Incorporating Micro Data into Differentiated Products Demand Estimation with PyBLP[*]

Christopher Conlon[†]   Jeff Gortmaker[‡]

August 16, 2023

Abstract

We provide a general framework for incorporating many types of micro data from summary statistics to full surveys of selected consumers into Berry, Levinsohn, and Pakes (1995)-style estimates of differentiated products demand systems. We extend best practices for BLP estimation in Conlon and Gortmaker (2020) to the case with micro data and implement them in our open-source package PyBLP. Monte Carlo experiments and empirical examples suggest that incorporating micro data can substantially improve the finite sample performance of the BLP estimator, particularly when using well-targeted summary statistics or "optimal micro moments" that we derive and show how to compute.

If Python is installed on your computer, PyBLP can be installed with the following command:

```
pip install pyblp
```

Up-to-date documentation for the package is available at `https://pyblp.readthedocs.io`.

Keywords: BLP; Micro data; Minimum distance estimation; Sugar tax
JEL Codes: C13; C18; C30; D12; L00; L66

[†]New York University, Stern School of Business: cconlon@stern.nyu.edu
[‡]Corresponding author. Harvard University: jgortmaker@g.harvard.edu

1. Introduction

Estimating supply and demand for differentiated products is a fundamental empirical challenge for a wide range of economic questions. Nearly thirty years ago, Berry, Levinsohn, and Pakes (1995) developed a class of estimators that allow for both flexible substitution patterns and endogenous prices. A key feature of the BLP approach is that it requires only "aggregate data" on prices and total sales of products at the market level, and exploits cross-market variation in prices, demographics, and product assortment in order to estimate flexible substitution patterns (Berry and Haile, 2014).

In many cases, researchers also have access to additional data on the decisions of individual consumers. These data may come from customer surveys (e.g., Maritz surveys recent automobile purchasers), or from tracking of individual purchasers (e.g., through loyalty cards or NielsenIQ panelists). These data are particularly useful when they link demographic information of individuals to characteristics of products, and when they contain information about the choices within individuals across time or product assortment. A growing literature has connected these "micro data" to the "aggregate data" of the classic BLP approach. Two prominent early examples of this "micro BLP" approach include Petrin (2002) and Berry, Levinsohn, and Pakes (2004), and it has been used in a wide variety of applications, 28 of which we list in Table 1.

Despite the popularity of incorporating micro data into BLP-style estimation, the literature lacks a standardized framework that is sufficiently general to encompass most use cases. Except for a few recent papers that use our framework,[1] the authors of nearly every paper in Table 1 implement the BLP estimator on their own, use different notation, and extend the model to incorporate micro data in a problem-specific manner. Not only does this make replication difficult, but the lack of corresponding formal econometric results makes it challenging to evaluate the statistical properties of micro BLP-style estimators. As an example, a key practical question is how one should weight the contributions of "aggregate data" versus "micro data" in the resulting GMM estimator. Different choices may result in substantially different parameter estimates. One advantage of using a standardized framework is this guarantees that such decisions are made in a consistent way.

Along with a standardized framework, we also systematize the types of "micro moments" that researchers can construct from "micro datasets." That is, one could attempt to match:

---

[1]Backus, Conlon, and Sinkinson (2021), Armitage and Pinter (2022), and Conlon and Rao (2023) use our software package PyBLP to estimate micro BLP models. More papers use PyBLP to estimate BLP-style models with only aggregate data, but since our focus in this paper is on incorporating micro data, we do not collect a list of these other papers in this article.

(a) the correlation or covariance between price paid and income; (b) the average income of consumers who purchase particular products; (c) the average price paid for consumers of different income levels; or (d) the probability of purchasing certain cheap or expensive products for consumers of specific income levels. All of these moments are ways to measure similar features of the same joint distribution. Which moments researchers ultimately employ may largely be driven by convenience or necessity. Surveys tend to report a series of marginal distributions or "crosstabs" without providing the underlying individual responses, and industry reports (or other academic papers) may provide only simple summary statistics, instead of a complete dataset with individual choices. One issue we address is the extent to to which simple moments can approximate the information contained in a complete sample of individual decisions. In doing so, we also provide a characterization of the "optimal micro moments" in the spirit of Chamberlain (1987).

A second challenge, *compatibility*, arises when "aggregate data" and "micro data" are sampled from different populations or according to different sampling schemes (as in, e.g., Imbens and Lancaster, 1994). For example, researchers might have over a decade of purchase data on automobiles, but a consumer survey from only a single year. Alternatively, survey data may oversample individuals who are likely to purchase vehicles, suggest a different distribution of income than the overall population, or simply have variables that are measured differently than in the aggregated purchase data. In these cases, adding certain forms of micro moments may make estimates worse rather than better. Certain forms of micro moments may be more or less robust to these issues. Systematizing the types of micro moments researchers can construct allows us to be explicit about these challenges, and to discuss the pros and cons of different approaches to addressing them.

In our prior work, Conlon and Gortmaker (2020), we presented best practices for BLP-style estimation with aggregate data, and provided a common framework, PyBLP, which implements these best practices in an open-source Python package.[2] The goal of this article is to extend these best practices to the case with micro data and make recommended techniques accessible to a wider range of researchers through PyBLP. For brevity's sake, we will refer to Conlon and Gortmaker (2020) whenever possible, particularly for more in-depth discussion of computation and simulation. In this article, after building up enough notation to define the aggregate BLP estimator in Section 2, we focus more on the applied econometrics that come with combining different sources of data into a single estimator.

---

[2]We recommend installing PyBLP on top of an Anaconda distribution, which comes pre-packaged with PyBLP's primary dependencies. Users of other languages such as MATLAB, Julia, and R can use PyBLP with packages that allow for between-language interoperability (e.g., reticulate for R).

Our work builds on a growing literature aimed at improving and better understanding the econometric properties of BLP-style estimation. Particularly important papers are Berry and Haile (2014, 2022), which develop a nonparametric framework for studying identification of BLP-style models using aggregate and micro data. Complementary to nonparametric results, in Section 3 we rely on economists' intuition from linear IV problems and econometric results in Salanié and Wolak (2022) to illustrate when aggregate data may be insufficient to accurately estimate key demand parameters. In Section 7 we run large-scale Monte Carlo experiments, which provide insights into how nonparametric identification results translate to finite samples.

A key contribution of this article is to provide a standardized econometric framework for how to incorporate many different types of micro data into BLP-style estimation. In Section 4 we characterize micro datasets as information from statistically independent surveys of potentially selected consumers. Conditional on aggregate data (product characteristics and underlying demographic distributions), a survey administrator selects a finite set of underlying consumers with known sampling probabilities. Information from the resulting dataset (consumer choices and demographics) can be incorporated into estimation by adding "micro moments," which match observed statistics with their model counterparts.

In Section 5 we demonstrate how our framework encompasses essentially the same micro moments used by all of the prior literature in Table 1. To demonstrate this, in Section 8 we show how using our framework, PyBLP can estimate the model in Petrin (2002) with only a few lines of code. We also provide a more in-depth empirical example estimating demand for soft drinks with NielsenIQ data. For estimating parameters that govern how consumers with different demographics value different product characteristics, we point to micro moments that contain information about the covariance between demographics and product characteristics. For estimating parameters that govern the degree of unobserved preference heterogeneity, we point to second choice data about what consumers would have chosen had their first choice been unavailable (as in, e.g., Berry, Levinsohn, and Pakes, 2004).

Our framework, however, is much more general, and supports matching a wide range of statistics computed from surveys with many forms of selection. Our goal is to cover most empirical use-cases, including using all the information in a micro sample. Supported statistics include smooth functions of sample means, including correlations and regression coefficients. Underlying samples of consumers can be selected based on their market, demographics, or even endogenous product choices. Supporting general forms of choice-based sampling is particular important because many surveys are targeted at consumers who have

3

purchased certain products.

We provide asymptotic analysis of our standardized framework under each asymptotic thought experiment that seems reasonable. This builds on Myojo and Kanazawa (2012), who extend the many products asymptotics of Berry, Linton, and Pakes (2004) to a specific type of micro moments originally used by Petrin (2002).[3] Through deriving asymptotic results, we highlight a few cases in which the micro BLP estimator can break down, but in general find that the estimator can perform well under many sizes of aggregate and micro data. In Section 7's Monte Carlo experiments we confirm that desirable asymptotic properties translate to finite samples.

A potential concern is that the standard error estimators used by a number of papers, including Petrin (2002), require knowledge of the sample covariance matrix of micro summary statistics. Although a survey may report the average income by purchase group, it is unlikely to report the sample covariances between these averages. Thankfully, knowledge of this additional information is not necessary for inference. In Appendix D we derive novel analytic expressions for the asymptotic covariance matrix of a very broad class of micro moments, which allow researchers to form consistent standard error estimators with only the micro summary statistics themselves and information about the number of underlying observations.

In addition to a new framework for micro moments, in Section 6 we also contribute a novel characterization of the "optimal micro moments" and a simple procedure for computing them. In a best-case scenario when we observe and are willing to use all the results from a consumer survey that is fully compatible with the aggregate data, we can construct micro moments that match a consistent estimate of the average score function of the micro data. Along with consistent estimates of an optimal weighting matrix and Chamberlain's (1987) optimal instruments,[4] the resulting estimator is statistically efficient within the class of all possible micro BLP estimators.[5] Computing each of these components requires only a few lines of code with PyBLP.

Characterizing the optimal micro moments also allows us to explore what types of summary statistics researchers may wish to collect if unable or unwilling to use a full micro

---

[3]For the many markets case, Freyberger (2015) and Hong, Li, and Li (2021) study asymptotics for the aggregate BLP estimator. Grieco, Murry, Pinkse, and Sagl (2023) study many market asymptotics for their closely-related estimator that also combines aggregate and micro data.

[4]In Conlon and Gortmaker (2020) we discuss optimal instruments at length and how to obtain computationally-cheap approximations to them.

[5]After defining relevant notation in Sections 2 and 4, we delineate this class more clearly in Section 6 and prove efficiency in Appendix E. It does not contain estimators that relax the Berry et al. (1995) share constraint, such as Grieco, Murry, Pinkse, and Sagl's (2023) CLER estimator.

dataset in estimation. Inspecting the functional form of micro data scores provides intuition about why some standard micro moments in the literature perform particularly well, and why so-far unused micro moments can perform better. In Section 7's Monte Carlo experiments, we study the relative performance of standard, less-standard, and optimal micro moments, while also building up intuition for best practices involving aggregate variation, pooling statistics across markets, and numerical integration.

In Section 8 we bring these best practices to a real-data example, in which we use NielsenIQ scanner and consumer survey data—two of the most used sources of data in the recent industrial organization literature—to estimate pre-2017 demand for soft drinks in Seattle. We then predict what would happen if prices increased by how much they did after the 2018 implementation of Seattle's sweetened beverage tax (SBT), and compare our substitution estimates to what actually happened. We expect that a structural approach to predicting policy effects is most useful in settings with limited reduced form evidence; however, we choose a SBT because we can compare our results with those from existing studies about the Seattle SBT.[6] We obtain similar results to what actually happened. Incorporating micro data allows us to break down our predictions by demographic group and achieve more realistic substitution patterns. Incorporating second choice data, which we show how to collect in a simple online survey, allows us to even better discipline substitution patterns, particularly to the outside good. By going through a sizable empirical exercise in detail, we hope to make clear what using our framework and implementing our recommendations looks like in practice.

Our work on optimal micro moments builds on literature that uses the likelihood of micro data in BLP-adjacent estimation, starting with Goolsbee and Petrin (2004) and Chintagunta and Dubé (2005). Most similar is Grieco, Murry, Pinkse, and Sagl's (2023) conformant likelihood-based estimator with exogeneity restrictions (CLER), which efficiently combines the full likelihood from individual choices with product-level aggregate moments. The main distinction is that our approach starts with relatively complete aggregate data and augments the BLP approach with additional statistics taken from surveys or other sources, while the CLER approach starts with the likelihood of relatively complete individual choices and augments this with moments from aggregate purchases. Compared to using optimal micro moments in our framework, there are costs and benefits to instead using the CLER approach to estimate demand, making each approach more appropriate in different settings.

_____

[6]For example, Powell and Leider (2020) uses a differences-in-differences approach, comparing with Portland, to measure price passthrough and substitution after the introduction of the tax.

CLER will be more statistically efficient for researchers who are interested in estimating demand with full individual choice data and fully-compatible aggregate data.[7] Since CLER requires optimizing over each product's mean utility as a separate nonlinear parameter, it may be less-well suited for researchers who are interested in matching summary statistics, incorporating supply-side information, or otherwise extending the BLP estimator with moments that introduce non-convexities into the objective function.[8]

There is also a recent literature of alternative approaches to BLP problems, some of which we discuss, and others of which are beyond the scope of this paper. In Appendix B we discuss how our results extend to the random coefficients nested logit (RCNL) model of Brenkers and Verboven (2006). PyBLP fully supports the RCNL model, as well as an approximation to the pure characteristics model of Berry and Pakes (2007). Similar to Salanié and Wolak's (2022) estimator that we discuss in Section 3, Lee and Seo (2015) also provide another approximate BLP estimator. Dubé, Fox, and Su (2012) propose an estimation algorithm for the aggregate BLP estimator based on the mathematical programming with equilibrium constraints (MPEC) method of Su and Judd (2012), which Conlon (2013) extends to generalized empirical likelihood (GEL) estimators. Hong, Li, and Li (2021) propose estimating a Laplace-type estimator with Hamiltonian Monte Carlo. In this article we follow Conlon and Gortmaker (2020) and focus on the more popular nested fixed point approach to computation.

2. Aggregate Data and Estimation Framework

In the left column of Table 2 we summarize the notation that we will introduce in this section. Throughout, we will use language that refers to consumers purchasing products in markets. However, the model is much more general, and can be used to study many types of decision-makers choosing from various types of choice sets.

Aggregate Data

Aggregate data are split into independent and identically distributed markets[9] that represent different realized choice sets for different consumers. Each market $t \in \mathcal{T}$ has a finite set of products $\mathcal{J}_t$, a finite set of consumer types $\mathcal{I}_t$, and a market size $\mathcal{M}_t \in \mathbb{R}$ that measures the

---

[7]CLER achieves efficiency gains by relaxing Berry, Levinsohn, and Pakes's (1995) constraint that observed shares equal those predicted by the model.

[8]Common extensions that we consider in this paper that CLER should easily support include using second choice data, adding a nesting structure, and most common forms of sampling selection.

[9]This can be relaxed in standard ways to incorporate various forms of cross-market dependence, which we can account for with, for example, clustered standard errors. See Appendix D for more details.

mass of consumers in the market.

Each product $j \in \mathcal{J}_t$ has characteristics $(x_{jt}, z_{jt}, \xi_{jt})$. There are $c = 1, \ldots, C$ observed characteristics $x_{jt} = (x_{1jt}, \ldots, x_{Cjt})' \in \mathbb{R}^{C \times 1}$ that directly affect consumer demand. Typically, x_{jt} includes both exogenous characteristics, which are uncorrelated with mean-zero unobserved quality $\xi_{jt} \in \mathbb{R}$, and endogenous characteristics, such as price, which we expect to be correlated with ξ_{jt}. Instruments $z_{jt} = (z_{1jt}, \ldots, z_{M_Ajt})' \in \mathbb{R}^{M_A \times 1}$ include the exogenous characteristics in x_{jt} along with other exogenous observables and will be interacted with ξ_{jt} to form M_A moments from the aggregate data.

