous case, and we use the same hyperparameters for our two stage estimator. We vary the values of the intercept β_0, the variance of the unobserved feature σ_ξ^2, and the covariance $\sigma_{e\xi}$, and compute our two stage estimate for each setting, and compare it to the corresponding results from BLP and the bound estimator.

4.2.2. Results We first consider a simple baseline case (Table 6), with a modest proportion of zero share products and low correlation between the unobservable and vertical features. For the sake of brevity, we again only display results for the intercept, vertical feature and first non-vertical feature (complete results are presented in Appendix H). As in the non-endogenous case, our two stage method outperforms the other two even in this simple baseline. We now make our problem more challenging in three ways: introducing a more popular outside option (Table 7), increasing the influence of the unobservable (Table 8), and increasing the degree of endogeneity (Table 9). Each of these three cases reflects the limitations of BLP, which yields mean estimates that are severely biased towards zero. While the bound estimator succeeds at correcting some of this inaccuracy, it still shows significant bias, producing estimates that are often 15-25% away from the true values. These biases are particularly pronounced for the coefficients of the non-vertical feature. Our method offers a significant improvement over both these methods: in almost all cases, our estimates are within 5-10% of the ground truth.

We conclude this section by considering a scenario that has all three of the previously discussed challenges: high endogeneity, a popular outside option, and an influential unobservable (Table 10). This problem is difficult, as it involves a large proportion of zeros as well as a weak instrument. It is thus unsurprising that our two stage method shows some bias in its estimate of $\bar{\beta}_1$. However, this bias is less than that shown by both the bound estimator and BLP; moreover, our method produces accurate estimates of all other coefficients, significantly outperforming both other approaches.

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-10.000	-10.079	-9.179	-9.345
		(0.022)	(0.004)	(0.003)
$\bar{\beta}^{(1)}$	1.000	1.010	0.834	0.922
		(0.003)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	0.977	0.875	0.878
		(0.001)	(0.001)	(0.001)
$\lambda^{(1)}$	0.500	0.462	0.643	0.453
		(0.005)	(0.003)	(0.001)
$\lambda^{(2)}$	0.500	0.500	0.490	0.450
		(0.002)	(0.002)	(0.001)

Table 6 Baseline Endogenous Case, 28% of products unsold
($\beta_0 = -10$, $\sigma_\xi = 0.5$, $\sigma_{e\xi} = 0.25$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-14.000	-13.762	-11.511	-12.200
		(0.028)	(0.012)	(0.007)
$\bar{\beta}^{(1)}$	1.000	0.996	0.329	0.854
		(0.011)	(0.009)	(0.002)
$\bar{\beta}^{(2)}$	1.000	0.936	0.782	0.781
		(0.002)	(0.002)	(0.001)
$\lambda^{(1)}$	0.500	0.458	0.715	0.402
		(0.008)	(0.006)	(0.001)
$\lambda^{(2)}$	0.500	0.479	0.452	0.415
		(0.002)	(0.003)	(0.002)

Table 7 Endogenous Case with Popular Outside Option, 59% of products unsold
($\beta_0 = -14$, $\sigma_\xi = 0.5$, $\sigma_{e\xi} = 0.25$)

While this example shows the limits of our method, it is encouraging that it is still able to produce state of the art results on a challenging estimation problem.

Overall, this set of simulations demonstrates that our estimator yields accurate results in the presence of endogeneity. As most empirical problems involve endogenous prices, this result is extremely encouraging. Note that as with the non-endogenous case, we include more simulations with other choices of parameter values in Appendix H, the results of which further demonstrate that the two stage estimator offers an elegant, accurate solution to the long tail problem.

5. Empirical Application

Now that we have shown that our method performs well in simulation experiments, we demonstrate its performance on real data. To do so, we use the publicly available Dominick's Finer Foods (DFF) dataset, which contains store-level scanner data collected at 93 DFF stores in and around Chicago over nearly eight years (1989-1997). The dataset includes sales, prices, and other key information

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-10.000	-10.676	-8.369	-8.735
		(0.060)	(0.007)	(0.005)
$\bar{\beta}^{(1)}$	1.000	0.968	0.734	0.828
		(0.008)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	0.989	0.758	0.777
		(0.006)	(0.002)	(0.002)
$\lambda^{(1)}$	0.500	0.570	0.603	0.346
		(0.009)	(0.005)	(0.002)
$\lambda^{(2)}$	0.500	0.516	0.449	0.339
		(0.004)	(0.004)	(0.004)

Table 8 Endogenous Case with Influential Unobservable, 33% of products unsold ($\beta_0 = -10$, $\sigma_\xi = 1.5$, $\sigma_{e\xi} = 0.25$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-10.000	-9.873	-9.062	-9.448
		(0.024)	(0.004)	(0.003)
$\bar{\beta}^{(1)}$	1.000	1.008	0.827	0.965
		(0.003)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	0.944	0.854	0.878
		(0.001)	(0.001)	(0.001)
$\lambda^{(1)}$	0.500	0.400	0.635	0.461
		(0.007)	(0.003)	(0.001)
$\lambda^{(2)}$	0.500	0.496	0.487	0.453
		(0.002)	(0.002)	(0.001)

Table 9 Highly Endogenous Case, 29% of products unsold ($\beta_0 = -10$, $\sigma_\xi = 0.5$, $\sigma_{e\xi} = 0.5$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-14.000	-13.864	-10.355	-11.531
		(0.051)	(0.015)	(0.009)
$\bar{\beta}^{(1)}$	1.000	0.920	0.324	0.854
		(0.018)	(0.010)	(0.003)
$\bar{\beta}^{(2)}$	1.000	0.911	0.641	0.686
		(0.004)	(0.003)	(0.002)
$\lambda^{(1)}$	0.500	0.504	0.610	0.299
		(0.013)	(0.009)	(0.002)
$\lambda^{(2)}$	0.500	0.474	0.389	0.319
		(0.005)	(0.006)	(0.005)

Table 10 Highly Endogenous Case with Popular Outside Option and Influential Unobservable, 60% of products unsold ($\beta_0 = -14$, $\sigma_\xi = 1.5$, $\sigma_{e\xi} = 0.5$)

at product-store-week level; we will specifically focus on the canned tuna category, which has

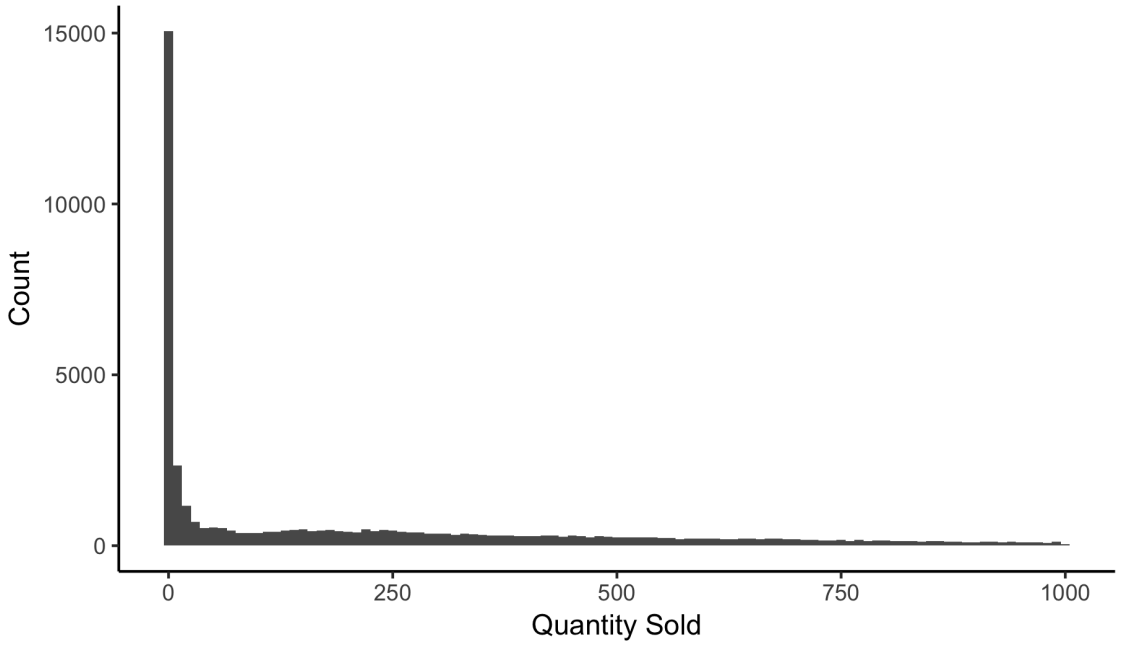


Figure 2 Histogram of weekly sales of tuna aggregated across DFF stores. Note that each bin covers 10 sales. The data clearly displays the long tail pattern, as a large number of UPCs sell less than ten units in a given week

previously been studied by Chevalier et al. (2003), Nevo and Hatzitaskos (2006) and Gandhi et al. (2020).

Before we present the results of our estimation, we determine whether the data does in fact display the long tail pattern. The dataset tracks the weekly sales of 278 canned tuna products, each identified by its Universal Product Code (UPC). In line with previous research (Nevo and Hatzitaskos 2006), we aggregate the sales of each of these products across all stores and conduct our analysis at the UPC-week level. This yields 53,888 market share observations over 373 weeks, where each week is considered a market. 10,401 of these product-week pairs–nearly 20% of the total–have zero sales, indicating that the canned tuna category does indeed exhibit the long tail pattern. Figure 2 further illustrates this point.

In this section, we use this long tail data to study how demand for tuna changes during Lent, a popular period for this product. Previous research has found that the price index for tuna (i.e. prices weighted by market share) falls significantly during Lent. While Chevalier et al. (2003) suggests that this decline is due to loss-leading behaviour from manufacturers, Nevo and Hatzitaskos (2006) and Gandhi et al. (2020) attribute this fall to a change in price sensitivity: demand becomes more elastic in the Lent period, leading to an increase in relative demand for cheaper products. Nevo and Hatzitaskos (2006) uses a multinomial logit demand model to substantiate their claims; however, they exclude the long tail of UPCs, including only the thirty best selling products. Gandhi et al. (2020) builds on the analysis in Nevo and Hatzitaskos (2006), and uses its proposed bound method

to estimate price elasticity using a subset of UPCs. However, neither of the analyses uses the random coefficients specification; they thus do not model customer heterogeneity, and are prone to the IIA restriction of the MNL model. We build on this prior work and investigate the same problem using a random coefficient MNL approach.

The simplest way to conduct our analysis would be to use BLP. However, this data shows the long tail pattern, and using BLP naively may yield biased estimates. We thus use our two stage method to analyze demand for canned tuna, and compare our results with those obtained by using BLP. For both methods, we use the following random coefficient MNL specification:

$$u_{ijt} = (\beta_p + \lambda_p \nu_i) p_{jt} + \beta_x x_{jt} + \xi_{jt} + \epsilon_{ijt}$$

where p_{jt} is the price of product j in week t, x_{jt} is a set of control variables that includes a product fixed effect and a time trend, and β_p and λ_p represent the two quantities of interest–the mean price coefficient and the variation in consumer price preferences. As the demand shock ξ_{jt} is likely correlated with price, we use wholesale costs as an instrument for price. This specification is the same as in Nevo and Hatzitaskos (2006) and Gandhi et al. (2020), only adding a random coefficient for price to avoid the undesirable IIA property of the pure MNL model.

Using this specification, we leverage our two stage method to estimate β_p and λ_p for the entire dataset. In the first stage, we use a neural network (as described in Section 3.1) to predict product market shares from the observed features: price, week, and a UPC fixed effect. Note that as price is unavailable for unsold products, we impute it from the price of the same product in the closest week that it was sold. For the vast majority of cases, this amounts to averaging the price of the product in the week before and the week after the unsold week[2]. We then use the predicted market shares to construct 350 bins, and conduct our re-weighted estimation as discussed in Section 3.2. We also run standard BLP on the same data, and compare our results in Table 11. Immediately, we notice that our two stage estimator has a significantly more negative coefficient for price, implying that demand for canned tuna is more elastic than BLP would suggest. This underestimation makes intuitive sense: BLP is forced to drop unsold products, which are likely to be more elastic than those with high sales. As a result, the estimated average price elasticity (1.47) is significantly less than that suggested by our estimator (1.77). Moreover, our method has a larger estimate for the variance term, which suggests that BLP may also underestimate customer heterogeneity.

Next, we consider the same question addressed in Nevo and Hatzitaskos (2006) and Gandhi et al. (2020): do prices become more elastic during Lent? To do so, we separately estimate each model on the Lent and non-Lent weeks, and compare the results in Table 12. Our method supports

[2] These imputations are only used in Stage I of the estimation and not Stage II, since Stage II only includes positive-sale products

	BLP	Two Stage Estimator
Price Coefficient: Mean	−0.71	−0.86
	(0.01)	(0.01)
Price Coefficient: Variance	0.29	0.37
	(0.02)	(0.03)
Avg. Own-Price Elasticity	1.47	1.77

Table 11 Results of demand estimation using the entire DFF dataset.

	BLP		Two Stage Estimator	
	Lent	Non-Lent	Lent	Non-Lent
Price Coefficient: Mean	−0.64	−0.71	−1.00	−0.87
	(0.03)	(0.01)	(0.05)	(0.01)
Price Coefficient: Variance	0.25	0.28	0.38	0.38
	(0.07)	(0.03)	(0.14)	(0.04)
Avg. Own-Price Elasticity	1.34	1.47	2.06	1.82

Table 12 Results of demand estimation on the DFF dataset, separately estimating the Lent and non-Lent weeks

the findings from prior research: price elasticity rises significantly during the Lent period of high demand. This suggests that the fall in price index observed by Chevalier et al. (2003) is indeed due to a reallocation of demand to less expensive products. However, if we were to use BLP for this analysis, we would arrive at the opposite conclusion! According to the BLP estimates, prices become less elastic during Lent. These estimates are very likely inaccurate: the long tail biases the mean price coefficient towards zero in both the Lent and non-Lent periods, but the magnitude of the bias in the Lent period may be larger due to the smaller sample size. Our method seems to correct for this error, and yields conclusions that are consistent with prior research.

Finally, we illustrate the managerial implications of our method's estimates. Suppose that we are in charge of setting the price of a specific tuna product during the Lent period: Chicken of the Sea Chunk Light Tuna in Oil. Based on prior research that demand becomes more elastic in Lent, we may consider reducing the price during Lent to increase revenue. We consider offering a 20% reduction in price during Lent period, and use both BLP and our two stage estimator to estimate the impact of this markdown. Computing counterfactual estimates using BLP paints a negative picture: slashing the product's price by 20% would lead to a 2.4% reduction in total revenue. However, our two stage method yields the opposite conclusion: such a price reduction would boost revenue by 3.6%. Given that BLP underestimates own-price elasticity in the long tail setting, it makes sense that our estimate would provide a more optimistic view of such a markdown.

We repeat this same exercise for all other products, and present our results in Table 13. Overall, we find that a 20% price markdown is a revenue-increasing strategy for the median product with

Avg. Weekly Lent Sales	Number of UPCs	Median Estimated Revenue Change	
		BLP	Two Stage Estimator
1 - 10	53	0.1%	5.5%
11 - 100	77	-2.6%	4.7%
101 - 1000	103	-7.5%	1.2%
$\geq$ 1001	26	-11.7%	-4.0%

Table 13 The median estimated revenue change from offering a 20% price reduction in Lent for a single product, categorized by average weekly sales.

small or medium weekly sales (≤ 1000), but a revenue-decreasing strategy for the median product with high sales (> 1000). This trend is intuitive: larger, more popular brands tend to be less elastic, and would likely lose revenue by cutting prices. However, a 20% markdown can be an effective strategy for smaller, more elastic products. BLP, on the other hand, indicates that a Lent discount is a revenue-decreasing strategy for all products, regardless of historical sales. Correcting the bias inherent in BLP in long tail settings is thus imperative to make effective managerial decisions.

Overall, we see that our method provides an easy and accurate way to estimate a random coefficient MNL model in a long tail setting. Our method produces results that are in line with prior research, and can be used to make more effective managerial decisions (e.g. on pricing) than BLP. We have thus demonstrated that our two stage estimator can be impactful in a real empirical setting.[3] Our empirical results, when added to our simulation experiments, provide strong evidence that our two stage estimator corrects for the bias induced in BLP in long tail data, and can be used to better inform managerial decisions in such settings.

6. Conclusion

In this paper, we propose a two-stage estimator for the widely used BLP model in demand estimation when the data has long tail. We combine machine learning methods with structural estimation to correct for the bias in the classic BLP estimator applied to long tail data. We show that our estimator is consistent in large sample and performs well in simulations. Moreover, the method is easily implementable in BLP-type of settings. Our method demonstrates how machine learning tools can be applied to improve commonly used structural estimation methods with applications in economics, marketing, and operations management.

[3] Ideally, we would like to have compared the results of our method to those of the bound estimator (Gandhi et al. 2020); however, their estimator failed to converge on this dataset due to the large number of fixed effects. We also find that their method leads to unreasonable estimates when using a much smaller subset of products in the data and without random coeffecient. Results are available upon request.

References

Aksoy-Pierson M, Allon G, Federgruen A (2013) Price competition under mixed multinomial logit demand functions. *Management Science* 59(8):1817–1835.

Allenby GM, Rossi PE (1998) Marketing models of consumer heterogeneity. *Journal of econometrics* 89(1-2):57–78.

Allon G, Federgruen A, Pierson M (2011) How much is a reduction of your customers' wait worth? an empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management* 13(4):489–507.

Amano T, Rhodes A, Seiler S (2019) Large-scale demand estimation with search data .

Bajari P, Nekipelov D, Ryan SP, Yang M (2015) Machine learning methods for demand estimation. *American Economic Review* 105(5):481–85.

Berry S, Levinsohn J, Pakes A (1995) Automobile Prices in Market Equilibrium. *Econometrica: Journal of the Econometric Society* 841–890.

Besbes O, Sauré D (2016) Product assortment and price competition under multinomial logit demand. *Production and Operations Management* 25(1):114–127.

Breiman L (2001) Random forests. *Machine learning* 45(1):5–32.

Cachon GP, Olivares M (2010) Drivers of finished-goods inventory in the us automobile industry. *Management Science* 56(1):202–216.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters.

Chevalier JA, Kashyap AK, Rossi PE (2003) Why don't prices rise during periods of peak demand? evidence from scanner data. *American Economic Review* 93(1):15–37.

Conlon C, Gortmaker J (2019) Best practices for differentiated products demand estimation with pyblp. *Unpublished Manuscript* .

Dubé JP, Fox JT, Su CL (2012) Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica* 80(5):2231–2267.

Dubé JP, Hortaçsu A, Joo J (2021) Random-coefficients logit demand estimation with zero-valued market shares. *Marketing Science* 40(4):637–660.

Freyberger J (2015) Asymptotic theory for differentiated products demand models with many markets. *Journal of Econometrics* 185(1):162–181.

Gandhi A, Lu Z, Shi X (2020) Estimating demand for differentiated products with zeroes in market share data. *Available at SSRN 3503565* .

Guajardo JA, Cohen MA, Netessine S (2015) Service competition and product quality in the us automobile industry. *Management Science* 62(7):1860–1877.

He P, Zheng F, Belavina E, Girotra K (2018) Customer preference and station network in the london bike share system. *Columbia business school research paper* (18-20).

Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67.

Jagabathula S, Subramanian L, Venkataraman A (2020) A conditional gradient approach for nonparametric estimation of mixing distributions. *Management Science* .

Jiang ZZ, Li J, Zhang D (2020) A high-dimensional choice model for online retailing. *Jun and Zhang, Dennis, A High-Dimensional Choice Model for Online Retailing (September 6, 2020)* .

Kabra A, Belavina E, Girotra K (2019) Bike-share systems: Accessibility and availability. *Management Science* .

Nevo A (2001) Measuring market power in the ready-to-eat cereal industry. *Econometrica* 69(2):307–342.

Nevo A, Hatzitaskos K (2006) Why does the average price paid fall during high demand periods? Technical report, CSIO working paper.

Quan TW, Williams KR (2018) Product variety, across-market demand heterogeneity, and the value of online retail. *The RAND Journal of Economics* 49(4):877–913.

Rusmevichientong P, Shmoys D, Tong C, Topaloglu H (2014) Assortment optimization under the multinomial logit model with random choice parameters. *Production and Operations Management* 23(11):2023–2039.

Sudhir K (2001) Competitive pricing behavior in the auto market: A structural analysis. *Marketing Science* 20(1):42–60.

Appendix A. Algorithm

Here, we succinctly summarize our method's key steps. The following algorithm defines our two stage estimator:

1. **Stage I**. Train a neural network to predict a product's market share s_{jt} given only its observed covariates x_{jt}. Use k-fold cross-fitting to obtain the predicted market share for each product, $\hat{\pi}_{jt}$
2. **Stage II**. Drop all products with zero sales. Then
   (a) Sort all products across markets into B bins based on their predicted shares, grouping products with similar $\hat{\pi}$. For each bin B_k, compute every product's weight as follows
   $$\hat{w}_{jt} = \frac{\sum_{b \in B_k, s_b > 0} \hat{\pi}_b}{\sum_{b \in B_k, s_b > 0} s_b}$$
   (b) Compute re-weighted market shares for each product
   $$\dot{s}_{jt} = \hat{w}_{jt} s_{jt}$$
   (c) Compute orthogonalization matrix $\hat{\mu}$ (see Appendix F)
   (d) Use existing BLP routines to estimate the causal parameters $\theta_0 = (\bar{\beta}, \Sigma_0)$
   $$\hat{\theta}_{TS} \equiv \arg\min_{\theta} \Big(\frac{1}{T} \sum_{t=1}^{T} \hat{\mu}\, z_t \xi_t(\theta, x_t, \dot{s}_t) \Big)' W \Big(\frac{1}{T} \sum_{t=1}^{T} \hat{\mu} z_t \xi_t(\theta, x_t, \dot{s}_t) \Big)$$

In practice, we incorporate $\hat{\mu}$ into step 2 (d) by redefining the weighting matrix used by the generalized methods of moments (GMM) estimator. Specifically, if W represents our original positive semi-definite moment weighting matrix, we define $\bar{W} = \hat{\mu}' W \hat{\mu}$. We can now use existing BLP implementations to solve for $\hat{\theta}_{TS}$. All we need to do is replace the observed shares s_{jt} with the weighted shares $\dot{s}_{jt}$ and the moment weighing matrix W with $\bar{W}$, and then use an existing BLP package to compute $\hat{\theta}_{TS}$. Examples of BLP packages include PyBLP in Python (Conlon and Gortmaker 2019) and the MATLAB implementation of the MPEC method (Dubé et al. 2012). Stage II thus requires minimal additional work compared to base BLP, and is easy to implement.

Appendix B. Proof of Lemma 2.1

Recall that for each product j in market t, $S_{jt}|x_t, \xi_t \sim \frac{1}{N} Binomial(N, \pi_{jt})$, where N is the total number of consumers in the market. We can use this fact to express the probability of a product going unsold,

$$\begin{aligned} \mathbb{P}(S_{jt} = 0 | z_t, \xi_t) &= \int (1 - \pi_{jt})^N dF_{x_t|z_t} \\ &= \int \left(1 - \int \frac{e^{x'_{jt}\bar{\beta} + x'_{jt}\Sigma_0 \nu + \xi_{jt}}}{1 + \sum_{k=1}^{J_t} e^{x'_{kt}\bar{\beta} + x'_{kt}\Sigma_0 \nu + \xi_{kt}}} dP_0(\nu) \right)^N dF_{x_t|z_t}, \end{aligned} \tag{A.13}$$

where $F_{x_t|z_t}$ is the cumulative density function of $x_t|z_t$. Notice that the inner integral term of the right-hand side of Equation A.13 is increasing in ξ_{jt} (the more utility product j has, the larger its market share), and so $\mathbb{P}(S_{jt}=0|z_t,\xi_t)$ is strictly decreasing in ξ_{jt}. This strict monotonicity is the key insight that yields the lemma. Since $\mathbb{P}(S_{jt}>0|z_t,\xi_t)=1-\mathbb{P}(S_{jt}=0|z_t,\xi_t)$, $\mathbb{P}(S_{jt}>0|z_t,\xi_t)$ is strictly increasing in ξ_{jt}.