Each consumer type $i \in \mathcal{I}_t$ has characteristics $(w_{it}, y_{it}, \nu_{it})$ and constitutes a known share $w_{it} \in [0, 1]$ of consumers in the market, where $\sum_{i \in \mathcal{I}_t} w_{it} = 1$. Each consumer type has $r = 1, \ldots, R$ observed demographics $y_{it} = (y_{1it}, \ldots, y_{Rit})' \in \mathbb{R}^{R \times 1}$ and unobserved preferences $\nu_{it} = (\nu_{1it}, \ldots, \nu_{Cit})' \in \mathbb{R}^{C \times 1}$ for the observed characteristics x_{jt}. Typically, demographics y_{it} will be sampled from census data or some other representative survey and ν_{it} will be a numerical approximation to an independent standard normal distribution. Types can be interpreted as a fixed set of Monte Carlo draws or another numerical approximation to a continuous distribution with integration weights w_{it}.[10]

We focus on a discrete set of consumer types for notational convenience and practical relevance. In practice, most researchers use a fixed number of Monte Carlo draws from the distribution of demographics and unobserved preferences. For brevity's sake, in this paper we do not discuss the econometric implications of simulation error resulting from only using a finite number of draws. Instead, in our Monte Carlo experiments and replication exercise, we use best practices from Conlon and Gortmaker (2020), which involve either using a large number of scrambled Halton draws (Owen, 2017) or an appropriate quadrature rule, and do not account for simulation error when computing standard errors.[11]

The mass $\mathcal{M}_t$ of consumers are differentiated by type $i \in \mathcal{I}_t$ and by idiosyncratic preferences $\varepsilon_{ijt} \in \mathbb{R}$ for each product $j \in \mathcal{J}_t$ and the outside alternative, denoted $j = 0$. Indirect utility from selecting $j \in \mathcal{J}_t \cup \{0\}$ is[12]

$$u_{ijt} = \delta_{jt} + \mu_{ijt} + \varepsilon_{ijt}, \quad u_{i0t} = \varepsilon_{i0t}. \tag{1}$$

---

[10]Simple Monte Carlo draws are equally-weighted, with $w_{it} = 1/|\mathcal{I}_t|$. Types may have different weights, for example, when demographics are sampled from a survey with sampling weights or when quadrature is used to approximate a continuous distribution.

[11]For those who are concerned about simulation error, PyBLP does support resampling consumer types to compute an estimate of the contribution of simulation error to the BLP or micro BLP estimator's asymptotic covariance matrix.

[12]Identification requires two normalizations. We follow standard practice by normalizing $\delta_{0t} = \mu_{i0t} = 0$. Levels of δ_{jt} and μ_{ijt} are then interpreted as relative to those of the outside option.

Mean utility $\delta_{jt} \in \mathbb{R}$ is common across consumer types and depends on product characteristics (x_{jt}, ξ_{jt}). Typically, an additivity assumption is made so that

$$\delta_{jt} = x'_{jt}\beta + \xi_{jt}. \tag{2}$$

The heterogeneous component of utility $\mu_{ijt} \in \mathbb{R}$ differs across types and will additionally depend on demographics and preferences (y_{it}, ν_{it}). A popular functional form is

$$\mu_{ijt} = x'_{jt}(\Pi y_{it} + \Sigma \nu_{it}). \tag{3}$$

With normally distributed unobserved preferences $\nu_{it} \sim N(0, I)$, indirect utility can be written as $u_{ijt} = x'_{jt}\beta_{it} + \varepsilon_{ijt}$ with random coefficients distributed $\beta_{it} \sim N(\beta + \Pi y_{it}, \Sigma\Sigma')$. We focus on this functional form because it is by far the most popular, but we also discuss the three most common variants, which we also support in PyBLP. First, to guarantee downward sloping demand for all consumers, one can replace the random coefficient β_{cit} on price $x_{cjt} = p_{jt}$ with a lognormal random coefficient (see Appendix A). Second, to parsimoniously incorporate geographic distance or other important product-specific demographics y_{ijt}, one can replace interactions between product dummies in x_{jt} and demographics in y_{it} with y_{ijt}. Third, one can use other parametric distributions for ν_{it}, such as exponential or χ^2 distributions.

Each consumer chooses among the discrete alternatives $j \in \mathcal{J}_t \cup \{0\}$ and selects the option that maximizes indirect utility. With type I extreme value idiosyncratic preferences ε_{ijt}, the logit probability that a consumer of type $i \in \mathcal{I}_t$ chooses a product $j \in \mathcal{J}_t$ is[13]

$$s_{ijt} = \frac{\exp(\delta_{jt} + \mu_{ijt})}{1 + \sum_{k \in \mathcal{J}_t} \exp(\delta_{kt} + \mu_{ikt})}. \tag{4}$$

Again, we focus on this distribution for ε_{ijt} because it is by far the most popular.[14] In Appendix B we discuss the most common variant, which is to assume that ε_{ijt} follows the assumptions of a two-level nested logit. The resulting random coefficients nested logit (RCNL) model of Brenkers and Verboven (2006), which we also support in PyBLP, is popular in applications where the most important product characteristic governing substitution is

---

[13]The one in the denominator is from the outside alternative normalization $\delta_{0t} = \mu_{i0t} = 0$.

[14]The pure characteristics model of Berry and Pakes (2007), which PyBLP can approximate, eliminates idiosyncratic preferences ε_{ijt} altogether. Although our focus is on more tractable models with ε_{ijt}, incorporating micro data does allow for the estimation of more flexible models in which heterogeneous utility μ_{ijt} dominates, reducing dependence on ε_{ijt} that can otherwise contribute to unrealistic substitution patterns.

categorical.

Aggregate market shares are given by integrating over the mass of consumers. The mixed logit market share of product $j \in \mathcal{J}_t$ is

$$s_{jt} = \sum_{i \in \mathcal{I}_t} w_{it} \cdot s_{ijt}. \tag{5}$$

We use s_{jt} to refer to generic market shares, potentially evaluated at different parameters $s_{jt}(\theta)$. We use $\mathcal{S}_{jt} = s_{jt}(\theta_0)$ to refer to the observed market shares generated by the true parameters θ_0.

The goal is to recover the true parameters $\theta_0 = (\beta_0, \Pi_0, \Sigma_0)$ that characterize the demand system. Since we will frequently refer to our earlier work, it is worth pointing out a difference in notation. In Conlon and Gortmaker (2020), we partitioned θ into three parts: θ_1 referred to β; θ_2, to (Π, Σ); and θ_3, to supply-side parameters.[15] Since our focus here is on the demand side, we use the notation $\theta = (\beta, \Pi, \Sigma)$, not $\theta = (\theta_1, \theta_2, \theta_3)$.

Towards recovering θ_0, the researcher first makes an assumption about how to define markets $t \in \mathcal{T}$ and their sizes $\mathcal{M}_t$. For each product, the researcher collects characteristics, instruments, and market shares: $\{(x_{jt}, z_{jt}, \mathcal{S}_{jt})\}_{j \in \mathcal{J}_t}$. Typically, market shares are observed quantities divided by the assumed number of consumers in the market. Finally, the researcher makes an assumption about consumer types: $\{(w_{it}, y_{it}, \nu_{it})\}_{i \in \mathcal{I}_t}$.

Aggregate BLP Estimator

Since unobserved quality $\xi_{jt} \in \mathbb{R}$ is mean-zero and uncorrelated with the M_A instruments z_{jt}, our assumptions about the aggregate data deliver M_A moment conditions $\mathbb{E}[\xi_{jt} \cdot z_{jt}] = 0$.[16] Using these, we can construct a GMM estimator for θ from $N_A = \sum_{t \in \mathcal{T}} |\mathcal{J}_t|$ aggregate observations and a weighting matrix $\hat{W}_A$ where hats denote sample approximations:[17]

$$\hat{\theta}_A = \underset{\theta}{\operatorname{argmin}}\, \hat{g}_A(\theta)' \hat{W}_A \hat{g}_A(\theta), \quad \hat{g}_A(\theta) = \frac{1}{N_A} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}_t} (\underbrace{\hat{\delta}_{jt}(\Pi, \Sigma) - x'_{jt}\beta}_{\hat{\xi}_{jt}(\theta)}) \cdot z_{jt}. \tag{6}$$

The key insight of Berry, Levinsohn, and Pakes (1995), building on Berry (1994), is that

---

[15]With a supply side, the parameter in β on price would instead be in θ_2.

[16]It is common to assume $\mathbb{E}[\xi_{jt} \mid z_{jt}] = 0$ and convert these conditional moments into unconditional ones.

[17]Typically, we solve this problem twice. Once to obtain a consistent estimator for the optimal weighting matrix—and for the optimal instruments, if appropriate—and a second time to obtain the efficient estimator. The most common choice for the initial weighting matrix is the 2SLS weighting matrix, which would be efficient if ξ_{jt} were homoskedastic.

we can invert the demand system and recover δ_{jt} from $s_{jt}(\Pi, \Sigma, \delta_t)$ by matching the observed shares $\mathcal{S}_{jt}$. In each market $t \in \mathcal{T}$, we can solve a system of $|\mathcal{J}_t|$ nonlinear equations to find the unique $|\mathcal{J}_t|$ mean utilities $\hat{\delta}_{jt}(\Pi, \Sigma)$ that equate observed market shares $\mathcal{S}_{jt}$ with their model counterparts:

$$\mathcal{S}_{jt} = s_{jt}(\Pi, \Sigma) \equiv \sum_{i \in \mathcal{I}_t} w_{it} \cdot \frac{\exp[\hat{\delta}_{jt}(\Pi, \Sigma) + x_{jt}'(\Pi y_{it} + \Sigma \nu_{it})]}{1 + \sum_{k \in \mathcal{J}_t} \exp[\hat{\delta}_{kt}(\Pi, \Sigma) + x_{kt}'(\Pi y_{it} + \Sigma \nu_{it})]}, \quad \forall j \in \mathcal{J}_t. \quad (7)$$

The econometric properties of this estimator under many products are discussed in Berry, Linton, and Pakes (2004); many markets, in Freyberger (2015) and Hong, Li, and Li (2021).[18] In Conlon and Gortmaker (2020) we discuss modern best practices for this type of estimation and implement them as defaults in PyBLP: fast and stable algorithms for solving the inner problem for $\hat{\delta}_{jt}(\Pi, \Sigma)$ and the outer problem for $\hat{\theta}_A$, fast and accurate ways to integrate over consumer types $i \in \mathcal{I}_t$, robust solutions to various numerical challenges, and when appropriate, the use of fixed effect absorption and Chamberlain's (1987) optimal instruments. Throughout this article, we continue to use all of these best practices for the aggregate portion of estimation.

A common extension, which we discuss at length in Conlon and Gortmaker (2020), is to derive an additional set of aggregate moment conditions from the first-order pricing conditions of firms and to append the sample analogues of these moments to those in $\hat{g}_A(\theta)$. Especially when using an approximation to the optimal instruments, incorporating well-specified supply-side moments can substantially improve the performance of the aggregate estimator. However, in this article, we primarily focus on the demand-only model to highlight the contribution of micro data.[19]

3. Aggregate Variation Only

Absent a well-specified supply side, it can be difficult to flexibly estimate the nonlinear parameters (Π, Σ) governing heterogeneous tastes without strong instruments and substantial cross-market variation. The frustrating result is an estimator with poor econometric performance or an estimated demand system with unreasonable substitution patterns. In this section, we discuss the identification of the aggregate model to motivate incorporating micro data.

---

[18]In Appendix D we discuss both many products, $|\mathcal{J}_t| \to \infty$, and many markets, $|\mathcal{T}| \to \infty$.

[19]Of course, PyBLP supports combining micro and supply-side moments. We provide an example of this in Section 8, where we use both to replicate Petrin (2002).

Intuitively, identification of (Π, Σ) requires cross-market variation in demographic distributions and choice sets. For a fully nonparametric treatment of identification with aggregate, market-level data, see Berry and Haile (2014) or the summary in Section 5 of Berry and Haile (2021). Our experience is that a good starting point for understanding whether there is sufficient aggregate variation is the typical economist's strong intuition about linear IV regression models.

Intuition from Linear Regression

To leverage this intuition, we use results in Salanié and Wolak (2022), who approximate the aggregate estimator in (6) with a linear IV regression. We write down the full approximation using this paper's notation in Appendix C but here consider only the simplest scalar case with $C = 1$ product characteristic, $R = 1$ demographic, and three parameters, $\theta = (\beta, \pi, \sigma)$. A second-order Taylor expansion around $\pi = \sigma = 0$ gives the following linear model with four regressors:[20]

$$\log \frac{s_{jt}}{s_{0t}} \approx \beta x_{jt} + \sigma^2 a_{jt} + \pi m_t^y x_{jt} + \pi^2 v_t^y a_{jt} + \xi_{jt}, \quad a_{jt} = \left(\frac{x_{jt}}{2} - \sum_{k \in \mathcal{J}_t} s_{kt} \cdot x_{kt} \right) \cdot x_{jt} \quad (8)$$

where $m_t^y = \sum_{i \in \mathcal{I}_t} w_{it} \cdot y_{it}$ is the within-market demographic mean, $v_t^y = \sum_{i \in \mathcal{I}_t} w_{it} \cdot (y_{it} - m_t^y)^2$ is its variance, and a_{jt} is an "artificial regressor" that reflects within-market differentiation of the product characteristic x_{jt}.[21] If $\pi = \sigma = 0$, the approximation is exact, and collapses to a simple logit regression: $\log(s_{jt}/s_{0t}) = \delta_{jt} = \beta x_{jt} + \xi_{jt}$.

The linear model in (8) is only an approximation, but its intuition about identification translates fairly well to the full model. First, without an instrument for the artificial regressor a_{jt} we should expect our estimate for σ^2 to be asymptotically biased—a_{jt} is a function of endogenous market shares s_{kt}, which are correlated with unobserved quality ξ_{jt}.[22] The "differentiation IVs" proposed by Gandhi and Houde (2020) and further evaluated in Conlon and Gortmaker (2020) look similar to a_{jt} and work well in practice compared to other types of "BLP instruments" that are functions of other products' exogenous characteristics.[23] Indeed,

---

[20]Here, s_{0t} does not refer to a "true" share, like the true θ_0, but just the outside share for $j = 0$.

[21]Salanié and Wolak (2022) give additional intuition for the functional form of a_{jt}. A quadratic form is unsurprising because x_{jt} multiplies ν_{it}. The $\frac{1}{2}$ comes from the symmetric shape of the logistic distribution.

[22]In a fully nonparametric model, a different instrument is needed for each of the $|\mathcal{J}_t|$ market shares (Berry and Haile, 2014).

[23]In Section 7 we use the "quadratic" version of differentiation IVs in our Monte Carlo experiments. For this example, differentiation IVs would be $z_{jt} = (x_{jt}, \hat{a}_{jt}, m_t^y x_{jt}, v_t^y \hat{a}_{jt})'$ where $\hat{a}_{jt} = \sum_{k \neq j} (x_{kt} - x_{jt})^2$. Expanded, $\hat{a}_{jt} = x_{jt}^2 - 2x_{jt} \sum_{k \neq j} x_{kt} + \sum_{k \neq j} x_{kt}^2$ and $a_{jt} = x_{jt}^2/2 - x_{jt} \sum_{k \neq j} s_{kt} x_{kt}$. The main difference is the share s_{kt}-weighted average of x_{kt} in a_{jt} instead of the unweighted average in $\hat{a}_{jt}$. Weighting by share is

the first stage of an IV regression using differentiation IVs implements precisely the "IIA test" recommended by Gandhi and Houde (2020): estimate the simple logit regression controlling for differentiation IVs and consider a richer model if the IVs are statistically relevant.

Second, absent significant cross-market variation in assortment $\mathcal{J}_t$, the artificial regressor a_{jt} will be nearly collinear with x_{jt} and x_{jt}^2, and it will be difficult to separately identify σ^2 from linear coefficients. This aligns with the standard intuition that with only aggregate data, the degree of unobserved preference heterogeneity, here measured by σ^2, is identified by how consumers substitute between products when faced with cross-market variation in choice sets.

Third, if the distribution of the demographic y_{it}, here measured by its mean m_t^y and variance v_t^y, does not vary much across markets, the regressors $m_t^y x_{jt}$ and $v_t^y a_{jt}$ will be nearly collinear with x_{jt} and a_{jt}, and it will be difficult to separately identify π and π^2 from β and σ^2. Absent cross-market variation in (m_t^y, v_t^y), distinctions between taste variation from demographics versus unobserved heterogeneity will be solely driven by functional form. With only aggregate data, separate identification of (Π, Σ) requires cross-market variation in demographics.

Even when using appropriate instruments, a lack of cross-market choice set and demographic variation will either result in poor estimators of (Π, Σ) or leave researchers with no alternative other than to estimate a more restrictive demand system. Supply restrictions aside, the typical solution is to exploit within-market variation from micro data that links demographics to individual choices, rather than aggregate market shares.

In practice, our recommendation when considering estimating a demand system with only aggregate data aligns with those of Salanié and Wolak (2022) and Gandhi and Houde (2020). We recommend first running a version of the IV regression in (8), with the full version written out in Appendix C, to get a sense of whether aggregate variation will be sufficient to estimate a flexible demand system.[24] If so, the estimates from this regression will give a sense of what reasonable starting values and parameter bounds may look like.

4.   A Unified Framework for Micro Data and Estimation

In the right column of Table 2, we summarize additional notation that we will introduce in this section. We begin by explaining our notation and the framework we will use to charac-

---

infeasible because market shares are endogenous. The other difference is the "BLP instrument" $\sum_{k \neq j} x_{kt}^2$.