Without loss of generality, we assume ξ_{jt} follows a discrete distribution. ξ_{jt}^+ denotes the ξ_{jt}'s which take on non-negative values, and ξ_{jt}^- denotes those with negative values. We can write

$$\begin{aligned}
\mathbb{E}[\xi_{jt}\,|\,z_t,S_{jt}>0] &= \sum_{\xi_{jt}^+}\xi_{jt}^+\mathbb{P}(\xi_{jt}^+|z_t,S_{jt}>0)+\sum_{\xi_{jt}^-}\xi_{jt}^-\mathbb{P}(\xi_{jt}^-|z_t,S_{jt}>0)\\
&=\sum_{\xi_{jt}^+}\xi_{jt}^+\mathbb{P}(\xi_{jt}^+|z_t)\mathbb{P}(S_{jt}>0|\xi_{jt}^+,z_t)\frac{1}{\mathbb{P}(S_{jt}>0|z_t)}\\
&+\sum_{\xi_{jt}^-}\xi_{jt}^-\mathbb{P}(\xi_{jt}^-|z_t)\mathbb{P}(S_{jt}>0|\xi_{jt}^-,z_t)\frac{1}{\mathbb{P}(S_{jt}>0|z_t)}\\
&=\frac{1}{\mathbb{P}(S_{jt}>0|z_t)}\left[\sum_{\xi_{jt}^+}\xi_{jt}^+\mathbb{P}(\xi_{jt}^+|z_t)\mathbb{P}(S_{jt}>0|\xi_{jt}^+,z_t)+\sum_{\xi_{jt}^-}\xi_{jt}^-\mathbb{P}(\xi_{jt}^-|z_t)\mathbb{P}(S_{jt}>0|\xi_{jt}^-,z_t)\right]\\
&>\frac{\mathbb{P}(S_{jt}>0|\xi_{jt}^+,z_t)}{\mathbb{P}(S_{jt}>0|z_t)}\left[\sum_{\xi_{jt}^+}\xi_{jt}^+\mathbb{P}(\xi_{jt}^+|z_t)+\sum_{\xi_{jt}^-}\xi_{jt}^-\mathbb{P}(\xi_{jt}^-|z_t)\right],
\end{aligned}$$

since $\mathbb{P}(S_{jt}>0|z_t,\xi_t)$ is strictly increasing in ξ_{jt}.

Since $\mathbb{E}[\xi_{jt}\,|\,z_t]=\sum_{\xi_{jt}^+}\xi_{jt}^+\mathbb{P}(\xi_{jt}^+|z_t)+\sum_{\xi_{jt}^-}\xi_{jt}^-\mathbb{P}(\xi_{jt}^-|z_t)=0$,

$$\mathbb{E}[\xi_{jt}\,|\,z_t,S_{jt}>0]>0.$$

Given the markets are independent, by the law of large numbers, $\mathbb{E}_T[\xi_{jt}(\theta_0,x_t,s_t)\,|\,z_t,s_{jt}>0]\to\mathbb{E}[\xi_{jt}|z_t,S_{jt}>0]>0$, as $T\to\infty$.

Appendix C. Hyperparameter Selection

There are two important hyperparameters in our estimation procedure: the number of individual predictors F in the first stage and the number of bins B in the second stage. Here, we briefly discuss how to choose values for these parameters in practice.

We first focus on the the number of individual predictors F. Recall that in the first stage of our method, we average predictions over F trained models to give our final estimates of market share. Overall, F should be chosen by cross-validation to maximise the out-of-sample accuracy of the first stage predictions. Setting $F>1$ allows the neural network to model heterogeneity in customer preferences; the larger the F, the more heterogeneity the model can capture. In theory, the larger

F	Out-of-sample R^2
2	0.179
10	0.642
50	0.772
100	0.775
150	0.769

Table A.14 Predictive accuracy of Stage I for different values of F, the number of individual predictors. (Simulation settings: $\beta_0 = -12$, $\sigma_\xi = 1.0$)

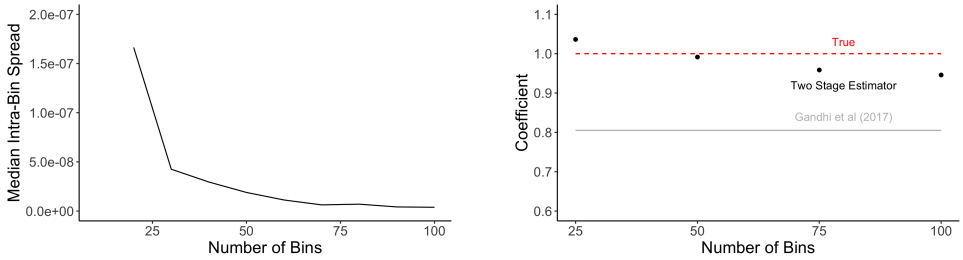
the F the better; however in practice, increasing F beyond a point does not improve predictive accuracy. Table A.14 summarizes the relationship between F and out-of-sample predictive accuracy in our simulation study. We clearly see that increasing F improves accuracy to a point, but above $F = 50$, all models perform equally well. As increasing F increases computational load, it is not worth increasing F beyond the point at which predictive accuracy plateaus. In our experience, setting F between 50 and 200 is a reasonable choice, though we recommend that practitioners choose the best value for their application by cross-validation.

The second important hyperparameter in our estimation procedure is the number of bins B. Recall that in Stage II of our method, we sort all positive sale products into B bins based on their predicted market share, then use these bins to construct weights that produce unbiased inference. Overall, we want these bins to be small enough that predicted share does not vary much within the bin, ensuring that products in the same bin derive similar amounts of utility from the observed features. However, we also want the bins to be large enough that $\frac{1}{|B_k|}\sum_{b\in B_k, s_b>0}\hat{\pi}_b$ and $\frac{1}{|B_k|}\sum_{b\in B_k, s_b>0} s_b$ are meaningful (i.e. low variance) estimates of $\mathbb{E}[\pi_{jt}|x_t]$ and $\mathbb{E}[\pi_{jt}|x_t, s_{jt} > 0]$ for each bin B_k.

One way of achieving this balance is by plotting the median intra-bin spread of predicted share. For a given value of B, we calculate the squared difference between the highest and lowest values of $\hat{\pi}$ in each bin, then compute the median of these values across all bins. We can plot these median values for a few different values of B, as shown in the left panel of Figure A.3. As we see, the median intra-bin spread is a decreasing function of the number of bins; however, the spread declines quickly, then flattens out beyond a point. The point at which the curve flattens out is a good candidate for B. At this point, we have achieved most of the benefit in terms of maximising the similarity of products within each bin; increasing the number of bins may only reduce the quality of our estimates. This approach is conceptually similar to the Elbow method often used for hyperparameter selection in clustering and unsupervised machine learning.

Figure A.3 uses our simulation experiment to demonstrate that this heuristic is a good approach for choosing the number of bins. On the left, we plot the spread curve for two different simulation

(a) Bin Selection with $\beta_0 = -10, \sigma_\xi = 1.5$



(b) Bin Selection with $\beta_0 = -14, \sigma_\xi = 0.5$

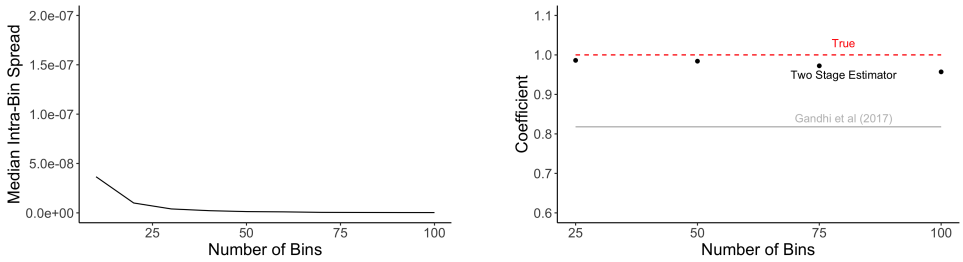


Figure A.3 Bin Selection. Setting B as the number of bins after which the spread curves (on the left) flatten yields the most accurate estimates of $\bar{\beta}^{(1)}$. However, in both simulations (a) and (b), our method is robust to the exact choice of B: any value produces less biased estimates than the bound method and BLP (not pictured: see Tables A.20 and A.24).

settings, while on the right, we plot the mean estimate of the customer preference coefficient $\bar{\beta}^{(1)}$ (see section 4.1.1 for a detailed discussion of the simulation setup). In the left figure of part (a), the spread curve flattens out around 50 bins; this setting also corresponds to the least biased estimate of $\bar{\beta}^{(1)}$, as shown in the figure on the right. Part (b) has a longer tail, but shows a similar story: the spread curve flattens around 25 bins, which produces the most accurate results. However, it is important to note that our estimates are robust to the selection of B. In both cases, no matter the chosen value of B, our estimates are far more accurate than that of the bound estimator (Gandhi et al. 2020) and BLP. Thus, any reasonable setting of B will yield high quality estimates of the structural parameters.

Appendix D. Comparison with Other Prediction Methods

In section 3.1, we propose a deep learning method to estimate each product's market share using only its observed features. It is not strictly necessary to use this approach: any statistical model that yields high quality predictions can be used to construct weights in the second stage. However, this prediction task is challenging for two reasons: the long tail of zero products and intra-market product dependence. In this section, we discuss these challenges, and demonstrate that our deep learning method greatly outperforms alternative approaches.

The first key challenge for prediction in this setting is the long tail of zero sale products. Recall that in the first stage, our goal is to accurately predict the market share of all products in the dataset. A natural choice for this prediction is a simple MNL choice model. However, MNL choice models cannot handle products that have zero sales, and thus may struggle to provide accurate predictions in the long tail setting. To train an MNL model, we either need to throw out all products with zero sales–which can lead to significant bias and loss of power–or use a workaround, such as adding one to all product sales to artificially remove the zeroes. As we will demonstrate, neither of these provides satisfactory results. The long tail also presents challenges for non-choice statistical models, as market shares in the long tail setting are highly right skewed. Many simple methods (e.g. linear regression) will struggle to deal with this skewness, even after using appropriate data transformations (e.g. a log transform). Our deep learning approach is not only able to include all products in its estimation, but also is unaffected by this skewness.

The second key challenge for prediction in this setting is intra-market share dependence. Each product's market share depends not only on its own features, but also on the features of every other product in the same market. An MNL choice model is the only simple method that is designed to handle such dependence explicitly. Generic machine learning methods (e.g. random forests) cannot easily account for this dependence. One possible workaround is to expand each product's covariate set to include the features of all other products in its market. However, this is an ineffective solution, as even a modest number of products per market greatly increases the dimensionality of the dataset. Moreover, this approach cannot easily handle different markets having different numbers of products, which is almost always the case in practice. Our deep learning based approach explicitly models the intra-market dependence without such cumbersome feature engineering.

We demonstrate the superiority of our proposed prediction method in a simulation experiment. Consider the baseline case discussed in section 4.1.2 and Table 2. We evaluate seven prediction strategies for Stage I:

- **Linear Regression (LR)**: A simple linear regression that uses a product's own features to predict its market share. This approach uses a log transform to deal with skewness. It ignores the dependence of a product's share on the features of other products in its market.
- **Expanded Linear Regression (ELR)**: This approach expands a product's covariate set to include the features of all other products in its market, and train a linear model as before. Note that this greatly increases the dimensionality of the dataset: with five features per product and 50 products per market, the covariate set now has 250 features (20% of the total number of observations). We thus add L2 regularization (i.e. the ridge penalty) (Hoerl and Kennard 1970) to improve predictive performance. It is worth noting that this is only possible as every market has the same number of products in this simulation. This pattern is highly atypical in practice.

- **Random Forest (RF)** A random forest (Breiman 2001) that predicts a product's market share using only its own features. This method is tree-based, and can effectively handle skewed data. It is an easy to use, nonparametric prediction technique.
- **Expanded Random Forest (ERF)**. As in the linear case, we also train the random forest on the expanded covariate set.
- **MNL** A multinomial logit choice model. This approach cannot handle zero share products, and is thus trained only on products with positive observed sales.
- **MNL-Add** An alternate multinomial logit choice model. This approach uses all products (i.e. with positive and zero share), but adds one to the quantity sold of each product to artificially remove zeroes.
- **Deep Learning (DL)** Our proposed deep learning method, described in section 3.1

We first evaluate the predictive performance of these methods using out-of-sample R^2 (Table A.15). Our deep learning approach clearly provides the most accurate predictions, followed by the simple MNL model. The random forest in particular seems to struggle. However, overall R^2 is not the best measure of performance in this setting. The predictions from the first stage of our method will be used to construct bins and weights in the second stage. For our method to work, the predicted share in each bin must be close to the true share. We use the predicted share from each method to construct $B = 50$ equally sized bins of products (as described in section 3.2), and compare the mean predicted share to the mean true share in each bin in Figure A.4. The closer the plotted points lie to the 45 degree line, the better the predictions. From the figure, it is clear that our deep learning method produces the most accurate predictions. Despite its low R^2, the random forest produces reasonably accurate predictions within each bin. However, it tends to overestimate low shares and underestimate high shares. Moreover, expanding the covariate set greatly worsens the predictions of the random forest; the high-dimensionality seems to outweigh the benefits of modeling intra-market dependence. The MNL model performs better than the RF for low share products, but greatly underestimates high shares. This inaccuracy is made much worse by the add-one trick in MNL-Add; it appears to be better to just drop the zero-share products than to use this workaround. Finally, both linear models–LR and ELR–perform poorly due to the right skew of the data.

Finally, we consider the impact of using other prediction methods on the estimation in Stage II. For clarity, we only present results for RF, MNL, and DL; as discussed, the other methods do not produce reasonable first stage predictions. As Table A.16 demonstrates, the deep learning based approach produces significantly more accurate inference for the structural parameters of interest. The skew observed in the RF and MNL predictions (Figure A.4) greatly impact their estimation accuracy. MNL is particularly inaccurate: this is likely due to its greater inaccuracy in predicting

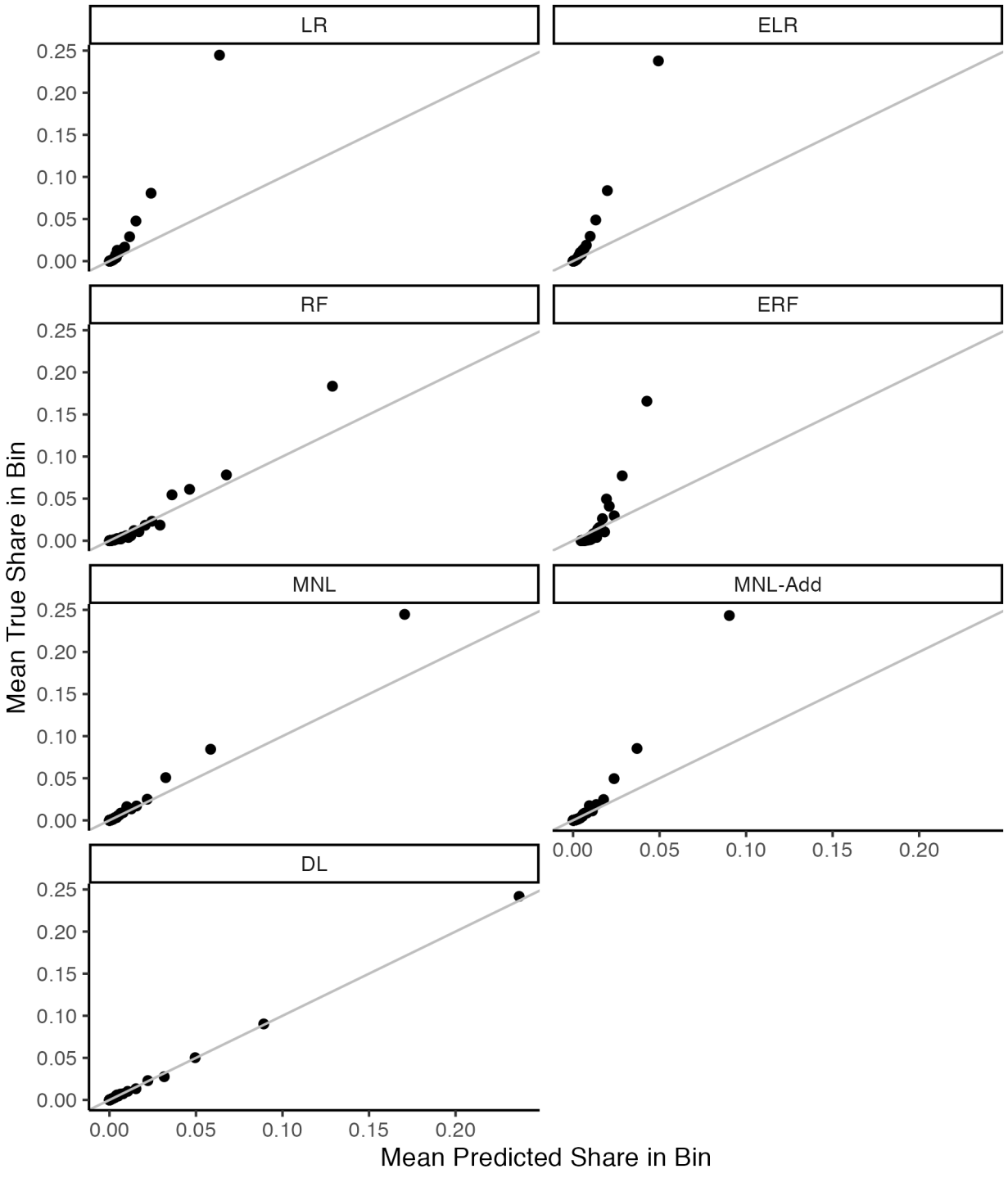


Figure A.4 **Mean predicted market share vs. true market share in each bin for the seven considered first stage methods. Points closer to the 45-degree line (grey) reflect more accurate predictions. We display results for the baseline simulation case ($\beta_0 = -10$, $\sigma_\xi = 0.5$).**

Prediction Method	Out-of-sample R^2
LR	0.85
ELR	0.85
RF	0.52
ERF	0.37
MNL	0.90
MNL-Add	0.88
DL	0.92

Table A.15 First stage out-of-sample R^2 for alternate prediction methods in baseline case ($\beta_0 = -10$, $\sigma_\xi = 0.5$)

Parameter	True Value	RF	MNL	DL
β_0	-10.000	-8.399	-9.537	-10.331
		(0.005)	(0.020)	(0.016)
$\bar{\beta}^{(1)}$	1.000	0.812	0.788	1.016
		(0.001)	(0.003)	(0.002)
$\bar{\beta}^{(2)}$	1.000	0.872	0.833	1.012
		(0.001)	(0.001)	(0.001)
$\lambda^{(1)}$	0.500	0.563	0.429	0.497
		(0.002)	(0.007)	(0.003)
$\lambda^{(2)}$	0.500	0.491	0.422	0.504
		(0.002)	(0.003)	(0.001)

Table A.16 Stage II estimates of structural parameters using three different prediction methods in stage 1: a random forest (RF), simple multinomial logit (MNL), and our proposed deep learning approach (DL). We display results for the baseline case, with 26% of products unsold ($\beta_0 = -10$, $\sigma_\xi = 0.5$)

high shares (compared to RF and DL). Overall, it is clear that our deep learning based approach offers a significant improvement over alternative first stage prediction methods.

Appendix E. Proof of the Weighting Scheme

Here, we prove our weighting scheme's key property. Specifically, we show that our two stage approach finds weights w_t such that

$$\mathbb{E}_T[\xi_{jt}(\theta_0, x_t, w_t s_t) \,|\, s_{jt} > 0, z_t] \to 0 \tag{A.14}$$

First, consider the functional form of ξ_{jt} in the unweighted case. Consider a single consumer with preferences β_i. In this case, ξ has the following functional form

$$\xi_{jt}(\beta_i, x_t, s_t) = \log s_{ijt} - \log s_{i0t} - x'_{jt}\beta_i \tag{A.15}$$

where s_{ijt} is the market share of product (j,t) for consumer i, and s_{i0t} is the probability that i does purchase a product. Note that $s_{i0t} = 1 - \sum_{j=1}^{J_t} s_{ijt}$. Thus, for the given value of β_i, we have that

$$\mathbb{E}_T[\xi_{jt}(\beta_i, x_t, s_t) | s_{jt} > 0, z_t] = \mathbb{E}_T[\log s_{ijt} - \log s_{i0t} - x'_{jt}\beta_i | s_{jt} > 0, z_t] \tag{A.16}$$

Our proposed weights are defined as

$$w_{jt} = \frac{\mathbb{E}[S_{jt}|x_t]}{\mathbb{E}[S_{jt}|x_t, s_{jt} > 0]} \tag{A.17}$$

Recall that S_{jt} is a multinomial random variable drawn from the true choice probabilities π_{jt}, and that the expectation reflected here is over ξ (i.e. conditioning only on x_t).

First, consider the model with exogenous covariates and no random coefficients, and let the true customer preference be given by β.

$$\begin{aligned}
&\mathbb{E}_T[\log w_{jt}s_{jt} - \log w_{0t}s_{0t} - x'_{jt}\beta | s_{jt} > 0, x_t] \\
&= \mathbb{E}_T[\log \mathbb{E}[S_{jt}|x_t] - \log \mathbb{E}[S_{jt}|x_t, s_{jt} > 0] + \log s_{jt} - \log s_{0t} - x'_{jt}\beta | s_{jt} > 0, x_t] \\
&= \mathbb{E}_T[\log \mathbb{E}[S_{jt}|x_t] - \log \mathbb{E}[S_{jt}|x_t, s_{jt} > 0] + \tilde{\xi}_{jt} | s_{jt} > 0, x_t],
\end{aligned}$$

where $\tilde{\xi}_{jt} = \xi_{jt}(\beta, x_t, s_t)$, the setting of the unobservable that satisfies the share equation with the true parameter β for the observed share s_t. Note that these are not equal the true values of the unobservable ξ_{jt}, but converge to it in large samples, as $\xi_{jt} = \xi_{jt}(\beta, x_t, \pi_t)$. Since $\log \mathbb{E}[S_{jt}|x_t] = \log \pi_{jt} + \frac{1}{\mathbb{E}[S_{jt}|x_t]}(\pi_{jt} - \mathbb{E}[S_{jt}|x_t]) + o(\pi_{jt} - \mathbb{E}[S_{jt}|x_t])$,

$$\begin{aligned}
&\mathbb{E}_T[\log \mathbb{E}[S_{jt}|x_t] - \log \mathbb{E}[S_{jt}|x_t, s_{jt} > 0] + \tilde{\xi}_{jt} | s_{jt} > 0, x_t] \\
&= \mathbb{E}_T \Big\{ \log \pi_{jt} - \log \pi_{jt} | s_{jt} > 0 + \tilde{\xi}_{jt} \\
&+ O(\pi_{jt} - \mathbb{E}[S_{jt}|x_t]) + o(\pi_{jt} - \mathbb{E}[S_{jt}|x_t]) \\
&+ O(\pi_{jt}|s_{jt} > 0 - \mathbb{E}[S_{jt}|x_t, s_{jt} > 0]) + o(\pi_{jt}|s_{jt} > 0 - \mathbb{E}[S_{jt}|x_t, s_{jt} > 0]) | s_{jt} > 0, x_t \}
\end{aligned}$$

We can rewrite the first three terms as follows,

$$\begin{aligned}
&= \log \frac{exp(x'_{jt}\beta + \xi_{jt})}{1 + \sum_k exp(x'_{kt}\beta + \xi_{kt})} - \log \frac{exp(x'_{jt}\beta + \xi_{jt})}{1 + \sum_k exp(x'_{kt}\beta + \xi_{kt})} | s_{jt} > 0 + \tilde{\xi}_{jt} \\
&= \xi_{jt} - \xi_{jt} | s_{jt} > 0 + \tilde{\xi}_{jt}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_T \Big\{ \xi_{jt} - \xi_{jt} | s_{jt} > 0 + \tilde{\xi}_{jt} \\
&+ O(\pi_{jt} - \mathbb{E}[\pi_{jt}|x_t]) + o(\pi_{jt} - \mathbb{E}[\pi_{jt}|x_t]) \\
&+ O(\pi_{jt}|s_{jt} > 0 - \mathbb{E}[\pi_{jt}|x_t, s_{jt} > 0]) + o(\pi_{jt}|s_{jt} > 0 - \mathbb{E}[\pi_{jt}|x_t, s_{jt} > 0]) | s_{jt} > 0, x_t \} \\
&\to 0.
\end{aligned}$$

Next, we show that similar reasoning generalizes to the random coefficient case with instruments z_t. Let G_T be the moment condition $G_T(\beta, s_t) = \mathbb{E}_T(z_t' \xi_t(\beta, s_t))$. Then,

$$\begin{aligned}
&G_T(\beta, w_t s_t) - G_T(\beta, \pi_t) \\
&= \mathbb{E}_T \left\{ z_t' \xi_t(\beta, \frac{\mathbb{E}[S_t|x_t]}{\mathbb{E}[S_t|x_t, s_t > 0]} s_t) - z_t' \xi_t(\beta, \pi_t) \right\} \\
&= \mathbb{E}_T \left\{ z_t' H_{0t}^{-1} \left(\frac{\mathbb{E}[S_t|x_t]}{\mathbb{E}[S_t|x_t, s_t > 0]} s_t - \pi_t \right) - \frac{1}{2} z_t' I_{0t} \left(\frac{\mathbb{E}[S_t|x_t]}{\mathbb{E}[S_t|x_t, s_t > 0]} s_t - \pi_t \right)^2 + o_p(\frac{1}{N}) \right\},
\end{aligned}$$

where H_{0t}^{-1} and I_{0t} are constants as in Freyberger (2015). Since $\mathbb{E}_T[s_t] \to \mathbb{E}[S_t|x_t, s_t > 0]$ and $\mathbb{E}_T[\pi_t] \to \mathbb{E}[S_t|x_t]$ as $T \to \infty$, $G_T(\beta, w_t \tilde{s}_t) - G_T(\beta, s_t) \to 0$

We use the sample average of $\mathbb{E}_T[\pi_{jt}|x_{jt}]$ and $\mathbb{E}_T[s_{jt}|s_{jt} > 0, x_{jt}]$ for each matched sample to construct the estimator for w_t: $\hat{w}_b = \frac{\frac{1}{|B_k|} \sum_{b' \in B_k, s_{b'} > 0} \hat{\pi}_{b'}}{\frac{1}{|B_k|} \sum_{b' \in B_k, s_{b'} > 0} s_{b'}}$.