[24]With a reasonably small number of characteristics and demographics, is perhaps simplest to treat π^2 as an unconstrained fourth parameter, say γ, and to estimate π only from cross-market variation in demographic means m_t^y, while "controlling" for $v_t^y a_{jt}$.

terize "micro datasets," indexed by d. We then build up additional notation to incorporate "micro moments" into the BLP estimator. Our "micro moments," indexed by m, are smooth functions of "micro parts," indexed by p, which are in turn conditional expectations of scalar functions we call "micro values."

Survey Data

We begin with the assumption that micro data are split into datasets $d \in \mathcal{D}$ that report results from statistically independent consumer surveys. Statistically, micro data are generated conditional on all aggregate data: products $\mathcal{J}_t$, consumer types $\mathcal{I}_t$, and sizes $\mathcal{M}_t$ of all markets $t \in \mathcal{T}$.[25] We use the notation $\mathbb{P}_A$, $\mathbb{E}_A$, and $\mathbb{V}_A$ to denote probabilities, expectations, and variances conditional on all aggregate data.

Each consumer n is defined by a 3-tuple (t_n, i_n, j_n) and chooses $j \in \mathcal{J}_t \cup \{0\}$ with (mixed) logit choice probability $\mathbb{P}_A(j_n = j \mid t_n = t, i_n = i) = s_{ijt}$ following (4). Likewise, within a market, the weight corresponding to each consumer type $i \in \mathcal{I}_t$ is the same as in the aggregate demand model (5), and is given by $\mathbb{P}_A(i_n = i \mid t_n = t) = w_{it}$. These types i and weights w_{it} include both observed demographics y_{it} and unobserved preferences ν_{it}.

However, not all consumers need be observed in a micro dataset. Instead, we assume that a survey administrator selects a finite set of consumers $n \in \mathcal{N}_d$ with independent sampling probabilities $\mathbb{P}_A(n \in \mathcal{N}_d \mid t_n = t, i_n = i, j_n = j) = w_{dijt}$. Most common survey designs can be represented with different sampling probabilities w_{dijt}, including arbitrary stratification by the consumer's market, type, and even choice. For a survey to be useful, we need to know how it was conducted, so we will assume that the researcher knows the sampling probabilities w_{dijt} for each dataset $d \in \mathcal{D}$.

Consider some simple examples. If the survey randomly samples from all consumers in different markets, sampling probabilities should be proportional to the number of consumers in each market, $w_{dijt} \propto \mathcal{M}_t$. An alternative would be to stratify across markets so that consumers are sampled from each market with equal probability, $w_{dijt} \propto 1/|\mathcal{T}|$. Other common sampling schemes might only sample individuals conditional on making a purchase, $w_{dijt} \propto 1\{j \neq 0\}$, or on purchasing a particular brand b, with $w_{dijt} \propto 1\{j \in \mathcal{J}_b\}$. It is also common to sample individuals whose income y_{rit} is above or below some level (such as households eligible for WIC), for example $w_{dijt} \propto 1\{y_{rit} < \$50{,}000\}$. We can combine these into a more detailed example: $w_{dijt} \propto \mathcal{M}_t \cdot 1\{y_{rit} < \$50{,}000, t \in \mathcal{T}_d, j \neq 0\}$ would generate a random sample of consumers from a few markets $\mathcal{T}_d \subset \mathcal{T}$ with income below \$50,000 who

---

[25]Depending on which asymptotic thought experiment from Appendix D is most appropriate, we may also include survey sampling probabilities, defined shortly, in the aggregate data.

make a purchase.

Micro Statistics

Ideally, the researcher would observe a matched dataset of all sampled consumers' markets, choices, and demographics: $\{(t_n, j_n, y_{i_n t_n})\}_{n \in \mathcal{N}_d}$.[26] For example, the NielsenIQ panelist data tracks the products purchased by households, which stores they visit, and the demographics of the corresponding household. In this scenario, we can make full use of all the information in the micro dataset.[27] In many other cases, we will have incomplete data from a limited number of consumers, or summary statistics for subsets of individuals. The extent of our micro data will determine which "micro moments" we can and cannot compute.

We will use each micro summary statistic that we observe to define one of $m = 1, \ldots, M_{\mathrm{M}}$ micro moments. Each micro moment m matches a single summary statistic, which could be a simple average, a weighted average, a conditional average, or even a covariance or regression coefficient.

Consider a simple example. We are interested in capturing the relationship between having children and purchasing a minivan. Suppose we have access to summary statistics from a representative survey of households that purchased a car in $d = 2023$. Specifically, suppose we observe two summary statistics: the average number of kids across surveyed households, as well as the average number of kids across minivan purchasers,

$$\overline{\mathrm{kids}}_{2023} = \frac{1}{N_{2023}} \sum_{n \in \mathcal{N}_{2023}} \mathrm{kids}_{i_n t_n}, \tag{9}$$

$$\overline{\mathrm{kids}}_{2023}^{\mathrm{mini}} = \frac{\frac{1}{N_{2023}} \sum_{n \in \mathcal{N}_{2023}} \mathrm{kids}_{i_n t_n} \cdot 1\{j_n \in \mathcal{J}_{\mathrm{mini}}\}}{\frac{1}{N_{2023}} \sum_{n \in \mathcal{N}_{2023}} 1\{j_n \in \mathcal{J}_{\mathrm{mini}}\}}. \tag{10}$$

We can use these two summary statistics to define $M_{\mathrm{M}} = 2$ micro moments. The first, $\overline{\mathrm{kids}}_{2023}$, is a simple average, and the second, $\overline{\mathrm{kids}}_{2023}^{\mathrm{mini}}$, is the ratio of two simple averages. To cover both of these cases (and many more), our framework supports summary statistics that are smooth functions of simple averages.[28]

We call each simple average a "micro part." Each of the $p = 1, \ldots, P_{\mathrm{M}}$ micro parts is an

---

[26]By definition, the researcher does not know unobserved preferences ν_{it}.

[27]In Section 6 we discuss optimal micro moments that make full use of the information in a micro dataset. In Section 8 we demonstrate how to do so with NielsenIQ data.

[28]Below, we explain how to write weighted averages as simple averages using this same framework.

average over all $N_{d_p} = |\mathcal{N}_{d_p}|$ observations in its micro dataset $d_p \in \mathcal{D}$:

$$\bar{v}_p = \frac{1}{N_{d_p}} \sum_{n \in \mathcal{N}_{d_p}} v_{p i_n j_n t_n}. \tag{11}$$

Each part p is defined as the average of a function $v_p(t_n, j_n, y_{i_n t_n})$, or $v_{p i_n j_n t_n}$ for short, that may depend on the choice conditions (e.g., prices, assortment, and product characteristics) in the market t_n, the consumer demographics $y_{i_n t_n}$, and the selected choices j_n. The choice of $v_p(\cdot)$ is determined both by what statistics are available in our data, and which model parameters we are trying to estimate.

To match $\overline{\text{kids}}_{2023}$ and $\overline{\text{kids}}_{2023}^{\text{mini}}$, we will need to define $P_M = 3$ micro parts: the average number of kids in the micro data, the share of households who purchased a minivan, and the average number of kids multiplied by a dummy for purchasing a minivan,

$$\bar{v}_1 = \frac{1}{N_{2023}} \sum_{n \in \mathcal{N}_{2023}} v_{1 i_n j_n t_n}, \quad v_{1ijt} = \text{kids}_{it}, \tag{12}$$

$$\bar{v}_2 = \frac{1}{N_{2023}} \sum_{n \in \mathcal{N}_{2023}} v_{2 i_n j_n t_n}, \quad v_{2ijt} = 1\{j \in \mathcal{J}_{\text{mini}}\}, \tag{13}$$

$$\bar{v}_3 = \frac{1}{N_{2023}} \sum_{n \in \mathcal{N}_{2023}} v_{3 i_n j_n t_n}, \quad v_{3ijt} = \text{kids}_{it} \cdot 1\{j \in \mathcal{J}_{\text{mini}}\}. \tag{14}$$

We have assumed that in the survey, we directly observe $\bar{v}_1 = \overline{\text{kids}}_{2023}$, but that we do not directly observe $\bar{v}_2$ or $\bar{v}_3$, only their ratio $\bar{v}_3/\bar{v}_2 = \overline{\text{kids}}_{2023}^{\text{mini}}$. To express the simple average $\overline{\text{kids}}_{2023}$, the ratio $\overline{\text{kids}}_{2023}^{\text{mini}}$, and any other smooth function of averages, such as covariances or even regression coefficients, we need to define slightly more notation.

Each micro moment m matches a scalar summary statistic denoted $f_m(\bar{v}) \in \mathbb{R}$, which is a smooth function $f_m : \mathbb{R}^{P_M \times 1} \to \mathbb{R}$ of potentially all micro moment parts $\bar{v} = (\bar{v}_1, \ldots, \bar{v}_{P_M})'$. In our example, our $M_M = 2$ two summary statistics can be written as

$$\overline{\text{kids}}_{2023} = f_1(\bar{v}) = \bar{v}_1, \tag{15}$$

$$\overline{\text{kids}}_{2023}^{\text{mini}} = f_2(\bar{v}) = \bar{v}_3/\bar{v}_2. \tag{16}$$

In general, each micro moment is defined by both its underlying micro values v_{pijt} and its smooth function $f_m(\cdot)$. We expect that most useful summary statistics are smooth functions of averages, so we think that our definition of micro moments is fairly nonrestrictive. We discuss common summary statistics in Section 5.

Model Analogues

Our assumptions about consumer and survey sampling allow us to compute the model analogue for each observed summary statistic $f_m(\bar{v})$. Under the model, each micro moment part $\bar{v}_p(\theta)$ is defined as the expectation of v_{pijt} conditional on the aggregate data and the parameters θ:

$$v_p(\theta) \equiv \mathbb{E}_A^\theta[v_{pi_nj_nt_n}] = \frac{\sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{J}_t\cup\{0\}} w_{it} \cdot s_{ijt}(\theta) \cdot w_{d_pijt} \cdot v_{pijt}}{\sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}_t} \sum_{j\in\mathcal{J}_t\cup\{0\}} w_{it} \cdot s_{ijt}(\theta) \cdot w_{d_pijt}}. \tag{17}$$

Notice that we aggregate over all markets t, individuals i, and products j using the same $s_{ijt}(\theta)$ from (4), and the same w_{it} we use to compute the aggregate shares s_{jt} in (5). We rely on the sampling weights w_{d_pijt} and micro values v_{pijt} to limit each part's calculation to sub-populations of individuals and to calculate conditional expectations. Likewise, by varying the model parameters θ, we are implicitly re-weighting v_{pijt} so that the objective is to choose θ such that the model average from (17) matches the survey average from (11).

The model analogue of the observed micro summary statistic $f_m(\bar{v})$ is $f_m(v(\theta))$ where $v(\theta) = (v_1(\theta), \ldots, v_{P_M}(\theta))'$. At the true θ_0, iterated expectations and the continuous mapping theorem give $m = 1, \ldots, M_M$ conditions $f_m(\bar{v}) - f_m(v(\theta_0)) \xrightarrow{P} 0$.[29] Slightly abusing the definition of a statistical moment, we will call each of these conditions a "micro moment."[30]

---

[29]Here, convergence in probability is not conditional on the aggregate data, so these conditions are statistically compatible with the aggregate moments $\mathbb{E}[\xi_{jt} \cdot z_{jt}] = 0$.

[30]If $f_m(\bar{v}) = \bar{v}_p$ is a simple average, condition m can be interpreted without abusing terminology as a moment $\mathbb{E}[v_{pijt} - v_p(\theta_0)] = 0$. If all summary statistics were simple averages, the minimum distance estimator we will define shortly would instead be a GMM estimator.

Micro BLP Estimator

We can extend the aggregate GMM estimator in (6) with M_M new micro moments and a larger weighting matrix $\hat{W} = \mathrm{diag}(\hat{W}_\mathrm{A}, \hat{W}_\mathrm{M})$.[31] This gives a minimum distance estimator:[32]

$$\hat{\theta} = \underset{\theta}{\mathrm{argmin}}\ \hat{g}(\theta)'\hat{W}\hat{g}(\theta), \quad \hat{g}(\theta) = \begin{bmatrix} \hat{g}_\mathrm{A}(\theta) \\ \hat{g}_\mathrm{M}(\theta) \end{bmatrix}, \quad \hat{g}_\mathrm{M}(\theta) = \begin{bmatrix} f_1(\bar{v}) - f_1(v(\theta)) \\ \vdots \\ f_{M_\mathrm{M}}(\bar{v}) - f_{M_\mathrm{M}}(v(\theta)) \end{bmatrix}. \quad (18)$$

In practice, we can concentrate out the linear parameters β and only optimize over the nonlinear parameters (Π, Σ). For each guess of (Π, Σ), we need to solve the nested fixed point for all mean utilities $\hat{\delta}_{jt}(\Pi, \Sigma)$. The micro BLP estimation algorithm is given below.

Algorithm 1 Nested Fixed Point with Micro Moments

For each guess of the nonlinear parameters (Π, Σ):

1. For each market $t \in \mathcal{T}$, solve (5) for $\hat{\delta}_{jt}(\Pi, \Sigma)$ for all products $j \in \mathcal{J}_t$. In Conlon and Gortmaker (2020) we describe and evaluate different solvers in Sections 3 and 5.

2. For each micro moment $m = 1, \dots, M_\mathrm{M}$, compute $f_m(v(\theta)) = f_m(v(\hat{\delta}(\Pi, \Sigma), \Pi, \Sigma))$ in (17). Stack the micro sample moments $\hat{g}_\mathrm{M}(\theta) = (f_1(\bar{v}) - f_1(v(\theta)), \dots, f_{M_\mathrm{M}}(\bar{v}) - f_{M_\mathrm{M}}(v(\theta)))'$.

3. Recover linear parameters $\hat{\beta}(\Pi, \Sigma)$ from the linear IV GMM regression $\hat{\delta}_{jt}(\Pi, \Sigma) = x'_{jt}\beta + \xi_{jt}$. In Conlon and Gortmaker (2020) we describe fixed effect absorption in Section 3 and the regression in Appendix A.

4. Compute residual unobserved qualities $\hat{\xi}_{jt}(\theta) = \hat{\delta}_{jt}(\Pi, \Sigma) - x'_{jt}\hat{\beta}(\Pi, \Sigma)$. Construct the aggregate sample moments $\hat{g}_\mathrm{A}(\theta) = \frac{1}{N_\mathrm{A}} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}_t} \hat{\xi}_{jt}(\theta) \cdot z_{jt}$.

5. Stack sample moments into $\hat{g}(\theta) = (\hat{g}_\mathrm{A}(\theta)', \hat{g}_\mathrm{M}(\theta)')'$ and construct the objective $\hat{g}(\theta)'\hat{W}\hat{g}(\theta)$.

Since the first use of the micro BLP estimator in Petrin (2002), a wide range of papers, many of which we reference in Section 1, have extended the aggregate BLP estimator with various forms of moments based on micro data. Although each paper uses its own notation and language, in Section 5 we describe how most of these cases fit into our framework.

Although variants of the micro BLP estimator have been used extensively in practice, its

---

[31]The optimal weighting matrix is block diagonal because the aggregate and micro moments are uncorrelated (see Appendix D).

[32]Again, we typically solve this problem twice, once with an initial weighting matrix and again with the optimal one, and, if appropriate, optimal instruments and optimal micro moments. With micro moments, there is no "canonical" choice for the initial weighting matrix, like the 2SLS weighting matrix for the aggregate estimator. Instead, we prefer to compute and invert all moments' covariances at some initial guess for θ_0, which could be informed by estimators based on aggregate data.

econometric properties have received less attention than those of the aggregate estimator. Appendices sometimes provide heuristic discussions of asymptotic covariances (e.g., in Petrin, 2002; Berry, Levinsohn, and Pakes, 2004), and Grieco, Murry, Pinkse, and Sagl (2023) provide formal analysis of many markets asymptotics for their closely-related estimator. However, the only formal asymptotic analysis of a special case of the micro BLP estimator in (18) of which we are aware is in Myojo and Kanazawa (2012), which extends the many products asymptotics of Berry, Linton, and Pakes (2004) with micro moments of the specific form used by Petrin (2002).[33] Both of these papers also study the effect of simulation error, which, again, we omit from this article, but think may be an interesting direction for future research.