Appendix F. Orthogonalization

In Appendix E, we establish the conditions our weighting scheme needs to fulfil to yield unbiased inference. Our method estimates these weights w_0 using the neural network predictions from Stage I. Though these estimates $\hat{w}$ will be close to the true weights, they may suffer from biases due to regularization. To ensure that our Stage II results are $\sqrt{T}$ consistent, we must make our estimates less sensitive to biased estimates of w_0. In this paper, we accomplish this using the "double machine learning" approach proposed by Chernozhukov et al. (2018). This approach modifies the moment condition function to be Neyman orthogonal to the nuisance parameters, that is, parameters that are not of causal interest, but still effect the observed outcomes.

Specifically, we introduce the orthogonalization parameter $\mu \in \mathbb{R}^{d_\theta \times d_z}$ with true value μ_0. The nuisance set η consists both of the weights w and μ, that is, $\eta = (w, \mu)$ with true value $\eta_0 = (w_0, \mu_0)$. We use μ to modify our GMM moment score from $m(\theta, z_t, w_t s_t) = z_t \xi_t(\theta, x_t, w_t s_t)$ to $\psi(\theta, z_t, s_t, \eta) = \mu z_t \xi_t(\theta, x_t, w_t s_t)$ such that the following two conditions hold

$$\mathbb{E}\psi(\theta_0, z_t, w_t s_t, \eta_0) = 0 \tag{A.18}$$

$$\partial_\eta \mathbb{E}\psi(\theta_0, z_t, w_t s_t, \eta_0)[\eta - \eta_0] = 0 \tag{A.19}$$

The first condition ensures the moment score function ψ still identifies θ_0, while the second condition posits that the Gateaux derivative of ψ with respect to η vanishes at the true parameter values. As the nuisance set contains both w and μ, this vanishing derivative ensures that our estimation is robust to small biases not only in estimates of w_0, but also μ_0

Constructing an appropriate μ follows from Chernozhukov et al. (2018), and relies mostly on values we already use in the BLP procedure. The true value μ_0 is given by

$$\mu_0 = A' \Omega^{-1} - A' \Omega^{-1} G_w (G_w' \Omega G_w)^{-1} G_w' \Omega^{-1} \tag{A.20}$$

where $\Omega \in \mathbb{R}^{d_z \times d_z}$ is an arbitrary positive definite weighting matrix, $G_w \in \mathbb{R}^{d_z \times d_\theta}$ is given by

$$G_w = \partial_w \mathbb{E}[m(\theta, z_t, w_t s_t)]\big|_{\theta=\theta_0, w=w_0}, \tag{A.21}$$

and $A \in \mathbb{R}^{d_z \times d_\theta}$ is an arbitrary moment selection matrix who's optimal value is

$$A = \partial_\theta \mathbb{E}[m(\theta, z_t, w_t s_t)]\big|_{\theta=\theta_0, w=w_0}, \tag{A.22}$$

The functional form of BLP allows us to estimate each of these quantities individually and μ_0 a a whole. Assume we have $\hat{\theta}_{init}$, an initial estimate of θ_0. $\hat{\theta}_{init}$ can derived from prior knowledge, running standard BLP, or by running our two stage estimator without orthogonalization. Recall that in BLP estimation, we often define $\delta_{jt} = \xi_{jt} + x'_{jt}\bar{\beta}$, and express the model-estimated share σ (from equation 3) as a function of δ_t and Σ,

$$\sigma_{jt}(\delta_t, \Sigma, x_t) = \int \frac{e^{\delta_{jt} + x'_{jt}\Sigma\nu}}{1 + \sum_{k=1}^{J_t} e^{\delta_{kt} + x'_{kt}\Sigma\nu}} dP_0(\nu) \tag{A.23}$$

We then optimize over free parameters δ and θ, subject to the constraint that $\sigma_{jt}(\delta_t, \Sigma, x_t) = w_{jt}s_{jt}$. This form allow us to estimate the orthogonalization parameter. We first present our estimates, then explain how we derive them. We estimate μ_0 with $\hat{\mu}$ such that

$$\hat{\mu} = \hat{A}'\Omega^{-1} - \hat{A}'\Omega^{-1}\hat{G}_w(\hat{G}'_w\Omega\hat{G}_w)^{-1}\hat{G}'_w\Omega^{-1} \tag{A.24}$$

where Ω is an arbitrary $d_z \times d_z$ matrix (e.g. the identity),

$$\hat{G}_w = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t} z_{jt}s_{jt}\Big(\frac{\partial \sigma_{jt}}{\partial \delta_{jt}}\Big)^{-1}\Big|_{\theta=\hat{\theta}_{init}} \tag{A.25}$$

and

$$\hat{A} = \left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t} z_{jt}x_{jt}\,,\ \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t} z_{jt}\Big(\frac{\partial \sigma_{jt}}{\partial \Sigma}\Big)\Big(\frac{\partial \sigma_{jt}}{\partial \delta_{jt}}\Big)^{-1}\right]\Bigg|_{\theta=\hat{\theta}_{init}} \tag{A.26}$$

We can then incorporate this estimate $\hat{\mu}$ into our moment score function as described in section 3.2 and Appendix A. Note that $\frac{\partial \sigma_{jt}}{\partial \delta_{jt}}$ and $\frac{\partial \sigma_{jt}}{\partial \Sigma}$ usually need to be computed to solve the GMM optimization problem in BLP estimation (see Dubé et al. (2012)). As a result, orthogonalizing our two stage estimator involves minimal computational overhead as it primarily uses values we already have access to. Also note that we compute $\hat{\mu}$ once before estimation, not in every iteration of the optimizer.

The reparametrization of ξ_{jt} makes deriving $\hat{G}_w$ and $\hat{A}$ straightforward. First, we estimate $\mathbb{E}[m(\theta, x_t, w_t s_t)]$ with its sample equivalent

$$\hat{m} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t} z_{jt}\xi_{jt}(\theta, x_t, w_t s_t) \tag{A.27}$$

which can be written as

$$\hat{m} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t} z_{jt}(\delta_{jt}(\Sigma, x_t, w_t s_t) - x'_{jt}\beta) \tag{A.28}$$

We can then estimate G_w with

$$\hat{G}_w = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t} z_{jt}\frac{\partial}{\partial w}\delta_{jt}(\Sigma, x_t, w_t s_t) \tag{A.29}$$

To calculate the derivative in this expression, we use the fact that $\sigma_{jt}(\delta_t, \Sigma, x_t) = w_{jt}s_{jt}$, which in turn implies that $\partial_w \sigma_{jt} = s_{jt}$ By the chain rule, we know that

$$\frac{\partial \sigma_{jt}}{\partial w} = \frac{\partial \sigma_{jt}}{\partial \delta_{jt}}\frac{\partial \delta_{jt}}{\partial w} \tag{A.30}$$

and can thus write

$$\hat{G}_w = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t} z_{jt}s_{jt}\Big(\frac{\partial \sigma_{jt}}{\partial \delta_{jt}}\Big)^{-1} \tag{A.31}$$

We derive $\hat{A}$ using similar reasoning. Though we can use an arbitrary matrix to estimate A, its optimal value is estimated by

$$A = \Big[\partial_\beta \hat{m}(\theta, x_t, w_t s_t)\ ,\ \partial_\Sigma \hat{m}(\theta, x_t, w_t s_t)\Big]\Big|_{\theta=\theta_0, w=w_0}, \tag{A.32}$$

Reparametrizing $\xi_{jt} = \delta_{jt} - x'_{jt}\bar{\beta}$ and using similar chain rule arguments as before, we obtain

$$\hat{A} = \left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t} z_{jt}x_{jt}\ ,\ \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J_t}\sum_{j=1}^{J_t} z_{jt}\Big(\frac{\partial \sigma_{jt}}{\partial \Sigma}\Big)\Big(\frac{\partial \sigma_{jt}}{\partial \delta_{jt}}\Big)^{-1}\right] \tag{A.33}$$

Appendix G. Asymptotic Results

Here, we present two keys results about the asymptotic properties of our estimator. Overall, our estimate is $\sqrt{T}$ consistent and asymptotically normal as the number of markets T goes to infinity. To prove each of these properties, we rely heavily on the theory provided in Freyberger (2015) and Chernozhukov et al. (2018). In this section, we present key assumptions and results, as well as sketch out the proofs of the estimator's large sample properties.

We assume all the standard conditions for the BLP model hold true. These are presented and discussed at length in section 2.1 of Freyberger (2015). We then assert the following assumptions

Assumption 1. *Let $\hat{\eta} = (\hat{w}, \hat{\mu})$ denote our estimate of the true nuisance parameter $\eta_0 = (w_0, \mu_0)$, where w denotes the product weights and μ denotes the orthogonalization parameter. $\psi(\theta, \eta)$ is the re-weighted, orthogonalized moment score in our estimation, and T is the number of markets. We assume the following conditions hold*

A1.1 The matrix $\mu'\mu$ has full rank and is stochastically bounded, that is, for every $\epsilon > 0$, there exists an $M(\epsilon)$ such that $\mathbb{P}(||\mu'\mu|| > M(\epsilon)) < \epsilon$

A1.2 $\psi(\theta, \eta)$ is smooth. For all values of η, $r \in [0, 1)$, and some positive, finite constants C_0 and ω

(a) $\mathbb{E}[||\psi(\theta_0, \eta) - \psi(\theta_0, \eta_0)||^2] \leq C_0||\eta - \eta_0||^\omega$

(b) $\partial_r^2 \mathbb{E}[\psi(\theta_0, \eta_0 + r(\eta_0 - \eta)) \leq C_0||\eta - \eta_0||^2$

A1.3 The observed shares converge to their true values, that is

$$\max_{1 \leq t \leq T} \max_{1 \leq j \leq J_t} |s_{jt} - \pi_{jt}| \xrightarrow{p} 0$$

A1.4 $\hat{\eta}$ converges to η_0 sufficiently quickly. More specifically,

$$||\hat{\eta} - \eta_0|| = o_p(T^{-1/4})$$

A1.5 The moment score $\psi(\theta, \eta)$ obeys the Neyman near-orthogonality condition at $\eta = \hat{\eta}$

$$||\partial_\eta \psi(\theta_0, \eta_0)[\eta_0 - \eta]|| \leq \delta_T T^{-1/2}$$

where $(\delta_T)_{T \geq 1}$ is a series of positive constants converging to zero.

Assumption A1.1 is a standard regularity condition for the orthogonalization parameter μ. A1.2 is a standard smoothness condition that is typically fulfilled in practice. A1.3 and A1.4 relate to the quality of our estimates of the true nuisance parameters w_0 and μ_0. A1.3 requires that as the number of markets goes to infinity, our estimated weights $\hat{w}$ approach the weights w_0. This condition is weak and easily satisfied by our neural network-based estimates. A1.4 further requires our nuisance parameters to converge to their true values at a $T^{-1/4}$ rate. This condition is guaranteed by a variety of machine learning algorithms (Chernozhukov et al. 2018). Finally, A1.5 states that the double machine learning approach is effective, and our moment function is Neyman orthogonal to the nuisance parameters. This condition is guaranteed by construction of ψ, assumption A1.4, and the theory presented in Chernozhukov et al. (2018).

Having stated our key assumptions, we now present our main results in Theorems 1 and 2 below. We first present the two theorems, then sketch out their proofs later in the section.