In Appendix D we derive the econometric properties of the general micro BLP estimator under different asymptotic thought experiments: (a) many markets, including those covered by surveys; (b) many markets, few with surveys, but the surveys are large; and (c) few markets, but markets and surveys are both large. Asymptotic normality is straightforward to show for cases (a) and (b) because markets are independent. Without a growing number of markets, case (c) requires either ruling out micro moments with asymptotic variances that depend on specific products, or dropping markets covered by micro data from the aggregate moments.[34]

A convenient result in Appendix D is that the choice of asymptotic thought experiment does not affect how we compute the estimator or its asymptotic variance. Additionally, consistent estimators of standard errors can be formed without any external information about the sampling error in the summary statistics $f_m(\bar{v})$ other than the number of observations N_d.[35] This means that in order to do valid inference, researchers do not need to know sample covariances or standard errors for the summary statistics they are matching.[36]

The choice of asymptotic thought experiment does inform how we think about rates of convergence for $\hat{\theta}$. Parameter estimators will in general converge at the faster rate of

---

[33]Myojo and Kanazawa (2012) also incorporate supply-side moments and run a Monte Carlo experiment. In contrast, our focus in this article is on best practices for a broader class of micro BLP estimators under a variety of different asymptotic thought experiments.

[34]For example, matching the mean income for those who purchase a specific product j may make the asymptotic distribution of $\hat{\theta}$ depend on the potentially non-normal characteristics of j. As a robustness check, we could set instruments $z_{jt} = 0$ for all markets t in which we match this moment.

[35]Given N_d and sampling weights w_{dijt}, we can form a consistent estimate of the covariance $\mathbb{C}_A(v_{p i_n j_n t_n}, v_{q i_n j_n t_n})$ between each pair of micro parts p and q, and use the delta method to obtain the asymptotic covariance matrix for the micro moments.

[36]By default, PyBLP computes analytic asymptotic covariances. However, it does allow researchers to specify their own asymptotic covariance matrix for micro moments, so that if researchers can use alternative measures of this matrix, if desired.

the aggregate or micro data, depending on which sample size is larger. The exception is estimators of linear parameters β, which are only identified by aggregate variation in the linear IV regression. If some nonlinear parameters in (Π, Σ) are not identified or only weakly identified by variation from the larger sample, the other sample provides "backup variation" that may still guarantee strong identification at the slower rate.[37] For example, if there is no cross-market variation in the distribution of demographics whatsoever, Π will not be identified from aggregate variation, and its estimator will converge at the rate of the micro data.

The punchline is that under fairly mild and typical assumptions, the micro BLP estimator is asymptotically normal with reasonable rates of convergence. In Section 7 we run large-scale Monte Carlo experiments to confirm that these desirable asymptotic properties translate to finite samples.

Weighted Micro Data

So far, there are three places where weights can show up in our framework. It is worth clarifying their different roles. First, w_{it} measures the share of *all* consumers in market t who are of type i where the type contains both observed (demographic) and unobserved heterogeneity. The choice of w_{it} should be largely unaffected by the micro data.[38] Second, w_{dijt} is the probability that a consumer in market t of type i who chooses j is selected to be in micro dataset d. Third, although the notation in (11) suggests that micro parts $\bar{v}_p$ are simple averages, many surveys datasets involve weighting schemes in order to better approximate the demographics or choices of the target population.

As an example, the simple average of income among NielsenIQ panelists tends to be higher than the national average. NielsenIQ provides *projection factors* so that after weighting, the demographics of their panelist sample is broadly demographically similar to the entire US population (including incomes). This presents a choice to the researcher: define $\bar{v}_p$ as the simple average of panelist income or as the projection factor-weighted average of panelist income (or to construct custom projection factors that are better suited to one's setting). We expect that in many cases, the latter will be preferred as researchers are often interested in estimating the preferences of an overall population.

As in many surveys, the NielsenIQ projection factors can be interpreted as *inverse sampling weights* $\widetilde{w}_{dijt} \propto 1/w_{dijt}$, which adjust for non-random selection into the micro dataset.

---

[37] Grieco, Murry, Pinkse, and Sagl (2023) call this property of their related estimator "conformant."

[38] In most specifications the researcher will use equally weighted pseudo-random Monte Carlo draws so that $w_{it} = \frac{1}{|\mathcal{I}_t|}$ or quadrature rules over a (multivariate) standard normal distribution.

In this case we could multiply our "micro values" v_{pijt} by $\widetilde{w}_{d_pijt}$ to produce valid estimates of quantities across *all* consumers, not just consumers selected to be in the micro dataset.

For concreteness, consider the running example of minivans and kids. If we assume that the survey was representative, sampling weights should depend only on market size, whether the market is in 2023, and purchasing a car: $w_{dijt} \propto \mathcal{M}_t \cdot 1\{t \in \mathcal{T}_{2023}, j \neq 0\}$. In this case, $\overline{\text{kids}}_{2023}$ from (15) represents an unbiased estimate of the average number of children among households that purchased a car in 2023.

Another possibility is that the survey over-sampled high-income households (perhaps as "likely automobile buyers"), using sampling weights proportional to some known, increasing function of household income: $w_{dijt} \propto g(\text{income}_{it}) \cdot \mathcal{M}_t \cdot 1\{t \in \mathcal{T}_{2023}, j \neq 0\}$. In this case, a simple average $\overline{\text{kids}}_{2023}$ over $v_{1i_nj_nt_n} = \text{kids}_{i_nt_n}$ is biased for its population counterpart. However, if the survey administrator computed inverse sampling weights $\widetilde{w}_{dijt} \propto 1/w_{dijt}$ and instead reported a weighted average with $v_{1ijt} = \widetilde{w}_{dijt} \cdot \text{kids}_{it}$, then $\overline{\text{kids}}_{2023}$ would be unbiased.

When defining the model analogue $f_m(v(\theta))$ of a micro statistic $f_m(\bar{v})$, it is crucially important to know whether and how this statistic has already been weighted. If $f_m(\bar{v})$ has already been adjusted (e.g., with inverse sampling weights) so that it is a valid estimate of some quantity across *all* consumers, then we can drop the sampling probabilities w_{dijt} from the right-hand side of the model analogue in (17).[39] On the other hand, if $\bar{v}_p$ is a simple average over a selected sample, we certainly need to take the sampling probabilities w_{dijt} into account. A simple sanity check is to compare the distribution of each of the demographics under the the demand model (with just the w_{it} weights), to the demographics of the corresponding micro dataset. This comparison is feasible if we condition on demographics i and markets t, but not if we condition on choices j which depend on the unknown parameters θ.

The formula in (17) provides some ambiguity in how we define micro sampling weights w_{dijt} and micro part values v_{pijt}, particularly for conditional expectations. Suppose we were only interested in the average number of children among minivan buyers, $\overline{\text{kids}}_{2023}^{\text{mini}}$. Previously, we represented this with $f_m(\bar{v}) = \overline{v_3}/\overline{v_2}$ where

$$w_{dijt} \propto \mathcal{M}_t \cdot 1\{t \in \mathcal{T}_{2023}, j \neq 0\}, \quad \begin{aligned} v_{2ijt} &= 1\{j \in \mathcal{J}_{\text{mini}}\}, \\ v_{3ijt} &= \text{kids}_{it} \cdot 1\{j \in \mathcal{J}_{\text{mini}}\}. \end{aligned} \tag{19}$$

An alternative would be to instead condition the micro dataset on only minivan buyers, and

---

[39]Using PyBLP, this amounts to setting w_{dijt} equal to some constant.

use only a single micro part $f_m(\bar{v}) = \bar{v}_1$ instead of the ratio:

$$w_{dijt} \propto \mathcal{M}_t \cdot 1\{t \in \mathcal{T}_{2023}, \ j \in \mathcal{J}_{\text{mini}}\}, \quad v_{1ijt} = \text{kids}_{it}. \tag{20}$$

After plugging into (17) and evaluating $f_m(\bar{v})$, both of these will yield the same number: $\overline{\text{kids}}_{2023}^{\text{mini}}$. Though the two approaches contain identical information, we typically discourage the second approach even though it appears simpler.

The main disadvantage is that if, as before, we also wanted to include the average number of children among all car buyers, $\overline{\text{kids}}_{2023}$, these would now be defined over two different micro datasets. In order to correctly calculate weighting matrices and perform inference in Appendix D, we require that each micro dataset be *statistically independent*. This is impossible if one micro dataset is simply a subset of another. A substantive restriction of our general framework for micro moments is that we require some care in how datasets (and corresponding survey weights) are constructed in order to provide correct inference.

5. Standard Micro Moments

The empirical literature has used a variety of different micro moments. In Table 3 we list popular micro moments and the papers from Table 1 that use variants of them.

Demographic Information

Many surveys report information that links purchase behavior to demographic variables. Our running example will be Petrin (2002), which uses summary statistics from a random survey of consumers to help estimate parameters in Π on interactions between consumer demographics and product characteristics.

Petrin (2002) observes the share of consumers in a certain income group $i \in \mathcal{I}_m$ who purchase a new vehicle, and uses this information to incorporate a "$\mathbb{P}(j \neq 0 \mid i \in \mathcal{I}_m)$" moment. We develop this notation-abusing shorthand to refer to a micro moment m that matches $f_m(\bar{v}) = \bar{v}_1/\bar{v}_2$ with micro values $v_{1ijt} = 1\{j \neq 0\} \cdot 1\{i \in \mathcal{I}_m\}$ and $v_{2ijt} = 1\{i \in \mathcal{I}_m\}$.[40] Intuitively, this type of micro moment should help estimate a coefficient in Π that shifts utility for consumers in the income group.

To target a coefficient in Π on the interaction between family size and a minivan dummy, Petrin (2002) could discretize family size y_{rit} into groups of consumers $i \in \mathcal{I}_m$, collect mini-

---

[40]This assumes the underlying dataset d is not selected, $w_{dijt} = 1$. If based on a survey the samples only those in the income group, $w_{dijt} = 1\{i \in \mathcal{J}_m\}$, this shorthand would refer to a micro moment that simply matches the share $f_m(\bar{v}) = \bar{v}_3$ of inside purchases with $v_{3ijt} = 1\{j \neq 0\}$.

vans into a group of products $j \in \mathcal{J}_m$ (e.g., minivans), and incorporate similar "$\mathbb{P}(j \in \mathcal{J}_m \mid i \in \mathcal{I}_m)$" moments. Often, surveys only collect information by broad demographic groups like $\mathcal{I}_m$. However, Petrin (2002) observes the mean family size of those who purchase minivans, and uses this to incorporate a "$\mathbb{E}[y_{rit} \mid j \in \mathcal{J}_m]$" moment, which intuitively contains more information than a single discretized counterpart.

Similarly, surveys that collect data about individual products rather than just broad categories of choices can be even more informative. For a product characteristic x_{cjt} such as price or size that is more granular than $1\{j \in \mathcal{J}_m\}$, matching "$\mathbb{E}[x_{cjt} \mid i \in \mathcal{I}_m,\ j \neq 0]$" could be more useful for estimating a coefficient in Π on the corresponding characteristic.

Even more potentially informative is the covariance "$\mathbb{C}(x_{cjt}, y_{rit} \mid j \neq 0)$" between a product characteristic x_{cjt} and a demographic y_{rit}.[41] Unlike both "$\mathbb{E}[y_{rit} \mid j \in \mathcal{J}_m]$" and "$\mathbb{E}[x_{cjt} \mid i \in \mathcal{I}_m,\ j \neq 0]$," which discretize x_{cjt} and y_{rit} into broad categories, a covariance potentially contains more useful information about a coefficient in Π on the interaction between x_{cjt} and y_{rit}. Although more demanding on the available micro data, there have been a few papers that have matched covariances (see Table 5).[42]

Many useful summary statistics can be written as a function of simple averages. For example, correlations and regression coefficients are covariances scaled by smooth functions of variances. PyBLP supports all such forms of micro moments, requiring only that the user specify the function $f_m(\cdot)$, as well as its derivative for computing objective gradients and delta method-based covariances.

Second Choices

First incorporated in BLP-style estimation by Berry, Levinsohn, and Pakes (2004), "second choices" are a particularly useful form of micro data that requires additional notation. What choices consumers would have made had their first choice been unavailable provides a great deal of information about substitution patterns, and we show how to incorporate this information in our framework below.

Each consumer n in a micro dataset $d \in \mathcal{D}$ with second choices has an additional characteristic k_n. Given a market $t_n = t$ and type $i_n = i$, a consumer chooses $j \in \mathcal{J}_t \cup \{0\}$ first and $k \in \mathcal{J}_t \cup \{0\} \setminus \{j\}$ second with probability $\mathbb{P}_A(j_n = j,\ k_n = k \mid t_n = t,\ i_n = i) = s_{ijkt}$. Idiosyncratic preferences ε_{ijt} remain the same across first and second choices. With ε_{ijt} dis-

---

[41]In a dataset d that already conditions on inside purchase, $w_{dijt} = 1\{j \neq 0\}$, this shorthand refers to a micro moment m that matches $f_m(\bar{v}) = \bar{v}_1 - \bar{v}_2 \cdot \bar{v}_3$ with values $v_{1ijt} = x_{cjt} \cdot y_{rit}$, $v_{2ijt} = x_{cjt}$, and $v_{3ijt} = y_{rit}$.

[42]Nurski and Verboven (2016) match actual covariances, while Berry, Levinsohn, and Pakes (2004) match two moments: "$\mathbb{E}[x_{cjt} \cdot y_{rit} \mid j \neq 0]$" and "$\mathbb{E}[y_{rit} \mid j \neq 0]$." Since "$\mathbb{E}[x_{cjt} \mid j \neq 0]$" is equal to a fixed constant, these two moments span the single covariance.

tributed type I extreme value, the probability of the joint event can be written in a familiar form, $s_{ijkt} = s_{ijt} \cdot s_{ik(-j)t}$ where $s_{ik(-j)t} = s_{ikt}/(1 - s_{ijt})$ is the probability of choosing k when j is eliminated from the choice set.[43] In practice, we derive and use a less intuitive but more general expression $s_{ijkt} = s_{ik(-j)t} - s_{ikt}$, which also works for the previously-mentioned nested logit variant discussed in Appendix B.[44]

The survey sampling probability $\mathbb{P}_A(n \in \mathcal{N}_d \mid t_n = t, i_n = i, j_n = j, k_n = k) = w_{dijkt}$ can also depend on second choices. For example, $w_{dijkt} \propto \mathcal{M}_t \cdot 1\{j, k \neq 0\}$ would generate a random sample of consumers whose first and second choices were both inside alternatives.

Each micro moment part p based on a micro dataset d_p with second choices has micro values v_{pijkt} that can depend on second choices. For example, if $\bar{v}_p$ is the share of participants in a survey whose second choice was in some set $\mathcal{K}_p$ (e.g., Ford vehicles or light trucks), its micro values are $v_{pijkt} = 1\{k \in \mathcal{K}_p\}$. The conditional expectation of micro values based on second choices is

$$v_p(\theta) = \frac{\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t \cup \{0\}} \sum_{k \in \mathcal{J}_t \cup \{0\} \setminus \{j\}} w_{it} \cdot s_{ijkt}(\theta) \cdot w_{d_p ijkt} \cdot v_{pijkt}}{\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t \cup \{0\}} \sum_{k \in \mathcal{J}_t \cup \{0\} \setminus \{j\}} w_{it} \cdot s_{ijkt}(\theta) \cdot w_{d_p ijkt}}. \tag{21}$$

It is conceptually straightforward to incorporate third or fourth choices by adding more subscripts and sums. We limit our attention to second choices because additional sums severely increase computational cost and required notation.[45]

In papers such as Berry, Levinsohn, and Pakes (2004) that use second choice data, a popular statistic is the covariance "$\mathbb{C}(x_{cjt}, x_{ek(-j)t} \mid j, k \neq 0)$" between first and second choice characteristics x_{cijt} and x_{ekt}.[46] Intuitively, this should contain information about a parameter in Σ that measures the variance of unobserved preference heterogeneity ν_{cit} for x_{cijt} if $e = c$, or the covariance between unobserved preferences ν_{cit} and ν_{eit} for x_{cijt} and x_{eijt} if $e \neq c$. Holding mean preferences δ_{jt} equal, if when j is eliminated from the choice set consumers tend to select a second choice k that has a very similar characteristic $x_{ckt} \approx x_{cjt}$, it must be

---

[43]For more details see Conlon and Mortimer (2021) and the "individual diversion ratio". The expression $D_{j \to k, i} = s_{ik(-j)t} = s_{ikt}/(1 - s_{ijt})$ works for type I extreme value ε_{ijt}, but for other distributions such as that used by the nested logit model in Appendix B, $s_{ik(-j)t}$ can be computed numerically by removing j from the choice set and computing the probability of choosing k.