THEOREM 1. *If the conditions in Assumption 1 hold, then $\hat{\theta} \xrightarrow{p} \theta$ as $T \to \infty$.*

THEOREM 2. *Assume the conditions in Assumption 1 hold. If R denotes the number of random draws used to estimate the shares in Equation (3), $\lambda_1 = lim_{T \to \infty} \frac{\sqrt{T}}{R} < \infty$, Φ_1 is a $d_z \times d_z$ matrix, and $\bar{m}$ is a finite constant, then*

$$\sqrt{T}(\hat{\theta}_{TS} - \theta_0) \xrightarrow{d} \mathcal{N}((\Gamma' W \Gamma)^{-1} \Gamma' W \, \lambda_1 \bar{m}, V_{GMM})$$

where

$$V_{GMM} = (\Gamma' W \Gamma)^{-1} \Gamma' W \Phi_1 W \Gamma (\Gamma' W \Gamma)^{-1}, \quad \textit{and}$$
$$\Gamma = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\partial}{\partial \theta} \mathbb{E}[\mu_0 \, z_t' \xi_t(\theta_0, x_t, ws_t)]$$

Overall, these theorems establish that our estimator is $\sqrt{T}$ consistent and asympotically normal. As long as our Stage I estimates are predictive and the chosen number of draws R grows faster than $\sqrt{T}$, we obtain unbiased estimates of customer preferences even in the presence of the long tail. In the rest of this section, we sketch out how to prove these two theorems, relying heavily on the theory presented in Freyberger (2015) and Chernozhukov et al. (2018).

Proof of Theorem 1 The proof of this theorem follows as Theorem A.1 in Freyberger (2015). Define $\psi_T(\theta, w_t s_t, \mu) = \mathbb{E}_T[\psi(\theta, w_t s_t, \mu)]$. The first part of the proof shows that an estimator defined as a series that satisfies $||\psi_T(\tilde{\theta}, w_t s_t, \mu)|| = \inf_{\theta} ||\psi_T(\theta, w_t s_t, \mu)|| + o_p(1)$ is consistent. As this condition is completely general, it also holds for our moment score.

The second part of the proof establishes that BLP satisfies the condition mentioned above. To do so, it relies on a set of standard regularity assumptions that hold true for BLP. To modify the proof to fit our estimator, we need to further assume conditions A1.1 and A1.3. Once we make these assumptions, the proof follows as laid out by Freyberger (2015), and is thus not repeated here.

Proof of Theorem 2 This proof follows that of Theorem A.2 in Freyberger (2015). Specifically, if δ_T is a sequence of positive constants converging to zero such that $\delta_T \geq T^{-1/2}$, we can show that

$$\begin{aligned}\sqrt{T}(\hat{\theta}_{TS} - \theta_0) =& ((\Gamma' W \Gamma)^{-1} \Gamma' W + o_p(1)) \\ &\times \left(Q_{1T} + \frac{1}{\sqrt{R}} Q_{2T} + \frac{1}{\sqrt{T}} Q_{3T} + \frac{\sqrt{T}}{R} C_{1T} + \right. \\ &\left. o_p\Big(\frac{\sqrt{T}}{R}\Big) + O_p(\delta_T) \right) \end{aligned} \tag{A.34}$$

where $Q_{1T} \xrightarrow{d} N(0, \Phi_1)$, $Q_{2T} \xrightarrow{d} N(0, \Phi_2)$, $Q_{3T} \xrightarrow{d} N(0, \Phi_3)$, and $C_{1T} \xrightarrow{p} \bar{m}$. Furthermore, Q_{1T}, Q_{2T}, and Q_{3T} are asymptotically independent. This result is sufficient to prove the theorem, and is what we will focus on.

For this proof, we edit our notation slightly. The sample moment conditions are evaluated using $P_R(\nu)$, a distribution that approximates $P_0(\nu)$–the true distribution of customer heterogeneity– using R Monte Carlo samples. Our moment conditions are thus more adequately expressed as $\psi_T(\theta, \hat{w}_t s_t, \hat{\mu}, P_R)$. This approximation adds some error to the final estimate, which disappears as long as R grows faster than $\sqrt{T}$.

Recall that overall, our estimator's objective is to minimize

$$\psi_T(\theta, \hat{w}_t s_t, \hat{\mu}, P_R)' W \psi_T(\theta, \hat{w}_t s_t, \hat{\mu}, P_R)) \tag{A.35}$$

Let $G_T(\theta, \hat{w}_t s_t, P_R) = \mathbb{E}_T[m(\theta, \hat{w}_t s_t, P_R)]$ represent the unorthogonalized sample moment score. Then define

$$D\psi_T(\theta, \hat{w}_t s_t, \hat{\mu}, P_R) \equiv \frac{\partial}{\partial \theta} \psi_T(\theta, \hat{w}_t s_t, \hat{\mu}, P_R) \tag{A.36}$$

$$DG_T(\theta, \hat{w}_t s_t, P_R) \equiv \frac{\partial}{\partial \theta} G_T(\theta, \hat{w}_t s_t, P_R) \tag{A.37}$$

As Freyberger (2015) describes, we can write

$$\sqrt{T}(\hat{\theta}_{TS} - \theta_0) = (D\psi_T(\hat{\theta}_{TS}, \hat{w}_t s_t, \hat{\mu}, P_R)' W D\psi_T(D\psi_T(\hat{\theta}_{TS}, \hat{w}_t s_t, \hat{\mu}, P_R))^{-1} \tag{A.38}$$

$$\times D\psi_T(D\psi_T(\hat{\theta}_{TS}, \hat{w}_t s_t, \hat{\mu}, P_R) W \sqrt{T} \psi_T(D\psi_T(\theta_0, \hat{w}_t s_t, \hat{\mu}, P_R) \tag{A.39}$$

We first focus on the last term. Specifically, we show that

$$\begin{aligned} \sqrt{T}\psi_T(\theta_0, \hat{w}_t s_t, \hat{\mu}, P_R) &= \sqrt{T}\hat{\mu} G_T(\theta_0, \hat{w}_t s_t, P_R) \\ &= \sqrt{T}\hat{\mu} G_T(\theta_0, w_{0t} s_t, P_0) + O_p\Big(\frac{1}{\sqrt{R}}\Big) + O_p\Big(\frac{\sqrt{T}}{R}\Big) + \\ &\quad o_p\Big(\frac{\sqrt{T}}{R}\Big) + O_p(\delta_T) \end{aligned} \tag{A.40}$$

where w_{0t} represents the true weights that yield unbiased inference. To show A.40, we rewrite the left hand side as

$$\sqrt{T}\psi_T(\theta_0, \hat{w}_t s_t, \hat{\mu}, P_R) = \sqrt{T}\psi_T(\theta_0, w_{0t} s_t, \mu_0, P_0) + \tag{A.41}$$

$$\sqrt{T}[\psi_T(\theta_0, w_{0t} s_t, \mu_0, P_R) - \psi_T(\theta_0, w_{0t} s_t, \mu_0, P_0)] + \tag{A.42}$$

$$\sqrt{T}[\psi_T(\theta_0, \hat{w}_t s_t, \hat{\mu}, P_R) - \psi_T(\theta_0, w_{0t} s_t, \mu_0, P_R)] \tag{A.43}$$

As Freyberger (2015) describes, the first term (A.41) is $O_p(1)$ as it belongs to the GMM problem without simulation or estimation error. With Assumption A1.1, standard BLP regularity conditions, and Lyapunov's CLT, it converges to a normal variable.

Now consider the second term (A.42). We know that

$$\begin{aligned} \sqrt{T}[\psi_T(\theta_0, w_{0t} s_t, \mu_0, P_R) - \psi_T(\theta_0, w_{0t} s_t, \mu_0, P_0)] = \sqrt{T}\mu_0 [G_T(\theta_0, w_{0t} s_t, P_R) - \\ G_T(\theta_0, w_{0t} s_t, P_0)] \end{aligned} \tag{A.44}$$

Freyberger (2015) establishes that

$$\sqrt{T}[G_T(\theta_0, w_{0t} s_t, P_R) - G_T(\theta_0, w_{0t} s_t, P_0)] = O_p\Big(\frac{1}{\sqrt{R}}\Big) + O_p\Big(\frac{\sqrt{T}}{R}\Big) + o_p\Big(\frac{\sqrt{T}}{R}\Big) \tag{A.45}$$

Since μ_0 is stochastically bounded by assumption A1.1, we obtain

$$\sqrt{T}[\psi_T(\theta_0, w_{0t}s_t, \mu_0, P_R) - \psi_T(\theta_0, w_{0t}s_t, \mu_0, P_0)] = O_p\Big(\frac{1}{\sqrt{R}}\Big) + O_p\Big(\frac{\sqrt{T}}{R}\Big) + o_p\Big(\frac{\sqrt{T}}{R}\Big) \tag{A.46}$$

Finally, focus on the last term (A.43). Recall that $\eta = (\mu, w)$, and so we can rewrite this expression as

$$\sqrt{T}[\psi_T(\theta_0, \hat{\eta}, P_R) - \psi_T(\theta_0, \eta_0, P_R)] \tag{A.47}$$

This is exactly equivalent to equation (A.15) in Chernozhukov et al. (2018). As they lay out, as long as $\hat{\eta}$ converges to η_0 at a rate that is $o_p(T^{-1/4})$ (assumption A1.4) and the score function is smooth (assumption A1.2) and orthogonal to the nuisance parameters (assumption A1.5), then

$$\sqrt{T}[\psi_T(\theta_0, \hat{\eta}, P_R) - \psi_T(\theta_0, \eta_0, P_R)] = \delta_T + o_p(T^{-1/4}) + o_p(1) \lesssim \delta_T \tag{A.48}$$

Combining equations (A.46) and (A.48), we obtain the result described in equation (A.40). Now that we have established this property, the only thing left to do to prove the theorem is to show that $D\psi_T$ converges to Γ is probability. This follows from the reasoning in Freyberger (2015), and requires the boundedness of μ (Assumption A1.1) and the weighted share convergence (Assumption A1.3) as additional assumptions. As this proof is very similar to that provided in Freyberger (2015), it is not repeated here.

Finally, we also note that the estimation of the asymptotic variance follows the standard GMM variance estimation procedure.

Appendix H. Full Simulation Results

Here, we present the full results from our Monte Carlo simulations. The simulation set up was discussed at length in section 4. We first consider the case in which all product features are exogenous. We vary the true intercept β_0 between -10, -12, and -14 and the true standard deviation of the unobservable σ_ξ between 0.5, 1.0, and 1.5. In addition to analyzing every combination of these variables (Tables A.18 - A.26), we also consider an extremely long tail scenario with $\beta_0 = -20$ (Table A.27). We then allow for confounding, and consider a set of simulations in which the unobserved feature and the vertical product feature are endogenous. We vary β_0, σ_ξ, and the strength of confounding $\sigma_{e\xi}$ (Tables A.28 - A.34). In every setting, our two stage method produces significantly less biased estimates of β_0, $\bar{\beta}$, and λ than both BLP and the bound estimator proposed by Gandhi et al. (2020). The following tables demonstrate that our method not only outperforms the other methods, but is also more robust to increasing the size of the tail and the influence of the unobservable. The bias in the estimates of $\bar{\beta}^{(1)}$–the mean coefficient of the vertical feature–across all simulations is summarized in Table A.17.