[44]That is, $\mathbb{P}_\varepsilon(u_{ijt} > u_{ikt} > u_{i\ell t}, \forall \ell \neq j, k) = \mathbb{P}_\varepsilon(u_{ikt} > u_{i\ell t}, \forall \ell \neq j, k) - \mathbb{P}_\varepsilon(u_{ikt} > u_{i\ell t}, \forall \ell \neq k)$. The second term is simply s_{ikt}. The first term can be equivalently expressed as $\lim_{\delta_{jt} \to -\infty} s_{ikt}$, which equals $s_{ik(-j)t}$ for both the simple and nested logit models.

[45]With longer lists of ranked choice data, researchers often consider full maximum likelihood type approaches rather than aggregated moments (see, e.g., Agarwal and Somaini, 2020).

[46]In practice, Berry, Levinsohn, and Pakes (2004) split this covariance up and match two moments, "$\mathbb{E}[x_{cjt} \cdot x_{ek(-j)t} \mid j, k \neq 0]$" and "$\mathbb{E}[x_{ek(-j)t} \mid j, k \neq 0]$," to work with simple averages.

that ν_{cit} has a high variance. Otherwise, we would expect to see proportionate substitution to all remaining alternatives.

Relatively complete data on consumers' first and second choices is becoming more common in empirical research. In these cases, researchers may have survey data which measures $\mathbb{P}_A(j_n = j,\ k_n = k \mid n \in \mathcal{N}_d)$ directly. That is, they may observe first and second choices in aggregate, but not necessarily the corresponding demographic information for the consumers. For example, Grieco et al. (2021) have survey data from Maritz that surveys new car purchasers both on which car they purchased and what model they would purchase if their choice were unavailable. Conlon et al. (2023) use the same survey data, and show that it is possible to provide semi-parametric (mixed logit) estimates of utilities using only first and second choices from a single market. In our empirical example in Section 8, we demonstrate how to collect simple second choice data from an online survey. In the UK, a typical survey question asked by the Competition and Markets Authority (CMA) to evaluate a potential merger is "where would you have made your purchases today if this store were closed for six months?" (Reynolds and Walters, 2008).

As mentioned above, a common data constraint is that many surveys may not collect information about individual products or product characteristics, but only for groups of products. Conceptually, it is straightforward to incorporate information on how many consumers would substitute to another minivan or pickup truck without specifying the brand: '$\mathbb{P}(k(\text{-}b(j)) \in \mathcal{K}_m \mid j \in \mathcal{J}_m)$" with $v_{mijkt} = 1\{k \in \mathcal{K}_m\}$ and $w_{dijkt} = 1\{j \in \mathcal{J}_m\}$. These kinds of information might be especially useful if the goal is to estimate a random coefficient on a dummy for "pickup truck" or "minivan."

One extension available in PyBLP is elimination not only of the exact first choice j, but a group of products $h(j)$ containing j.[47] This extension is particularly useful because often, second choice data will be at a higher level of aggregation than products. For example, the researcher may have access to information about where consumers may substitute when their favorite brand $h(j) = b(j)$ is eliminated from the market (e.g., Coca Cola), which includes their first choice product j (e.g., a 2-liter bottle of Coca Cola). Alternatively, we might observe how consumers substitute when all hospitals from the Partners system were eliminated from the choice set.

---

[47]The only real difference is that we compute $s_{ik(\text{-}h(j))t}$ instead of $s_{ik(\text{-}j)t}$. With ε_{ijt} distributed type I extreme value, we can write $s_{ik(\text{-}h(j))t} = s_{ikt}/(1 - \sum_{h(\ell)=h(j)} s_{i\ell t})$.

Compatibility Issues

An important challenge with combining aggregate and micro data is compatibility. It is rare to find two datasets that are perfectly compatible. Variables may be measured or defined slightly differently, data may be collected at different frequencies or during different periods, and survey data may oversample individuals in unexpected ways.

A frequent source of incompatibilities arises when the distribution of characteristics x_{jt}, demographics y_{it}, or choices s_{jt} differs significantly between the aggregate purchase data and the micro survey data. This could arise because the income of shoppers in a survey differs from the income of shoppers at a particular store, or if surveyed consumers face a different set of products (or characteristics such as prices) than those in the aggregate data. It could also arise because of bad luck or poor survey design. For example, nationally, Coca-Cola has around a 48% market share, while Pepsi has around a 20% market share. If we surveyed individuals about their soft drink preferences (as we do in Section 8) and found that more consumers preferred Pepsi to Coca-Cola, this would present a potential incompatibility with the aggregate sales data.

One likely violation of compatibility that is likely to arise in practice is that many papers match micro moments averaged over the entire sample, rather than a subset of markets. An example of correctly addressing compatibility can be found in Grieco et al. (2021) where the authors observe aggregate purchase data from 1980 to 2018, as well as individual survey data from the years 1991, 1999, 2005, 2015. Because the distribution of prices and characteristics are quite different in 1991 and 2015, it is important to condition on the year when constructing micro datasets so that $w_{1991,ijt} = \mathcal{M}_t \cdot \{j \neq 0, t \in \mathcal{T}_{1991}\}$ and $w_{2015,ijt} = \mathcal{M}_t \cdot \{j \neq 0, t \in \mathcal{T}_{2015}\}$, rather than averaging over all years.

As another example, Backus et al. (2021) compute both the chain-year specific joint distribution of characteristics (income and presence of children) when forming w_{it} and calculate separate micro moments for each chain-year. This results in a very large number of micro moments, but guarantees compatibility in the sense that this correctly matches individual shoppers to the correct product assortment and prices. By conditioning on chain, this avoids the possibility that the NielsenIQ panelists systematically shop at a different set of supermarket chains than predicted by the aggregate sales patterns $\mathcal{M}_t$. This issue will arise frequently with the NielsenIQ data, where not all supermarket chains report scanner data sales, but panelists report purchases at all stores whether or not they are in the scanner dataset.

Another example of where compatibility of micro data presents a challenge can be found

in Conlon and Rao (2023). A well-known problem with survey data on alcohol consumption is that reported per capita consumption reflects only 30-40% of alcohol purchases. While it might be tempting to construct moments to match the probability of purchasing a unit of alcohol conditional on income within some range, "$\mathbb{P}(j \neq 0 \mid y_{it} \in [\underline{y}_a, \bar{y}_a])$," these moments are incompatible with aggregate sales data. That is, there is no set of parameters θ such that one could match both the aggregate no purchase share s_{0t} and the purchase or no-purchase shares by income. One approach would be to not include micro moments from an incompatible survey, but the other is to define compatible moments that are potentially less efficient. The authors apply Bayes Rule and match the probability that a given unit of alcohol is purchased by households of each income level, "$\mathbb{P}(y_{it} \in [\underline{y}_a, \bar{y}_a] \mid j \neq 0)$." This avoids the issue that the marginal distribution of purchasing alcohol "$\mathbb{P}(j \neq 0)$" is completely different across the aggregate and micro datasets, while still including information on the relationship between alcohol purchases and income.[48]

In general, researchers may wish to match summary statistics that are compatible with their full dataset, rather than use all the information in a dataset that they suspect is less compatible. In our Monte Carlo experiments in Section 7, we provide a typical example where income is measured differently across datasets, but being careful about what information to match can still deliver an unbiased estimator.

In other settings, researchers may be unsure about whether there are compatibility issues. The good news is that testing for compatibility is straightforward (Imbens and Lancaster, 1994). If there is sufficient variation from the aggregate data (or from micro datasets that are known to be compatible) to identify the model, we can use an overidentification test. Our preferred approach is to estimate the model without the potentially incompatible micro moments to obtain $\hat{\theta}$, and then form a test statistic from differences $\hat{\Delta}_{\mathrm{M}} = f(\bar{v}) - f(v(\hat{\theta}))$ between observed and estimated micro statistics,[49]

$$\mathrm{Wald} = N_{\mathrm{A}} \hat{\Delta}'_{\mathrm{M}} \hat{S}_{\mathrm{M}}^{-1} \hat{\Delta}_{\mathrm{M}} \rightsquigarrow \chi^2(M_{\mathrm{M}}) \tag{22}$$

where $\hat{S}_{\mathrm{M}}^{-1}$ is the properly-scaled asymptotic covariance matrix of the micro moments that we derive in Appendix D, and which is automatically reported by PyBLP for standard error calculations.

---

[48]This of course relies on the assumption that higher or lower income individuals aren't systematically under-reporting relative to other income groups.

[49]If there are some compatible micro moments, these can be used to obtain $\hat{\theta}$, should have their elements in $\hat{\Delta}_{\mathrm{M}}$ set to zero, and their count should be subtracted from the χ^2 degrees of freedom.

6. Optimal Micro Moments

Having built intuition for common forms of micro moments, we discuss optimality. How should we choose what statistics to match, given data availability, computational resources, compatibility, and interpretability requirements?

Matching Scores

In terms of data availability, a best-case scenario is observing not just a few micro statistics $\bar{v}_m$, but rather a complete dataset of all sampled consumers' markets, choices, and demographics $\{(t_n, j_n, y_{i_n t_n})\}_{n \in \mathcal{N}_d}$. When we say "complete" we must observe not only individual choices, but also all of the relevant demographics required to compute the choice probabilities in (4). Rather than use the standard micro moments from Section 5, we can use the scores from the individual data likelihood and combine them with the aggregate moments from aggregate estimator in (6). This has the advantage that it will efficiently use all of the information in the micro dataset, and also that it may reduce the overall number of micro moments used in estimation. The disadvantage is that the individual scores are infeasible and require an initial estimate $\hat{\theta}$ in order to compute them.

Our preferred approach is a two-step procedure that minimizes computational costs while still making full use of the micro data.[50] After obtaining a first-stage estimator $\hat{\theta}$ with sub-optimal micro moments, the researcher constructs optimal micro moments m that match the average score function $f_m(\bar{v}) = \bar{v}_m$ for each micro dataset d_m evaluated at each nonlinear parameter θ_m:[51]

$$v_{mijt}(\hat{\theta}) = \frac{\partial \log \mathbb{P}_A^{\hat{\theta}}(t_n = t, \, j_n = j, \, y_{i_n t_n} = y_{it} \mid n \in \mathcal{N}_{d_m})}{\partial \theta_m}. \tag{23}$$

In words, the score tells us how the log choice probability varies with parameters θ for an individual with demographics y_{it} in market t. Conveniently, we only need to calculate this for the option j that an individual chooses.

In Appendix F we provide full expressions for micro data scores and demonstrate how to compute them with PyBLP. Incorporating second choices into our procedure is a simple matter of adding additional subscripts and a more complicated score expression, so here we focus on first choices.

---

[50]Our approach is closely related to the unfortunately-named "one-step" method discussed, for example, in Section 3.4 of Newey and McFadden (1994).

[51]This can be done in PyBLP with only a few lines of code. See Figure F1 in Appendix F.

First, we compute the score $v_{m_i n_j n_t n}(\hat{\theta})$ in (23) evaluated at each observation $n \in \mathcal{N}_d$ and take their average over individuals, choices, and markets to get $\bar{v}_m(\hat{\theta})$. We also pre-compute $v_{mijt}(\hat{\theta})$ for each possible (i, j, t) so that scores only need to be computed a single time. After also constructing an estimator of the optimal weighting matrix and, if desired, an approximation to Chamberlain's (1987) optimal instruments, the researcher obtains the second-stage estimator.

Algorithm 2 Optimal Micro BLP Estimator

Given a sense for reasonable bounds for the nonlinear parameters (Π, Σ), for example from running a version of the IV regression in (8) and Appendix C:

1. Use sub-optimal micro moments to obtain a first-stage estimator $\hat{\theta}$ by minimizing the objective constructed by Algorithm 1. We recommend drawing a few different starting values from within reasonable parameter bounds. In Conlon and Gortmaker (2020) we describe and evaluate other best practices for nonlinear optimization in Sections 3 and 5.

2. Approximate Chamberlain's (1987) optimal instruments $\hat{z}_{jt}(\hat{\theta})$ by following Algorithm 2 in Conlon and Gortmaker (2020), originally proposed by Berry, Levinsohn, and Pakes (1999).

3. Approximate the optimal micro moment m for each dataset d_m and nonlinear parameter p_m pair by computing $\bar{v}_m(\hat{\theta}) = \frac{1}{N_{d_m}} \sum_{n \in \mathcal{N}_{d_m}} v_{m_i n_j n_t n}(\hat{\theta})$ in (23).

4. Estimate the optimal weighting matrix $\hat{W}(\hat{\theta})$ by inverting an estimator of the asymptotic covariance matrix of the moments in Appendix D.

5. Use approximations to the optimal IVs, micro moments, and weighting matrix to obtain the second-stage estimator by minimizing the objective constructed by Algorithm 1. Again, we recommend drawing a few different starting values from within reasonable parameter bounds.

In Appendix E we show that if the first-stage estimator is consistent, then the second-stage estimator is asymptotically efficient within the class of all possible micro BLP estimators. By this, we mean that Algorithm 2 delivers an estimator with an asymptotic variance that is no greater than that of another micro BLP estimator based on any weighting matrix $\hat{W}$, instruments z_{jt}, micro moment functions $f_m(\cdot)$, and micro values v_{pijt}. Restricting ourselves to this class of micro BLP estimators rules out efficiency gains from estimators outside this class, such as those that do not require that observed market shares $\mathcal{S}_{jt}$ exactly equal their model counterparts.[52]

Only needing to compute scores once makes our two-step approach particularly computationally efficient. The more familiar approach of stacking scores with the original moments

---

[52]For example, Grieco, Murry, Pinkse, and Sagl's (2023) CLER estimator obtains efficiency gains by relaxing the share constraint, particularly when the number of micro observations is a nontrivial proportion of the observations underlying aggregate market shares.

(e.g., in Imbens and Lancaster, 1994) would require re-computing all observations' scores for each optimization iteration over θ.[53] This is also the case for Grieco, Murry, Pinkse, and Sagl's (2023) approach of subtracting a properly-scaled log-likelihood from the original objective.[54]

Grieco et al. (2023) point to a potential downside of using scores instead of their approach of using the log-likelihood itself: population scores may have multiple zeros even when the population likelihood has a unique maximum.[55] Although the aggregate moments should help to smooth out such spurious local minima, this downside does highlight the importance of drawing a few different starting values when performing nonlinear optimization.

A final concern is with inconsistent first-stage estimates. In practice, we recommend using standard micro moments discussed in the last section, which should typically provide consistent and credible parameter estimates for the first stage. If standard micro moments in conjunction with aggregate variation seem to only weakly identify or not identify some parameters, another option is to also match scores in the first stage, but evaluated at an informed guess of the true θ_0 rather than a consistent estimate, which in some cases may be more informative about θ than standard micro moments.[56]

Intuition from Scores

Often, instead of having the full results from a survey, researchers will only have access to or be willing to use summary statistics because of cost, interpretability, compatibility, confidentiality, or other data limitations. For estimating a given model, the most efficient summary statistic would be the score, averaged across all surveyed individuals. Although survey administrators are unlikely to collect scores for different models, inspecting the functional form of scores for some simple models does motivate the functional form of some of the common micro moments discussed in Section 5.

We present full score expressions in Appendix F but here consider the simplest case with

---

[53]Technically, one only need compute scores for each distinct set of demographic values, product choice, and market. This speeds up computation when demographics take on only a few discrete values and purchases are not spread across many products.

[54]Grieco et al. (2023) emphasize that this approach makes the demand-only objective convex in each δ_{jt}, which makes nonlinear optimization over all N_A values of δ_{jt} computationally feasible.

[55]For example, when Grieco et al. (2023) treat δ_{jt} as a parameter, the score for Σ is zero both at the true Σ_0 and $\Sigma = 0$. This is not the case here because the micro BLP score also depends on $\frac{\partial \delta_{jt}}{\partial \theta'}$, but it does suggest that such a concern is more than a theoretical edge-case.

[56]For example, limited cross-market choice set variation and standard micro moments that do not use second choices may result in a poorly-identified Σ. Using more information in the full micro dataset may help provide a consistent first-stage estimator.

$C = 1$ observed characteristic, $R = 1$ demographic, three parameters $\theta = (\beta, \pi, \sigma)$, and a micro dataset d with no selection, $w_{dijt} = 1$. First consider the case without any unobserved heterogeneity, $\sigma = 0$. The score for β is zero,[57] and for π is

$$\frac{\partial \log \mathbb{P}_A(t_n = t, \, j_n = j, \, y_{i_n t_n} = y_{it} \mid n \in \mathcal{N}_d)}{\partial \pi} = \frac{\partial u_{ijt}}{\partial \pi} - \sum_{k \in \mathcal{J}_t} s_{ikt} \cdot \frac{\partial u_{ikt}}{\partial \pi}. \tag{24}$$

in which the derivative of indirect utility for $j \neq 0$ in (1) with respect to π is

$$\frac{\partial u_{ijt}}{\partial \pi} = \frac{\partial \mu_{ijt}}{\partial \pi} + \frac{\partial \delta_{jt}}{\partial \pi} = x_{jt} \cdot y_{it} + \frac{\partial \delta_{jt}}{\partial \pi}. \tag{25}$$

Since s_{ijt} and $\frac{\partial \delta_{jt}}{\partial \pi}$ are functions of π,[58] the only term directly observed in the micro data is $x_{jt} \cdot y_{it}$. This suggests that the "$\mathbb{C}(x_{jt}, y_{it} \mid j \neq 0)$" moment discussed above should be very informative about π because it is similar to the score.[59] If $x_{jt} = 1$, then the primary term is simply y_{it}, suggesting that "$\mathbb{E}[y_{it} \mid j \neq 0]$" should be informative about π in this simpler case. In Section 7 we confirm this intuition with Monte Carlo experiments.

Often, demographics will be discrete (e.g., levels of education, presence of children, or binned income). For example, Petrin (2002) matches "$\mathbb{E}[x_{jt} \mid y_{it} = 1] = \mathbb{P}(j \neq 0 \mid y_{it} = 1)$" where $x_{jt} = 1$ is an indicator for all inside goods and y_{it} is an indicator for high income consumers. Intuition about informativeness is similar in this case. Up to a denominator "$\mathbb{P}(y_{it} = 1)$," which is a constant scaling factor that only depends on demographic data, matching this moment is identical to matching a "$\mathbb{E}[x_{jt} \cdot y_{it}]$" moment, which is very similar to the score.

However, matching only a single covariance or expectation leaves some information on the table because it does not span the subsequent terms in the score. Similarly, for the case with $\sigma \neq 0$, the score for π becomes an integral over unobserved heterogeneity, further distancing a single "$\mathbb{C}(x_{jt}, y_{it} \mid j \neq 0)$" moment from the true score.

To focus on the value of second choices, next consider the case with observed heterogene-

---

[57]Micro data are uninformative about β because it enters into choice probabilities s_{ijt} only through mean utilities δ_{jt}, which are pinned down by the aggregate data share constraint in (7).

[58]Since mean utilities are pinned down by the share constraint in (7), their derivatives are given by invoking the implicit function theorem: $\frac{\partial \delta_t}{\partial \pi} = \left(\frac{\partial s_t}{\partial \delta_t} \right)^{-1} \frac{\partial s_t}{\partial \pi}$.

[59]Grieco et al. (2023) also note the similarity of "$\mathbb{C}(x_{jt}, y_{it} \mid j \neq 0)$" moments to the score for π. Their expression does not involve $\frac{\partial \delta_{jt}}{\partial \pi}$ because their estimator treats each δ_{jt} as a separate parameter rather than as an implicit function of π.

ity, $\sigma \neq 0$, but without any observed demographics, $\pi = 0$. The score for σ is

$$
\begin{aligned}
&\frac{\partial \log \mathbb{P}_A(t_n = t,\ j_n = j,\ k_n = k \mid n \in \mathcal{N}_d)}{\partial \sigma} \\
&= \sum_{i \in \mathcal{I}_t} \frac{w_{it} \cdot s_{ijkt}}{\sum_{\iota \in \mathcal{I}_t} w_{\iota t} \cdot s_{\iota jkt}} \left[\frac{\partial u_{ijt}}{\partial \sigma} + \frac{\partial u_{ikt}}{\partial \sigma} - \sum_{\ell \in \mathcal{J}_t} s_{i\ell t} \cdot \frac{\partial u_{i\ell t}}{\partial \sigma} - \sum_{\ell \in \mathcal{J}_t \setminus \{j\}} s_{i\ell(-j)t} \cdot \frac{\partial u_{i\ell t}}{\partial \sigma} \right],
\end{aligned}
\tag{26}
$$

in which the derivative of indirect utilities for $j, k \neq 0$ with respect to σ is

$$
\frac{\partial u_{ijt}}{\partial \sigma} + \frac{\partial u_{ikt}}{\partial \sigma} = \nu_{it} \cdot (x_{jt} + x_{kt}) + \frac{\partial \delta_{jt}}{\partial \sigma} + \frac{\partial \delta_{kt}}{\partial \sigma}.
\tag{27}
$$

The only term directly observed in the micro data is $(x_{jt} + x_{kt})$. This is scaled by the average unobserved preference ν_{it} among those who choose j first and k second, but the sum itself is similar to the "$\mathbb{C}(x_{jt}, x_{k(-j)t}, \mid j, k \neq 0)$" moment discussed above, suggesting that such a second choice covariance should indeed be very informative about σ. And if available, the average or sum "$\mathbb{E}[x_{jt} + x_{k(-j)t} \mid j, k \neq 0]$" of first- and second-choice characteristics could be even more informative. We confirm this intuition in Section 7. Of course, only matching a covariance or sum will not fully match the expression in (26), which involves even more terms after adding in observed demographics (see Appendix F).

Inspecting scores in this way can provide some intuition for which micro moments may be particularly informative. We provide additional examples for extensions with lognormal random coefficients and nesting parameters in Appendices A and B. In general, however, the best summary statistics to match will depend on the model specification and the true parameter values.

In Appendix G we provide a more formal approach for determining which summary statistics are most informative about the parameters in the model. Given a first-stage estimator $\hat{\theta}$ and a sampling scheme w_{dijt}, we recommend simulating a micro dataset and regressing simulated scores on candidate micro values,[60] keeping only those sets of micro values that maximize the R^2 of the regression. This type of procedure identifies a small number of maximally informative summary statistics that are more likely to be collected by survey administrators than model-specific average scores.

---

[60]In Appendix G we also demonstrate how this can be done in only a few lines of code with PyBLP.

7. Monte Carlo Experiments

We provide several Monte Carlo experiments to illustrate the performance of the micro BLP estimator with different micro moments. We also use our simulations to illustrate the importance of practical choices that need to be made when doing empirical research, which we will further expand upon in the empirical examples of Section 8.

Monte Carlo Configuration

Our simulation configurations build on those of Conlon and Gortmaker (2020), which are loosely based on those of Armstrong (2016). We first describe a baseline configuration, and in the following subsections describe how we modify this configuration to compare different aspects of the micro BLP estimator.

For each configuration, we construct and estimate the model on 1,000 different synthetic datasets. In each of $T = |\mathcal{T}| = 40$ markets we randomly choose either 2, 5, or 10 firms, and have each firm produce 3, 5, or 5 products in that market. The number of products is generally between $10 < |\mathcal{J}_t| < 30$. Across markets, the number of aggregate observations is generally between $400 < N_A < 1,200$.

There are $C = 3$ observed product characteristics $x_{jt} = (1, x_{2jt}, p_{jt})'$: a constant, an exogenous characteristic $x_{2jt} \sim U(2, 4)$, and endogenous prices p_{jt}. We generate a realistic correlation between p_{jt} and unobserved quality ξ_{jt} by drawing ξ_{jt} and cost shocks from a mean-zero bivariate normal distribution, by drawing a cost shifter, and by numerically solving for Bertrand-Nash equilibrium prices p_{jt} and shares s_{jt} with the fixed point approach of Morrow and Skerlos (2011).[61] Since our focus is not on weak cost shifters, our marginal cost parameterization generates a strong correlation between the cost shifter and price. Instruments z_{jt} are $(1, x_{2jt})'$, the cost shifter, and the differentiation IVs of Gandhi and Houde (2020) discussed in Section 3.[62] We parameterize mean utility in (2) to give "realistic" outside shares generally between $0.6 < \mathcal{S}_{0t} < 0.9$:

$$\delta_{jt} = \beta_1 + \beta_x x_{2jt} + \alpha p_{jt} + \xi_{jt}, \quad \beta_0 = (\beta_{01}, \beta_{0x}, \alpha_0)' = (-6, 3, -3)'. \tag{28}$$

In each market t, we generate different Monte Carlo draws to represent $|\mathcal{I}_t| = 1,000$

---

[61]Firms choose prices to maximize their products' profits $s_{jt}(p_t) \cdot (p_{jt} - c_{jt})$ subject to marginal costs $c_{jt} = 2 + 0.1 \times x_{2jt} + 1.0 \times w_{jt} + \omega_{jt}$. The cost shifter is distributed $w_{jt} \sim U(0, 1)$. Unobserved quality ξ_{jt} and the cost shock ω_{jt} are mean-zero bivariate normal with common variance 0.2 and covariance 0.1.

[62]As noted in Footnote 23, we use the "quadratic" version of differentiation IVs: $\hat{a}_{2jt} = \sum_k (x_{2jt} - x_{2kt})^2$ both alone (when we include unobserved heterogeneity) and interacted with the mean $m_t^y = \sum_i w_{it} \cdot y_{it}$ of a consumer demographic y_{it}, discussed shortly.

consumer types, each with an equal share $w_{it} = 1/|\mathcal{I}_t|$. Since income is the most common demographic to appear in demand systems, we randomly assign each market to a US state and draw $R = 1$ demographic y_{it} from a lognormal distribution fit to the 2019 American Community Survey (ACS) income distribution for that state. To start, we do not include unobserved heterogeneity when parameterizing heterogeneous utility in (3):

$$\mu_{ijt} = \pi_1 y_{it} + \pi_x x_{2jt} y_{it}, \quad \Pi_0 = (\pi_{01}, \pi_{0x}, 0)' = (-0.1, 0.1, 0)', \quad \Sigma_0 = 0. \tag{29}$$

Finally, we simulate a micro dataset d with an average of 1,000 observations per market. Since the most common type of consumer survey samples only those who select an inside alternative, we use selection probabilities $w_{dijt} = 1\{j \neq 0\}$.

To obtain an estimator $\hat{\theta} = (\hat{\beta}_1, \hat{\beta}_x, \hat{\alpha}, \hat{\pi}_1, \hat{\pi}_x)$ we follow the recipe in Algorithm 2 for the optimal micro BLP estimator, but to start we do not approximate the optimal micro moments. To solve the fixed point for $\hat{\delta}_{jt}(\Pi, \Sigma)$ and optimize over θ, we use best practices described in Conlon and Gortmaker (2020).[63] To numerically integrate over the distribution of income y_{it}, we resample 1,000 times from its true distribution.

Monte Carlo Results

When reporting results from our simulations, we focus on the median absolute error (MAE) and median bias of the parameter estimators. In Appendices H and I we provide additional results measuring the performance of standard error counterfactual calculations, which are generally in line with the performance of parameter estimators across configurations. Computation was done on the Harvard Business School compute cluster.[64]

Demographic Variation

In Table 4 we vary the amount of cross-market demographic variation and measure the performance of the aggregate BLP estimator. When in each market income y_{it} is drawn from a lognormal distribution fit to the same national distribution of income, there is no cross-market variation, so as discussed in Section 3, π_1 and π_x are not identified.

In the second row, randomly assigning each market to one of the 50 US states provides

---

[63]We accelerate the fixed point with the SQUAREM method of Varadhan and Roland (2008) and use an L^∞ tolerance of `1E-14`. To optimize, we supply objectives and analytic gradients to SciPy's trust region algorithm "trust-constr" and use an L^∞ gradient-based tolerance of `1E-5`. For each GMM step, we draw three sets of starting values from 100% above and below the true parameter values.

[64]For our configurations, six rounds of optimization (three sets of starting values for each GMM step) typically take 1–3 minutes, plus another 30 seconds for computing optimal micro moments. Using second choice moments typically takes 3–8 times longer.

some cross-market variation, which gives an estimator with very little finite sample bias. However, income distributions do not vary much across states, so the estimator still has high variance, even when using optimal instruments.[65] Assigning markets instead to the 982 Public Use Microdata Areas (PUMAs) increases the amount of cross-market income variation, further reducing the bias and variance of $\hat{\pi}_1$ and $\hat{\pi}_x$.

In the last three rows, we double the number of markets to $T = 80$ but keep the amount of cross-market demographic variation the same by re-using the demographic distribution in each $t \leq 40$ for market $t + 40$. As the amount of cross-market choice set variation increases, bias and variance of $\hat{\pi}_1$ and $\hat{\pi}_x$ decrease. In line with the linear regression intuition from Section 3, more variation in demand helps estimate Π, which is identified by how cross-market demographic variation shifts demand. However, without a great deal of demographic variation, the estimator is still fairly noisy.

Since we made the cost shifter a strong instrument and did not model preference heterogeneity for price (see Appendix A for simulation results for when we do), the coefficient on price $\hat{\alpha}$ has very little bias and variance across all configurations. The performance of the linear parameter estimators $\hat{\beta}_1$ and $\hat{\beta}_x$ track the performance of the nonlinear estimators $\hat{\pi}_1$ and $\hat{\pi}_x$, so for simplicity's sake, we focus only on estimators of nonlinear parameters in subsequent results.

Standard Micro Moments

Sticking with $T = 40$ markets and state-level income variation, in Table 5 we illustrate the impact of standard micro moments discussed in Section 5. Matching only the mean income of those who do not choose the outside alternative with a "$\mathbb{E}[y_{it} \mid j \neq 0]$" moment somewhat reduces variance, but not by much. A "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" moment contains more information and reduces variance a bit more, particularly for the π_x parameter whose score it approximates. However, it is only with the combination of both moments that we greatly reduce the variance of both estimators.

Since "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" equals "$\mathbb{E}[x_{2jt} \cdot y_{it} \mid j \neq 0] + \mathbb{E}[x_{j2t} \mid j \neq 0] \cdot \mathbb{E}[y_{it} \mid j \neq 0]$," when paired with a "$\mathbb{E}[y_{it} \mid j \neq 0]$" moment it contains essentially the same information as matching the first term in the score for π_x, the interaction "$\mathbb{E}[x_{2jt} \cdot y_{it} \mid j \neq 0]$." Accordingly, both perform almost identically. Of course, matching a covariance can be more appealing because it is more interpretable and is more likely to be reported by a survey.

A survey that does not report covariances may still report average characteristics by

---

[65]Optimal instruments are well-known to reduce the bias and variance of the aggregate BLP estimator (Reynaert and Verboven, 2014; Conlon and Gortmaker, 2020).

demographic groups, allowing us to use a "$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$" moment that matches the mean x_{2jt} for low-income consumers. Discretizing y_{it} discards some information, reducing correlation with the score for π_x, so the estimator has a higher variance. Since in this simple simulation the score for π_x is dominated by $x_{2jt} \cdot y_{it}$, adding the discretized moment on top of the continuous one does not particularly improve the performance of the estimator.[66]

To visualize the relationship between "$\mathbb{E}[x_{2jt} \cdot y_{it} \mid j \neq 0]$," "$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$," and the score, for each observation in the micro data underlying Table 5 we compute $x_{2jt} \cdot y_{it}$, $x_{2jt} \cdot 1\{y_{it} < \bar{y}_t\}$, and the score for π_x. We report their correlation matrix in Figure 1.[67] As expected, $x_{2jt} \cdot y_{it}$ and the score have strong correlation of 0.675.[68] Discretizing y_{it} reduces the correlation with the score by around 11% to 0.6.

This same approach can be used as a diagnostic: researchers can use the score contributions of simulated individuals under the model at the estimated parameters $\hat{\theta}$, and compare these to their micro statistics (see Appendix G). While this requires an estimate of $\hat{\theta}$, it provides a simple way to measure whether the micro statistics do a good job capturing the potential micro-level variation.

Optimal Micro Moments and Compatibility

In Table 6 we illustrate the performance of optimal micro moments. The first row is the same as the fourth row in Table 5. In the second row, we use these same standard moments to obtain a first-stage estimator, and in the second GMM step, use optimal micro moments that match scores of π_1 and π_x. This requires using the full micro dataset rather than two summary statistics, but it does, unsurprisingly, decrease the variance of the estimator. In the middle two rows, we use the slightly less-informative micro moment "$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$" in the sixth row of Table 5. When using this as a first-stage estimator, the finite sample performance of the optimal micro moments is slightly worse, although not by much.