Popularity of Outside Option (β_0)	Strength of Unobservable (σ_ξ)	Level of Endogeneity ($\sigma_{e\xi}$)	% Products Unsold	% Bias in Estimate of $\bar{\beta}^{(1)}$ Two Stage Estimator	BLP	Gandhi et al. (2020)
-10	0.5	0	26%	2%	-17%	-11%
-10	1	0	30%	1%	-22%	-14%
-10	1.5	0	33%	-1%	-28%	-20%
-12	0.5	0	42%	0%	-47%	-14%
-12	1	0	43%	0%	-51%	-16%
-12	1.5	0	45%	-3%	-55%	-20%
-14	0.5	0	59%	-2%	-88%	-18%
-14	1	0	59%	-1%	-87%	-18%
-14	1.5	0	61%	-3%	-86%	-21%
-20	0.5	0	94%	37%	-65%	-58%
-10	0.5	0.25	28%	1%	-17%	-8%
-14	0.5	0.25	59%	0%	-67%	-15%
-10	1.5	0.25	33%	-3%	-27%	-17%
-10	0.5	0.5	29%	1%	-17%	-4%
-14	0.5	0.5	60%	-1%	-64%	-10%
-10	1.5	0.5	34%	-2%	-27%	-14%
-14	1.5	0.5	60%	-8%	-68%	-15%

Table A.17 Summary of simulation results. Percentage of bias in the estimate $\bar{\beta}^{(1)}$ for our two stage method compared with BLP and Gandhi et al. (2020)

Table A.18 Baseline case with a moderate outside option and a moderately influential unobservable ($\beta_0 = -10$, $\sigma_\xi = 0.5$, 26% of products unsold)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-10.000	-10.331	-9.284	-9.296
		(0.016)	(0.004)	(0.003)
$\bar{\beta}^{(1)}$	1.000	1.016	0.830	0.890
		(0.002)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	1.012	0.898	0.886
		(0.001)	(0.001)	(0.001)
$\bar{\beta}^{(3)}$	1.000	1.011	0.897	0.888
		(0.001)	(0.001)	(0.001)
$\bar{\beta}^{(4)}$	1.000	1.018	0.898	0.886
		(0.001)	(0.001)	(0.001)
$\bar{\beta}^{(5)}$	1.000	1.016	0.898	0.887
		(0.001)	(0.001)	(0.001)
$\lambda^{(1)}$	0.500	0.497	0.664	0.450
		(0.003)	(0.002)	(0.001)
$\lambda^{(2)}$	0.500	0.504	0.496	0.449
		(0.001)	(0.002)	(0.001)
$\lambda^{(3)}$	0.500	0.503	0.496	0.450
		(0.001)	(0.002)	(0.001)
$\lambda^{(4)}$	0.500	0.490	0.493	0.453
		(0.001)	(0.002)	(0.001)
$\lambda^{(5)}$	0.500	0.499	0.494	0.452
		(0.001)	(0.002)	(0.001)

Table A.19 Moderate outside option with an influential unobservable ($\beta_0 = -10$, $\sigma_\xi = 1.0$, 30% of products unsold)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−10.000	−10.734	−8.926	−9.045
		(0.032)	(0.005)	(0.004)
$\bar{\beta}^{(1)}$	1.000	1.011	0.778	0.857
		(0.005)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	1.029	0.843	0.841
		(0.003)	(0.002)	(0.001)
$\bar{\beta}^{(3)}$	1.000	1.028	0.845	0.842
		(0.003)	(0.002)	(0.001)
$\bar{\beta}^{(4)}$	1.000	1.036	0.842	0.840
		(0.003)	(0.002)	(0.001)
$\bar{\beta}^{(5)}$	1.000	1.035	0.846	0.845
		(0.003)	(0.002)	(0.001)
$\lambda^{(1)}$	0.500	0.517	0.662	0.404
		(0.005)	(0.003)	(0.001)
$\lambda^{(2)}$	0.500	0.508	0.477	0.409
		(0.002)	(0.003)	(0.002)
$\lambda^{(3)}$	0.500	0.514	0.485	0.407
		(0.002)	(0.003)	(0.003)
$\lambda^{(4)}$	0.500	0.493	0.477	0.411
		(0.003)	(0.003)	(0.003)
$\lambda^{(5)}$	0.500	0.503	0.482	0.419
		(0.002)	(0.003)	(0.002)

Table A.20 Moderate outside option with an extremely influential unobservable ($\beta_0 = -10$, $\sigma_\xi = 1.5$, 33% of products unsold)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−10.000	−11.170	−8.429	−8.697
		(0.050)	(0.007)	(0.005)
$\bar{\beta}^{(1)}$	1.000	0.992	0.718	0.805
		(0.007)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	1.052	0.770	0.781
		(0.006)	(0.002)	(0.002)
$\bar{\beta}^{(3)}$	1.000	1.049	0.775	0.781
		(0.006)	(0.002)	(0.002)
$\bar{\beta}^{(4)}$	1.000	1.062	0.769	0.780
		(0.006)	(0.002)	(0.002)
$\bar{\beta}^{(5)}$	1.000	1.066	0.774	0.786
		(0.006)	(0.002)	(0.002)
$\lambda^{(1)}$	0.500	0.571	0.637	0.345
		(0.006)	(0.004)	(0.002)
$\lambda^{(2)}$	0.500	0.535	0.456	0.348
		(0.004)	(0.004)	(0.004)
$\lambda^{(3)}$	0.500	0.539	0.464	0.343
		(0.004)	(0.004)	(0.004)
$\lambda^{(4)}$	0.500	0.502	0.451	0.347
		(0.004)	(0.004)	(0.004)
$\lambda^{(5)}$	0.500	0.526	0.460	0.359
		(0.004)	(0.004)	(0.004)

Table A.21 Strong outside option with a moderately influential unobservable ($\beta_0 = -12$, $\sigma_\xi = 0.5$, 42% of products unsold)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−12.000	−12.268	−10.545	−10.851
		(0.022)	(0.007)	(0.004)
$\bar{\beta}^{(1)}$	1.000	1.005	0.526	0.858
		(0.007)	(0.004)	(0.001)
$\bar{\beta}^{(2)}$	1.000	1.006	0.861	0.849
		(0.002)	(0.001)	(0.001)
$\bar{\beta}^{(3)}$	1.000	1.004	0.859	0.850
		(0.002)	(0.001)	(0.001)
$\bar{\beta}^{(4)}$	1.000	1.018	0.859	0.847
		(0.002)	(0.001)	(0.001)
$\bar{\beta}^{(5)}$	1.000	1.019	0.859	0.850
		(0.002)	(0.001)	(0.001)
$\lambda^{(1)}$	0.500	0.506	0.783	0.428
		(0.005)	(0.004)	(0.001)
$\lambda^{(2)}$	0.500	0.497	0.473	0.437
		(0.002)	(0.003)	(0.002)
$\lambda^{(3)}$	0.500	0.503	0.474	0.435
		(0.002)	(0.003)	(0.002)
$\lambda^{(4)}$	0.500	0.478	0.471	0.437
		(0.002)	(0.002)	(0.002)
$\lambda^{(5)}$	0.500	0.489	0.475	0.436
		(0.002)	(0.002)	(0.002)

Table A.22 Strong outside option with an influential unobservable
($\beta_0 = -12$, $\sigma_\xi = 1.0$, 43% of products unsold)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-12.000	-12.587	-10.118	-10.551
		(0.035)	(0.009)	(0.005)
$\bar{\beta}^{(1)}$	1.000	1.000	0.490	0.840
		(0.010)	(0.004)	(0.001)
$\bar{\beta}^{(2)}$	1.000	1.019	0.801	0.803
		(0.003)	(0.002)	(0.001)
$\bar{\beta}^{(3)}$	1.000	1.013	0.801	0.803
		(0.003)	(0.002)	(0.001)
$\bar{\beta}^{(4)}$	1.000	1.035	0.798	0.800
		(0.003)	(0.002)	(0.001)
$\bar{\beta}^{(5)}$	1.000	1.036	0.801	0.806
		(0.003)	(0.002)	(0.001)
$\lambda^{(1)}$	0.500	0.511	0.759	0.377
		(0.007)	(0.005)	(0.001)
$\lambda^{(2)}$	0.500	0.500	0.454	0.397
		(0.003)	(0.004)	(0.003)
$\lambda^{(3)}$	0.500	0.513	0.459	0.393
		(0.003)	(0.003)	(0.003)
$\lambda^{(4)}$	0.500	0.475	0.452	0.399
		(0.003)	(0.004)	(0.003)
$\lambda^{(5)}$	0.500	0.494	0.458	0.405
		(0.003)	(0.004)	(0.003)

Table A.23 Strong outside option with an extremely influential unobservable ($\beta_0 = -12$, $\sigma_\xi = 1.5$, 45% of products unsold)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−12.000	−12.953	−9.551	−10.141
		(0.051)	(0.010)	(0.006)
$\bar{\beta}^{(1)}$	1.000	0.968	0.455	0.802
		(0.013)	(0.004)	(0.002)
$\bar{\beta}^{(2)}$	1.000	1.043	0.725	0.742
		(0.005)	(0.002)	(0.002)
$\bar{\beta}^{(3)}$	1.000	1.034	0.726	0.744
		(0.005)	(0.002)	(0.002)
$\bar{\beta}^{(4)}$	1.000	1.059	0.722	0.741
		(0.006)	(0.002)	(0.002)
$\bar{\beta}^{(5)}$	1.000	1.068	0.726	0.747
		(0.006)	(0.002)	(0.002)
$\lambda^{(1)}$	0.500	0.557	0.718	0.317
		(0.009)	(0.006)	(0.002)
$\lambda^{(2)}$	0.500	0.522	0.430	0.338
		(0.004)	(0.005)	(0.004)
$\lambda^{(3)}$	0.500	0.535	0.434	0.335
		(0.004)	(0.005)	(0.004)
$\lambda^{(4)}$	0.500	0.484	0.427	0.339
		(0.004)	(0.005)	(0.004)
$\lambda^{(5)}$	0.500	0.509	0.431	0.356
		(0.004)	(0.005)	(0.004)

Table A.24 Very strong outside option with a moderately influential unobservable ($\beta_0 = -14$, $\sigma_\xi = 0.5$, 59% of products unsold)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−14.000	−14.300	−11.565	−12.136
		(0.026)	(0.014)	(0.006)
$\bar{\beta}^{(1)}$	1.000	0.984	0.123	0.818
		(0.012)	(0.010)	(0.002)
$\bar{\beta}^{(2)}$	1.000	1.012	0.827	0.792
		(0.002)	(0.002)	(0.001)
$\bar{\beta}^{(3)}$	1.000	1.010	0.825	0.792
		(0.002)	(0.002)	(0.001)
$\bar{\beta}^{(4)}$	1.000	1.046	0.828	0.792
		(0.002)	(0.002)	(0.001)
$\bar{\beta}^{(5)}$	1.000	1.046	0.827	0.793
		(0.002)	(0.002)	(0.001)
$\lambda^{(1)}$	0.500	0.518	0.848	0.398
		(0.007)	(0.006)	(0.001)
$\lambda^{(2)}$	0.500	0.493	0.474	0.418
		(0.002)	(0.004)	(0.002)
$\lambda^{(3)}$	0.500	0.500	0.470	0.417
		(0.002)	(0.004)	(0.002)
$\lambda^{(4)}$	0.500	0.459	0.473	0.419
		(0.002)	(0.003)	(0.002)
$\lambda^{(5)}$	0.500	0.474	0.470	0.418
		(0.002)	(0.004)	(0.002)

Table A.25 Very strong outside option with an influential unobservable ($\beta_0 = -14$, $\sigma_\xi = 1.0$, 59% of products unsold)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−14.000	−14.526	−11.049	−11.802
		(0.037)	(0.014)	(0.007)
$\bar{\beta}^{(1)}$	1.000	0.992	0.131	0.817
		(0.015)	(0.010)	(0.002)
$\bar{\beta}^{(2)}$	1.000	1.007	0.757	0.746
		(0.003)	(0.002)	(0.002)
$\bar{\beta}^{(3)}$	1.000	1.003	0.757	0.747
		(0.003)	(0.002)	(0.002)
$\bar{\beta}^{(4)}$	1.000	1.054	0.756	0.745
		(0.003)	(0.002)	(0.002)
$\bar{\beta}^{(5)}$	1.000	1.066	0.760	0.750
		(0.003)	(0.003)	(0.002)
$\lambda^{(1)}$	0.500	0.511	0.795	0.346
		(0.008)	(0.006)	(0.002)
$\lambda^{(2)}$	0.500	0.493	0.444	0.382
		(0.003)	(0.005)	(0.003)
$\lambda^{(3)}$	0.500	0.509	0.446	0.378
		(0.003)	(0.005)	(0.003)
$\lambda^{(4)}$	0.500	0.451	0.444	0.379
		(0.003)	(0.005)	(0.004)
$\lambda^{(5)}$	0.500	0.469	0.449	0.391
		(0.003)	(0.005)	(0.003)

Table A.26 Very strong outside option with an extremely influential unobservable ($\beta_0 = -14$, $\sigma_\xi = 1.5$, 61% of products unsold)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−14.000	−14.768	−10.416	−11.318
		(0.050)	(0.014)	(0.008)
$\bar{\beta}^{(1)}$	1.000	0.974	0.144	0.790
		(0.018)	(0.010)	(0.003)
$\bar{\beta}^{(2)}$	1.000	1.017	0.675	0.685
		(0.005)	(0.003)	(0.002)
$\bar{\beta}^{(3)}$	1.000	1.007	0.677	0.687
		(0.004)	(0.003)	(0.002)
$\bar{\beta}^{(4)}$	1.000	1.066	0.674	0.684
		(0.005)	(0.003)	(0.002)
$\bar{\beta}^{(5)}$	1.000	1.087	0.679	0.690
		(0.005)	(0.003)	(0.002)
$\lambda^{(1)}$	0.500	0.524	0.733	0.282
		(0.010)	(0.007)	(0.002)
$\lambda^{(2)}$	0.500	0.501	0.409	0.332
		(0.005)	(0.006)	(0.005)
$\lambda^{(3)}$	0.500	0.517	0.413	0.324
		(0.004)	(0.006)	(0.005)
$\lambda^{(4)}$	0.500	0.448	0.407	0.332
		(0.005)	(0.006)	(0.005)
$\lambda^{(5)}$	0.500	0.472	0.427	0.345
		(0.005)	(0.006)	(0.004)