In the last two rows, we illustrate an example where the "optimal micro moments" can perform worse than matching simple summary statistics. We simulate a second, independent micro dataset that is configured the same as the first, except we replace income y_{it} with a

---

[66] In more complicated simulations, for example with unobserved heterogeneity, adding additional moments can help explain variation in the more complicated score.

[67] The score is evaluated at the true θ_0. We report the absolute value of correlations, taking a median across the 1,000 simulated micro datasets.

[68] How to interpret this number? With a single parameter and a single linear micro moment, the asymptotic standard deviation (SD) of the efficient GMM estimator is one over this correlation times the score's SD (see Appendix G). Since the Normal distribution's MAD is proportional to its SD, this correlation should hence equal the ratio of MAEs obtained under the optimal moment versus the sub-optimal moment. Even though we are not in the scalar case, we see approximately this result in Table 6.

censored version $\tilde{y}_{it}$, an indicator for whether an individual is above or below the median income $\bar{y}_t$. These new micro data $(t_n, j_n, \tilde{y}_{i_n t_n})$ are not "complete" in the sense that they do not contain all of the information necessary to compute the individual choice probabilities (which require the actual income y_{it}). To approximate what a researcher might do here, when computing the scores we replace $\tilde{y}_{it}$ with the 25th percentile of income if below the median or the 75th percentile if above. As we see in the last row of Table 6, the optimal micro moments from the incompatible micro dataset perform significantly worse than no micro moments at all, particularly for $\hat{\pi}_1$. Adding micro statistics of the form "$\mathbb{E}[x_{2jt} \mid \tilde{y}_{it} < \bar{y}_t, j \neq 0]$" contains relevant information and does not have the same compatibility problems, giving similar improvements as before.

While we focus on changing the set of moments to address the compatibility problem, an alternative would be to modify the model to match the observed moments. One option might be to consider two sets of coefficients (π_h, π_l), for high- and low-income individuals. This would eliminate the compatibility problem and allow us to use the scores.[69]

Pooling Markets

Often, a researcher may have the same type of micro statistic for different markets. A practical question is whether one should pool these into a single micro moment,[70] or match a separate micro moment for each market. Computationally, pooling is not particularly important, since micro values will still need to be computed in each market. Statistically, however, we should expect market-specific moments to contain more information, reducing the variance of the estimator.

However, it is well-known that adding many moment conditions asymptotically biases the standard GMM estimator (Han and Phillips, 2006; Newey and Windmeijer, 2009). In Figure 2 we illustrate this bias-variance tradeoff. From left to right, we increase the number of micro moments, pooling them across a decreasing number of markets. This reduces the variance of the estimator at the cost of some bias. In general, we prefer more micro moments to fewer, particularly if markets are very observably different, since this will reduce the variance of the estimator. However, much like adding many instruments to simple linear IV regressions can be problematic (see, e.g., Angrist, Imbens, and Krueger, 1999), it is important to be aware of bias or lack of interpretability that one might be introducing by

---

[69]Strictly speaking, without changing the data generating process, this model would be misspecified so we omit it from Table 6.

[70]Given summary statistics $\bar{v}_t$ each based on N_t observations, the pooled summary statistic would be $\sum_t N_t \cdot \bar{v}_t / \sum_t N_t$.

adding a large number of moments.

Numerical Integration

In Table 7 we consider another important choice: how to choose sets of consumer types $\mathcal{I}_t$ to numerically integrate over a population of consumers. In Conlon and Gortmaker (2020) we emphasize how bounded and continuously differentiable integrals for market shares can be well-approximated with a small number of quadrature nodes and weights.[71] In the first two rows of Table 7 we compare $|\mathcal{I}_t| = 7$ Gauss-Hermite quadrature nodes with $|\mathcal{I}_t| = 1,000$ Monte Carlo draws from the true distribution of income y_{it}. Statistical performance is comparable, but with quadrature, it takes two orders of magnitude less time to compute the estimator.[72]

In the bottom two rows, we provide a typical example for which quadrature should not be used. Instead of matching a "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" moment, which is continuously differentiable in income y_{it}, we match "$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$," which is not because of the low-income indicator. As already discussed, discretizing income discards information, so the estimator performs worse regardless of the integration rule. But more importantly, since quadrature rules are specific to the domain of integration (e.g., a normal density over $\mathbb{R}$), they will not correctly integrate sub-intervals. This becomes apparent in Table 7. Other than not using quadrature in these cases, there are no obvious solutions when computing s_{jt} requires integrating over the entire distribution and the micro moments require integration over a sub-interval.

Problem Scaling

In Section 4 and Appendix D we discuss the econometric properties of the micro BLP estimator under different asymptotic thought experiments: (a) many markets, including those covered by surveys; (b) many markets, few with surveys, but the surveys are large; and (c) few markets, but markets and surveys are both large. Still using "$\mathbb{E}[y_{it} \mid j \neq 0]$" and "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" micro moments, in Figure 3 we use our simulations to illustrate that the estimator's reasonable rates of convergence translate to finite samples. From left to right, each column corresponds to cases (a), (b), and (c), respectively. For all cases, the variance of the estimator decreases similarly as we increase the number of aggregate and

---

[71]Theoretically, the integrand is approximated with a polynomial and then integrated exactly.

[72]For more dimensions of integration—more demographics or unobserved preferences—this computational performance gap decreases, and quadrature, including more sophisticated sparse grids, becomes comparable to Monte Carlo methods. See Figure 1 in Conlon and Gortmaker (2020).

micro observations. In Appendix H we document that standard error estimators have good coverage and low bias in finite samples, reflecting the asymptotic normality of the estimator.

Unobserved Heterogeneity

So far, our simulations only model one source of observed heterogeneity: income. To discuss the role of unobserved heterogeneity, we draw unobserved preferences ν_{2it} for x_{2jt} from the standard normal distribution and use 1,000 scrambled Halton draws (Owen, 2017) to approximate this distribution during estimation. We then add a $\sigma_x x_{2jt} \nu_{2it}$ term to heterogeneous utility, and choose σ_x to make unobserved preferences fairly important:

$$\mu_{ijt} = \pi_1 y_{it} + \pi_x x_{2jt} y_{it} + \sigma_x x_{2jt} \nu_{2it}, \qquad \begin{aligned} \Pi_0 &= (\pi_{01}, \pi_{0x}, 0)' = (-0.1, 0.1, 0)', \\ \Sigma_0 &= \mathrm{diag}(0, \sigma_{0x}, 0) = \mathrm{diag}(0, 0.5, 0). \end{aligned} \qquad (30)$$

In Table 8 we illustrate how the standard "$\mathbb{E}[y_{it} \mid j \neq 0]$" and "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" moments still greatly improve the performance of the estimator. Optimal micro moments do even better.

Our default configuration has a great deal of cross-market variation in choice sets $\mathcal{J}_t$, including the number of products $|\mathcal{J}_t|$ and the values of product characteristics (x_{jt}, ξ_{jt}). This is precisely the type of aggregate variation that is needed to identify Σ (Berry and Haile, 2014). As a result, particularly because we are using optimal instruments, $\hat{\sigma}_x$ has very low bias and variance, even without any micro moments.

In the bottom three rows, we use the same choice set $\mathcal{J}_t = \mathcal{J}$ in each market. Even with optimal instruments, $\hat{\sigma}_x$ has a substantial amount of bias and variance, and including micro data that link demographics to choices does not particularly improve the performance of $\hat{\sigma}_x$. This illustrates an important insight of Berry, Levinsohn, and Pakes (2004) that is formalized nonparametrically by Berry and Haile (2022): cross-market choice set variation is still needed to nonparametrically identify Σ, even when using within-market variation that links demographics to choices.[73]

Second Choices

Some datasets will simply not exhibit much cross-market choice set variation, either because there is only a single or a few markets, or because product offerings are fairly uniform. An alternative is using second choice data. Intuitively, each second choice observation is similar

---

[73]In our simulations, identification of Π comes from both within- and cross-market variation in demographics, as well as our parametric assumptions about how demographics enter into utility.

to observing a counterfactual market in which the consumer's first choice is removed from the choice set.

In Table 9 we illustrate the benefits from second choice data for our configuration with no choice set variation. In addition to the main micro dataset, we simulate a second, independent micro dataset that conditions on inside choices as well, but also reports second choices.

Matching the covariance "$\mathbb{C}(x_{2jt}, x_{2k(-j)t} \mid j, k \neq 0)$" between the exogenous product characteristic for first and second choices greatly reduces the variance of $\hat{\sigma}_x$. Matching the sum "$\mathbb{E}[x_{2jt} + x_{2k(-j)t} \mid j, k \neq 0]$," which is closer to the score for σ_x in (26), reduces the variance even more. Figure 4 reports a correlation matrix between micro values underlying these moments and the score for σ_x. Since $x_{2jt} \cdot x_{2k(-j)t}$ and $x_{2jt} + x_{2k(-j)t}$ are highly correlated with one another, their correlations with the score are similar.

We also consider matching the share of consumers who divert from a low- or high-x_{2jt} first choice j to a low-x_{2kt} second choice k. The hope is that this type of diversion ratio is easier to measure or more likely to be collected than the covariance. Discretizing x_{2jt} in this way reduces the correlation of each individual diversion ratio with the score, but in our simulations, matching only two diversion ratios is comparable in terms of variance reduction with the standard "$\mathbb{C}(x_{2jt}, x_{2k(-j)t} \mid j, k \neq 0)$" moment.

If the full second choice micro data are available, we can do even better. In the bottom row of Table 9 we show that the estimator is further improved when we use optimal micro moments that for the second GMM step match the scores of π_1, π_x, and σ_x.

8. Empirical Examples

We provide two empirical examples to illustrate how to use micro moments with real data. We first replicate Petrin (2002) to highlight the importance of incorporating demographics. Second, we demonstrate how to use NielsenIQ data and how to collect second choices in a more modern empirical example estimating demand for soft drinks.

Each market is a different time period for the same geographic region: either the entire US in Petrin (2002) or the city of Seattle in our soft drink example. Although there are only a few time periods, and hence limited cross-market variation, there are many products and micro observations. Both examples are ones for which the many product/large survey asymptotics seem to be most appropriate.

Petrin (2002) Replication

We estimate the model of Petrin (2002) and replicate its primary counterfactual: quantifying the consumer welfare gain from the introduction of the minivan. This paper was the first to incorporate micro moments into the BLP framework, and its counterfactual highlights how important it can be to incorporate demographics. Like Berry, Levinsohn, and Pakes (1995), Petrin (2002) also derives an additional set of aggregate moment conditions from the first-order pricing conditions of firms. We demonstrate how to construct and solve the problem with PyBLP in Figure 5.

After confirming that we can exactly replicate the published estimates from the original paper's IV logit model, we estimate the paper's micro BLP model and calculate counterfactual welfare twice. First, we follow the original paper by using the sample covariance matrix of micro moments estimated from the full micro data. Second, we discard this matrix and let PyBLP estimate the moments' covariances at first-step parameter estimates. The appeal of the latter approach is that it only requires summary statistics from the micro data, not their covariances, which will often not be reported by surveys. The two approaches are asymptotically equivalent, and we get nearly identical estimates.

We report our results in Table 10. Compared with the published estimates, results are similar, particularly those for marginal costs, although there are some substantial differences for the price and random coefficients.[74] In particular, we estimate somewhat lower price elasticities. We do get a similar estimate for the headline 1984 compensating variation from the introduction of the minivan: $430 million (with a standard error of $250 million) compared with $367 million estimated by the original paper. In line with the original paper, a large difference compared to the estimate under the logit model highlights the importance of including demographics in this setting.[75] We do not report estimates with optimal micro moments because the original paper's replication package does not include the complete (proprietary) micro data, only summary statistics.

We cannot perfectly replicate the original paper because its replication package does not include the importance sampling nodes and weights used in the final specification. Instead, we use 1,000 scrambled Halton draws (Owen, 2017), and find that after this point,

---

[74]Results would also be similar for the base coefficients, but Petrin (2002) uses a truncated $\chi^2(3)$ distributions for unobserved preferences, which, unlike the more standard $N(0, 1)$ distributions, are not mean zero, so differences in random coefficients that scale unobserved preferences shift mean preferences.

[75]The original paper only reports compensating variation for the logit model across multiple years, so we compute compensating variation for 1984 ourselves in the first column of Table 10. The logit parameter estimates in Table 10 are our own, and match those in the original paper up to rounding error.

increasing this number does not much change our estimates. Another important difference is that instead of using the derivative-free Nelder-Mead algorithm, which can be slow and perform poorly (Conlon and Gortmaker, 2020) we supply analytic gradients to a BFGS-based optimizer, and confirm that we get the same estimates for different sets of starting values.

Predicting Substitution from Seattle's Sweetened Beverage Tax

In recent years, one of the most used sources of matched aggregate and micro data for consumer purchases are the NielsenIQ Retailer Scanner and Consumer Panel datasets as provided by the Kilts Center at the Chicago Booth School of Business. The scanner data contains product characteristics and weekly sales for a large sample of retailers across the US. The consumer data contains consumer demographics and purchase decisions for a large sample of participating US households.

To demonstrate how to use micro moments with NielsenIQ data, we estimate pre-2017 demand for soft drinks in Seattle. We then predict what would happen if prices increased by how much they did after the 2018 implementation of Seattle's sweetened beverage tax (SBT)—the most recent SBT implemented in the US—and compare our substitution estimates to what actually happened. We view this exercise as in-between answering a policy question and a Monte Carlo, since we are using real data but already know what happened. This type of exercise could be repeated for different cities to evaluate the potential effects of proposed taxes.[76]

In Appendix J we discuss all the decisions we make when constructing our data: market definition, demographic data, product data, instruments, market sizes, micro data, and a custom second choice survey. We also discuss other decisions we could have made, weighing their pros and cons. We hope Appendix J will be particularly helpful for researchers using NielsenIQ data or collecting second choice data to estimate their own demand systems.

We collect quarterly sales data from 2007 to 2016 on 2,672 soft drink UPCs sold at five large retailers in Seattle, for a total of $N_{\mathrm{A}} = 78,161$ product-retailers.[77] In each quarter t, the market share $\mathcal{S}_{jt}$ of product-retailer j is total ounces purchased divided by the market size $\mathcal{M}_t$. To compute market sizes, we estimate of the number of trips made to these retailers

---

[76]Similarly, Zhen et al. (2014) estimate an Exact Affine Stone Index (EASI) demand model (Lewbel and Pendakur, 2009) that includes 23 different categories related to soft drinks to evaluate the impact of SBTs. EASI is a product space approach to demand estimation, which we view as complementary to characteristics space approaches like BLP.

[77]This includes fruit drinks and diet drinks, but for simplicity we do not consider juice or other sugary product categories. We combine product-retailers in the bottom 5% of ounces sold with the outside good in each quarter.

and scale this by a maximum potential demand per trip of 720 ounces.[78] Later in this section we will discuss how such market size assumptions can affect estimates.

Our product characteristics x_{jt} are price and indicators for diet and small-sized drinks.[79] We use a Hausman (1996)-type instrument for prices (contemporaneous prices in cities other than Seattle) that is very similar to the one used by Allcott et al. (2019) and construct a Gandhi and Houde (2020)-style differentiation IV to identify the standard deviation of normally distributed unobserved preference heterogeneity for price.[80]

We report aggregate BLP estimates in the first column of Table 11. We include product-retailer and retailer-quarter fixed effects to account for product-specific preferences and time-varying demand for retailers. We cluster standard errors by brand $b(j)$. Across specifications, our estimated price elasticity of demand is around -1.3, which is on the high end of typical estimates in the existing literature between -0.8 and -1.4 (e.g., Powell et al., 2013).

We also report results from our counterfactual in which we increase the prices of taxed 2016 products by how much they seemed to have increased after the introduction of the 2018 SBT of 1.75 cents per ounce:[81] 1.15 cents for taxed small-sized drinks and 0.97 cents for taxed family-sized ones (Powell and Leider, 2020).[82] We use a manual classification of taxed and untaxed goods that was created and graciously provided to us by the authors and research team of Powell and Leider (2020). Although the Seattle tax excluded diet beverages, tax status is not one-to-one with our diet indicator, with a strong but imperfect correlation of -0.76 in our 2016 data. Our estimates of a decrease in taxed ounces purchased of -30% and a small increase in untaxed ounces of 1% are not too far from the -22% and 4% estimated by Powell and Leider (2020), but as we discuss below, could benefit from additional dimensions of preference heterogeneity.

The political discourse surrounding SBTs and related economic theory emphasizes their

---

[78]We have an in-depth discussion of market sizes and how we form these estimates in Appendix J.