Table A.27 Extremely Strong Outside Option, 94% of products unsold
($\beta_0 = -20$, $\sigma_\xi = 0.5$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-20.000	-21.420	-13.386	-13.159
		(0.107)	(0.039)	(0.034)
$\bar{\beta}^{(1)}$	1.000	1.368	0.346	0.420
		(0.022)	(0.014)	(0.007)
$\bar{\beta}^{(2)}$	1.000	0.951	0.475	0.488
		(0.009)	(0.005)	(0.004)
$\bar{\beta}^{(3)}$	1.000	0.923	0.467	0.484
		(0.011)	(0.005)	(0.004)
$\bar{\beta}^{(4)}$	1.000	1.158	0.476	0.483
		(0.011)	(0.005)	(0.004)
$\bar{\beta}^{(5)}$	1.000	1.287	0.481	0.481
		(0.011)	(0.005)	(0.004)
$\lambda^{(1)}$	0.500	0.342	0.348	0.345
		(0.008)	(0.006)	(0.002)
$\lambda^{(2)}$	0.500	0.411	0.352	0.309
		(0.008)	(0.008)	(0.006)
$\lambda^{(3)}$	0.500	0.465	0.376	0.310
		(0.008)	(0.007)	(0.006)
$\lambda^{(4)}$	0.500	0.386	0.360	0.319
		(0.009)	(0.007)	(0.005)
$\lambda^{(5)}$	0.500	0.327	0.354	0.317
		(0.008)	(0.007)	(0.006)

Table A.28 Baseline Endogenous Case, 28% of products unsold
($\beta_0 = -10$, $\sigma_\xi = 0.5$, $\sigma_{e\xi} = 0.25$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-10.000	-10.079	-9.179	-9.345
		(0.022)	(0.004)	(0.003)
$\bar{\beta}^{(1)}$	1.000	1.010	0.834	0.922
		(0.003)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	0.977	0.875	0.878
		(0.001)	(0.001)	(0.001)
$\bar{\beta}^{(3)}$	1.000	0.976	0.873	0.878
		(0.002)	(0.001)	(0.001)
$\bar{\beta}^{(4)}$	1.000	0.984	0.875	0.878
		(0.002)	(0.001)	(0.001)
$\bar{\beta}^{(5)}$	1.000	0.980	0.874	0.877
		(0.002)	(0.001)	(0.001)
$\lambda^{(1)}$	0.500	0.462	0.643	0.453
		(0.005)	(0.003)	(0.001)
$\lambda^{(2)}$	0.500	0.500	0.490	0.450
		(0.002)	(0.002)	(0.001)
$\lambda^{(3)}$	0.500	0.502	0.490	0.451
		(0.002)	(0.002)	(0.002)
$\lambda^{(4)}$	0.500	0.488	0.493	0.450
		(0.002)	(0.002)	(0.001)
$\lambda^{(5)}$	0.500	0.494	0.490	0.448
		(0.002)	(0.002)	(0.001)

Table A.29 Endogenous Case with Popular Outside Option, 59% of products unsold
($\beta_0 = -14$, $\sigma_\xi = 0.5$, $\sigma_{e\xi} = 0.25$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−14.000	−13.762	−11.511	−12.200
		(0.028)	(0.012)	(0.007)
$\bar{\beta}^{(1)}$	1.000	0.996	0.329	0.854
		(0.011)	(0.009)	(0.002)
$\bar{\beta}^{(2)}$	1.000	0.936	0.782	0.781
		(0.002)	(0.002)	(0.001)
$\bar{\beta}^{(3)}$	1.000	0.933	0.781	0.782
		(0.002)	(0.002)	(0.001)
$\bar{\beta}^{(4)}$	1.000	0.961	0.784	0.784
		(0.002)	(0.002)	(0.001)
$\bar{\beta}^{(5)}$	1.000	0.961	0.782	0.782
		(0.002)	(0.002)	(0.001)
$\lambda^{(1)}$	0.500	0.458	0.715	0.402
		(0.008)	(0.006)	(0.001)
$\lambda^{(2)}$	0.500	0.479	0.452	0.415
		(0.002)	(0.003)	(0.002)
$\lambda^{(3)}$	0.500	0.492	0.457	0.416
		(0.002)	(0.003)	(0.002)
$\lambda^{(4)}$	0.500	0.457	0.454	0.417
		(0.002)	(0.003)	(0.002)
$\lambda^{(5)}$	0.500	0.467	0.450	0.415
		(0.002)	(0.003)	(0.002)

Table A.30 Endogenous Case with Influential Unobservable, 33% of products unsold
($\beta_0 = -10$, $\sigma_\xi = 1.5$, $\sigma_{e\xi} = 0.25$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-10.000	-10.676	-8.369	-8.735
		(0.060)	(0.007)	(0.005)
$\bar{\beta}^{(1)}$	1.000	0.968	0.734	0.828
		(0.008)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	0.989	0.758	0.777
		(0.006)	(0.002)	(0.002)
$\bar{\beta}^{(3)}$	1.000	0.993	0.758	0.779
		(0.006)	(0.002)	(0.002)
$\bar{\beta}^{(4)}$	1.000	1.001	0.755	0.776
		(0.006)	(0.002)	(0.002)
$\bar{\beta}^{(5)}$	1.000	1.000	0.758	0.778
		(0.006)	(0.002)	(0.002)
$\lambda^{(1)}$	0.500	0.570	0.603	0.346
		(0.009)	(0.005)	(0.002)
$\lambda^{(2)}$	0.500	0.516	0.449	0.339
		(0.004)	(0.004)	(0.004)
$\lambda^{(3)}$	0.500	0.520	0.451	0.344
		(0.004)	(0.004)	(0.004)
$\lambda^{(4)}$	0.500	0.504	0.453	0.350
		(0.004)	(0.004)	(0.004)
$\lambda^{(5)}$	0.500	0.522	0.462	0.348
		(0.004)	(0.004)	(0.004)

Table A.31 Highly Endogenous Case, 29% of products unsold

($\beta_0 = -10$, $\sigma_\xi = 0.5$, $\sigma_{e\xi} = 0.5$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−10.000	−9.873	−9.062	−9.448
		(0.024)	(0.004)	(0.003)
$\bar{\beta}^{(1)}$	1.000	1.008	0.827	0.965
		(0.003)	(0.001)	(0.001)
$\bar{\beta}^{(2)}$	1.000	0.944	0.854	0.878
		(0.001)	(0.001)	(0.001)
$\bar{\beta}^{(3)}$	1.000	0.944	0.853	0.877
		(0.001)	(0.001)	(0.001)
$\bar{\beta}^{(4)}$	1.000	0.951	0.854	0.877
		(0.002)	(0.001)	(0.001)
$\bar{\beta}^{(5)}$	1.000	0.946	0.853	0.876
		(0.001)	(0.001)	(0.001)
$\lambda^{(1)}$	0.500	0.400	0.635	0.461
		(0.007)	(0.003)	(0.001)
$\lambda^{(2)}$	0.500	0.496	0.487	0.453
		(0.002)	(0.002)	(0.001)
$\lambda^{(3)}$	0.500	0.495	0.485	0.450
		(0.002)	(0.002)	(0.001)
$\lambda^{(4)}$	0.500	0.484	0.486	0.448
		(0.002)	(0.002)	(0.001)
$\lambda^{(5)}$	0.500	0.489	0.484	0.449
		(0.002)	(0.002)	(0.001)

Table A.32 Highly Endogenous Case with Popular Outside Option, 60% of products unsold ($\beta_0 = -14$, $\sigma_\xi = 0.5$, $\sigma_{e\xi} = 0.5$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-14.000	-13.194	-11.320	-12.375
		(0.030)	(0.012)	(0.007)
$\bar{\beta}^{(1)}$	1.000	0.990	0.356	0.905
		(0.012)	(0.008)	(0.002)
$\bar{\beta}^{(2)}$	1.000	0.864	0.752	0.782
		(0.002)	(0.002)	(0.001)
$\bar{\beta}^{(3)}$	1.000	0.863	0.752	0.782
		(0.002)	(0.002)	(0.001)
$\bar{\beta}^{(4)}$	1.000	0.879	0.753	0.783
		(0.002)	(0.002)	(0.001)
$\bar{\beta}^{(5)}$	1.000	0.876	0.750	0.780
		(0.002)	(0.002)	(0.001)
$\lambda^{(1)}$	0.500	0.401	0.683	0.408
		(0.008)	(0.006)	(0.001)
$\lambda^{(2)}$	0.500	0.469	0.446	0.423
		(0.002)	(0.003)	(0.002)
$\lambda^{(3)}$	0.500	0.474	0.445	0.415
		(0.002)	(0.003)	(0.002)
$\lambda^{(4)}$	0.500	0.453	0.448	0.421
		(0.002)	(0.003)	(0.002)
$\lambda^{(5)}$	0.500	0.461	0.440	0.418
		(0.002)	(0.003)	(0.002)

Table A.33 Highly Endogenous Case with Influential Unobservable, 34% of products unsold ($\beta_0 = -10$, $\sigma_\xi = 1.5$, $\sigma_{e\xi} = 0.5$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	−10.000	−10.424	−8.283	−8.811
		(0.059)	(0.008)	(0.005)
$\bar{\beta}^{(1)}$	1.000	0.976	0.733	0.861
		(0.008)	(0.002)	(0.001)
$\bar{\beta}^{(2)}$	1.000	0.952	0.740	0.777
		(0.005)	(0.002)	(0.002)
$\bar{\beta}^{(3)}$	1.000	0.959	0.744	0.779
		(0.006)	(0.002)	(0.002)
$\bar{\beta}^{(4)}$	1.000	0.965	0.741	0.775
		(0.006)	(0.002)	(0.002)
$\bar{\beta}^{(5)}$	1.000	0.964	0.743	0.778
		(0.006)	(0.002)	(0.002)
$\lambda^{(1)}$	0.500	0.539	0.589	0.353
		(0.010)	(0.006)	(0.002)
$\lambda^{(2)}$	0.500	0.502	0.439	0.334
		(0.004)	(0.004)	(0.004)
$\lambda^{(3)}$	0.500	0.510	0.450	0.340
		(0.004)	(0.004)	(0.004)
$\lambda^{(4)}$	0.500	0.496	0.454	0.348
		(0.004)	(0.004)	(0.004)
$\lambda^{(5)}$	0.500	0.510	0.455	0.350
		(0.004)	(0.004)	(0.004)

Table A.34 Highly Endogenous Case with Popular Outside Option and Influential Unobservable, 60% of products unsold

($\beta_0 = -14$, $\sigma_\xi = 1.5$, $\sigma_{e\xi} = 0.5$)

Parameter	True Value	Two Stage Estimator	BLP	Gandhi et al. (2020)
β_0	-14.000	-13.864	-10.355	-11.531
		(0.051)	(0.015)	(0.009)
$\bar{\beta}^{(1)}$	1.000	0.920	0.324	0.854
		(0.018)	(0.010)	(0.003)
$\bar{\beta}^{(2)}$	1.000	0.911	0.641	0.686
		(0.004)	(0.003)	(0.002)
$\bar{\beta}^{(3)}$	1.000	0.911	0.645	0.690
		(0.005)	(0.003)	(0.002)
$\bar{\beta}^{(4)}$	1.000	0.955	0.644	0.687
		(0.005)	(0.003)	(0.002)
$\bar{\beta}^{(5)}$	1.000	0.957	0.644	0.689
		(0.005)	(0.003)	(0.002)
$\lambda^{(1)}$	0.500	0.504	0.610	0.299
		(0.013)	(0.009)	(0.002)
$\lambda^{(2)}$	0.500	0.474	0.389	0.319
		(0.005)	(0.006)	(0.005)
$\lambda^{(3)}$	0.500	0.489	0.392	0.331
		(0.005)	(0.006)	(0.005)
$\lambda^{(4)}$	0.500	0.445	0.402	0.331
		(0.005)	(0.005)	(0.005)
$\lambda^{(5)}$	0.500	0.467	0.399	0.340
		(0.005)	(0.005)	(0.005)