[79]Following Powell and Leider (2020), we define small- or individual-sized products as single-unit beverages that are no more than one liter in volume. Diet classification is in Appendix J, and is particularly important for this setting because the Seattle tax excluded diet drinks. A more in-depth study of soft drink demand would incorporate random coefficients on more characteristics.

[80]As we discuss below, we do not attempt to identify the distributions of random coefficients on other diet and small-sized indicators with only aggregate data because along these dimensions, cross-quarter choice set variation is very limited.

[81]Although modeling the supply side and predicting passthrough is beyond the scope of this paper, doing so would be useful for informing SBT policy. To obtain reasonable passthrough estimates, one would likely need to model the joint pricing decisions of retailers and distributors, since both are important in the market for soft drinks.

[82]Powell and Leider (2020) estimate these passthrough rates of 66% and 55% with a differences-in-differences approach, using Portland as the control group. They also use NielsenIQ scanner data.

differential effects by income (see, e.g., Allcott et al., 2019; Conlon et al., 2022). To predict differential substitution by income group, we include an indicator in demographics y_{it} for households with income above the 2016 median in Washington. We also include an indicator for households with at least one child.[83] We construct demographic shares for each of the four bins from annual American Community Survey (ACS) data for Seattle, and re-weight NielsenIQ households in Seattle by these ACS shares.

Since at the city level these demographics vary little during our sample period,[84] we do not attempt to identify how preferences vary by demographic group with only cross-market variation.[85] Indeed, following our advice from Section 3, running a FRAC regression gives very noisy point estimates for Π. This is unsurprising because such estimates are essentially formed from $2016 - 2007 = 9$ observations.

Instead, we match two sets of standard micro moments: "$\mathbb{E}[y_{rit} \mid j \neq 0]$" and "$\mathbb{C}(x_{cjt}, y_{rit} \mid j \neq 0)$" for the $R = 2$ demographics and $C = 3$ characteristics. We use the Consumer Panel data and compute $m = 1, \ldots, M_M = 8$ micro moment sample values $f_m(\bar{v})$ from a sample of $N_d = 10{,}455$ grocery trips with an inside purchase $j \neq 0$.[86]

The second column of Table 11 reports micro BLP estimates. We estimate a slight decline of price sensitivity with income,[87] and households with children also tend to be more price sensitive. Both low income households and those with children tend to dislike diet drinks. Incorporating micro data allows us to predict how the tax counterfactual differentially affects consumers by demographic group. Slightly more elastic demand for households with low income or children results in slightly more substitution away from taxed goods. However, compared to a baseline reduction in taxed volume of 30%, we are able to reject predicted differences of more than 4 percentage points for low versus high income households and 7 percentage points for households with versus without children at a 5% significance level.

These predictions are generally in-line with those of Barker et al. (2022), who pool 529 households in the NielsenIQ Consumer Panel dataset together with data before and after the

---

[83]We limit our attention to two binary demographics for simplicity in this empirical example. A more in-depth study would incorporate more functions of demographics measured in Census and NielsenIQ data.

[84]The share of high income households increases from 35% in 2007 to only 40% in 2016. The share of households with at least one child increases from 10% to only 11%.

[85]If we try to do so by including instruments that interact moments of the demographic distribution with characteristics and differentiation IVs, we get very noisy estimates that severely corrupt our other estimates.

[86]We compute weighted averages and covariances to account for both non-random participation of households in the NielsenIQ panel and different numbers of total grocery trips per quarter. See Appendix J for more details.

[87]An unconditional negative covariance in the micro data between prices and high income is potentially misleading. High income households also tend to purchase cheaper family-sized products. This negative covariance switches sign after controlling for package size.

implementation of seven recent SBTs in the US between 2015 (Berkeley) and 2018 (Seattle), and struggle to find statistically significant differences in the impact of these taxes by income group and presence of children. We view a structural approach that incorporates micro data as complementary to approaches such as that of Barker et al. (2022), which makes different modeling assumptions but can be limited by small sample sizes.[88]

Incorporating demographics captures some heterogeneous preferences for the outside good and diet beverages. But the model is missing a great deal of potential unobserved heterogeneity. Unfortunately, with market fixed effects, the distribution of unobserved preferences for the outside good is not identified with only aggregate data,[89] and we find that cross-quarter aggregate variation in the number of diet drinks is also insufficient to precisely estimate the scale of unobserved preferences for diet drinks.

Instead, we use survey-based second choice data.[90] To demonstrate how researchers can run a second choice survey, we use Prolific Academic to recruit 100 participants who live in Washington State for an online survey.[91] Our survey design is similar to that used for choice-based conjoint analysis (e.g., Allenby et al., 2019), and we provide more details at the end of Appendix J, including discussion of potential biases that often show up in results from online surveys. We use the survey to compute two diversion ratios: the share of participants who would divert to the outside good or a diet soft drink if their first choice non-diet brand were unavailable.[92] Like in the NielsenIQ micro data, we weight observations by ounces typically purchased and adjust for non-random sampling by demographic group.

In the third column of Table 11, we match these two diversion statistics for the last quarter in our sample.[93] If respondents' non-diet first choice soft drink brand were unavailable,

---

[88] One country with much larger sample sizes for studying the impact of SBTs is the UK, through data collected by the National Child Measurement Programme (see, e.g., Rogers et al., 2023).

[89] Recalling the FRAC intuition from Section 3, the artificial regressor on a constant $x_{jt} = 1$ is $a_{jt} = s_{0t} - 1/2$, variation of which is absorbed by market t fixed effects.

[90] Another approach would be to compute first- and seconds shares from a single household's purchases over time in the NielsenIQ micro data. The validity of this approach will depend on what generated changes in product availability or characteristics that led to switching. Since we generally expect product availability to be correlated across products, we prefer self-reported second choices, but observational diversion can be useful in settings where survey data is unavailable or unreliable.

[91] A larger sample size would be appropriate for a more complete empirical study. Allenby et al. (2019) notes that many conjoint practitioners use sample sizes of 500 to 1,000.

[92] We increase the total number of survey participants to 139 until we get 100 participants who say they have purchased at least one of eight of the most popular non-diet brands in Seattle during the last 30 days: Coke, Pepsi, Gatorade, Powerade, Canada Dry, Dr Pepper, Mountain Dew, or Seven Up.

[93] By matching statistics computed for Washington residents, not just Seattle residents, and for consumers in 2023, not 2016, we are assuming that these diversion ratios would not be much different for Seattle in 2016. At a minimum, in Appendix J we check whether the statistics are different for the 25% of respondents who live in Seattle and do find some difference for diversion to the outside good, although they are noisy.

"$\mathbb{P}(\text{Diet}_{k(-b(j))t} \mid \text{Surveyed Non-diet}_{jt}) = 16\%$" of respondents said they would divert to a diet beverage, and "$\mathbb{P}(k(-b(j)) = 0 \mid \text{Surveyed Non-diet}_{jt}) = 17\%$" said they would divert to the outside good, which includes both non-soft drinks and no beverage. Without matching these two additional moments, the model predicts 92% and 3%, respectively, suggesting that there is a great deal of unobserved preference heterogeneity left unmodeled.

Indeed, we get large estimated standard deviations on normally distributed unobserved preferences for inside goods and the diet characteristic. As a result, the counterfactual predicts a smaller decrease in taxed volume purchased, -16%, somewhat undershooting the estimate of -22% in Powell and Leider (2020), and a larger increase in untaxed volume purchased, 9%, somewhat overshooting but not statistically different from the estimate of 4% in Powell and Leider (2020). Given the nature of an imperfect prediction exercise, we do not expect to perfectly predict what actually happened, but do view our second choice estimates as more credible than those that rely more heavily on strong assumptions about market size.

Finally, in the fourth column of Table 11, we replace the standard micro moments with optimal micro moments in the second GMM step. We do not replace our second choice moments because, as is often the case, our survey did not collect full micro data, only enough to compute our desired diversion ratios. Point estimates and counterfactual results are fairly similar, suggesting that most of the information in the NielsenIQ micro data is already spanned by the standard micro moments for this model. This should not be surprising because the model discretized observed heterogeneity into four types: high and low income, and with and without children. We provide more in-depth discussions of how to compute optimal micro moments with PyBLP at the end of Appendix E and how to do so with NielsenIQ micro data near the end of Appendix J.

There are a number of extensions that would improve a more complete policy exercise. Incorporating more product characteristics, more consumer demographics, and more second choice data would help to better explain substitution patterns. Discussed in Appendix A, a lognormal random coefficient on price often provides a better fit, and can be helpful for modeling a supply side.[94] In Appendix B we discuss adding a nesting structure, which could be useful for explaining substitution between categorical characteristics such as brand or store. PyBLP also supports inclusion of product-specific demographics such as geographic distance to stores, which could allow researchers to predict cross-border shopping effects.[95]

---

[94]This guarantees downward sloping demand for all consumers, which can help guarantee pricing equilibrium existence and uniqueness.

[95]We discuss adding geographic distance in more depth in Appendix J.

9. Conclusion and Practical Advice

This article was motivated with a frustration experienced by many researchers with the aggregate BLP estimator: aggregate variation is usually very limited, leading to poor estimates of demand. Coupled with the best practices from Conlon and Gortmaker (2020) for the aggregate side of estimation, we confirm in this article that incorporating micro data can substantially improve the finite-sample performance of the BLP estimator. Our hope is that going forward, a standardized framework for doing so will encourage more researchers to use or collect micro data, particularly second choices, which can be very useful for estimating the degree of unobserved preference heterogeneity.

This article makes a number of contributions that we believe to be novel. Perhaps most importantly, we provide a flexible econometric framework for incorporating many different types of micro data into BLP-style estimation, which we subject to a number of different asymptotic thought experiments. These include cases where we observe relatively complete data on individual choices, demographics, and characteristics, and cases where we observe only limited statistics from surveys of individuals. Characterizing the asymptotic covariance matrix also allows us to clarify that researchers do not need to observe sample covariances between micro summary statistics to do valid statistical inference. Finally, we contribute a novel characterization of the optimal micro moments in the spirit of Chamberlain (1987) and a computationally straightforward procedure for computing them, which can be done with only a few lines of code when using PyBLP. These have the advantage of not only reducing bias and increasing efficiency, but can also significantly reduce the overall dimension of the problem.

We also provide some practical tips to researchers. First, researchers can and should check how much cross-sectional (or time series) variation there is in the aggregate data using the FRAC estimator of Salanié and Wolak (2022). Second, researchers can measure how much of the variation in the (infeasible) optimal micro moments from the score contributions can be captured using their micro statistics, even if complete individual data is not available. Third, researchers should be mindful of compatibility issues across datasets. The marginal distribution of demographics like income, or the purchase probabilities of particular choices may vary significantly between aggregate and micro datasets. In this case, blindly matching moments from micro datasets (including the "optimal micro moments") may be worse than using only aggregate data. However, we illustrate that alternative micro statistics can be designed to be more robust in this scenario. Fourth, while quadrature rules are often the best choice for evaluating numerical integrals of mixed logit models with aggregate data,

most quadrature rules are not designed to accurately integrate sub-intervals; in this case, less accurate (but unbiased) Monte Carlo rules may be preferred. Finally, researchers should think about which model parameters are most relevant for the policies they are interested in, and carefully consider designing surveys or experiments to help better estimate those objects. Here we provide a proof of concept showing how a small and inexpensive survey could be designed to better understand the effects of a sugary beverage tax.

Our goal has been to extend the best practices in Conlon and Gortmaker (2020) to the case with micro data, not only through this paper but also in a single software package, PyBLP. We have provided a list of best practices, evaluated them with large-scale simulations, and made them either defaults or easy to use in PyBLP. Our hope is that these practices can now be made available to a wider range of researchers, including those already using PyBLP. For researchers who wish to incorporate micro data into similar econometric frameworks that are not yet supported by PyBLP, we hope that the framework and results developed in this article, along with PyBLP's well-documented code, serve as a useful starting point.

Table 1: Empirical Literature

Paper	Demand Estimation		
	Industry	Country	Years
Petrin (2002)	Automobiles	United States	1981–1993
Berry, Levinsohn, and Pakes (2004)	Automobiles	United States	1993
Thomadsen (2005)	Fast Food	United States	1999
Goeree (2008)	Personal Computers	United States	1996–1998
Ciliberto and Kuminoff (2010)	Cigarettes	United States	1993–2002
Nakamura and Zerom (2010)	Coffee	United States	2000–2004
Beresteanu and Li (2011)	Automobiles	United States	1999–2006
Li (2012)	Automobiles	United States	1999–2006
Copeland (2014)	Automobiles	United States	1999–2008
Starc (2014)	Health Insurance	United States	2004–2008
Ching, Hayashi, and Wang (2015)	Nursing Homes	United States	1999
Li, Xiao, and Liu (2015)	Automobiles	China	2004–2009
Nurski and Verboven (2016)	Automobiles	Belgium	2010–2011
Barwick, Cao, and Li (2017)	Automobiles	China	2009–2011
Murry (2017)	Automobiles	United States	2007–2011
Wollmann (2018)	Commercial Vehicles	United States	1986–2012
Li (2018)	Automobiles	China	2008–2012
Li, Gordon, and Netzer (2018)	Digital Cameras	United States	2007–2010
Backus, Conlon, and Sinkinson (2021)	Cereal	United States	2007–2016
Grieco, Murry, and Yurukoglu (2021)	Automobiles	United States	1980–2018
Neilson (2021)	Primary Schools	Chile	2005–2016
Armitage and Pinter (2022)	Automobiles	United States	2009–2017
Döpper, MacKay, Miller, and Stiebale (2022)	Retail	United States	2006–2019
Durrmeyer (2022)	Automobiles	France	2003–2008
Weber (2022)	Trucks	United States	2010–2018
Bodéré (2023)	Preschools	United States	2010–2018
Montag (2023)	Laundry Machines	United States	2005–2015
Conlon and Rao (2023)	Distilled Spirits	United States	2007–2013

This table collects a non-exhaustive list of empirical papers that use the micro BLP estimator, along with the industry, country, and years for which each paper estimates demand. Some papers estimate demand for the listed broad industry and subsequently focus on a sub-industry. We only list published and recent working papers that do not diverge too much from the standard demand-side BLP model. In Table 3 we reorganize these papers by which micro moments they use.

Table 2: Notation

Notation for aggregate data and estimation (Section 2)		Notation for micro data and estimation (Section 4)	
$t \in \mathcal{T}$	Markets	$d \in \mathcal{D}$	Micro datasets
$\mathcal{M}_t \in \mathbb{R}_+$	Market size	$w_{dijt} \in [0,1]$	Sampling probability
		$w_{dijkt} \in [0,1]$	Joint sampling probability
$j \in \mathcal{J}_t$	Products		
$j = 0$	Outside alternative	$n \in \mathcal{N}_d$	Micro observations
$c = 1, \ldots, C$	Observed product characteristics	$t_n \in \mathcal{T}$	Micro observation market
$m = 1, \ldots, M_{\mathrm{A}}$	Instruments	$i_n \in \mathcal{I}_{t_n}$	Micro observation type
$x_{cjt} \in \mathbb{R}$	Observed product characteristic	$j_n \in \mathcal{J}_{t_n} \cup \{0\}$	Micro observation choice
$x_{jt} \in \mathbb{R}^{C \times 1}$	All observed product characteristics	$k_n \in \mathcal{J}_{t_n} \cup \{0\} \setminus \{j_n\}$	Micro observation second choice
$z_{mjt} \in \mathbb{R}$	Instrument		
$z_{jt} \in \mathbb{R}^{M_{\mathrm{A}} \times 1}$	All instruments	$p = 1, \ldots, P_{\mathrm{M}}$	Micro parts
$\xi_{jt} \in \mathbb{R}$	Mean-zero unobserved product quality	$d_p \in \mathcal{D}$	Micro part dataset
		$v_{pijt} \in \mathbb{R}$	Micro part value
$i \in \mathcal{I}_t$	Consumer types	$v_{pijkt} \in \mathbb{R}$	Second choice micro part value
$r = 1, \ldots, R$	Consumer demographics		
$w_{it} \in [0,1]$	Consumer type share	$m = 1, \ldots, M_{\mathrm{M}}$	Micro moments
$y_{rit} \in \mathbb{R}$	Consumer demographic	$f_m : \mathbb{R}^{P_{\mathrm{M}} \times 1} \to \mathbb{R}$	Micro moment function
$y_{it} \in \mathbb{R}^{R \times 1}$	All consumer demographics		
$\nu_{cit} \in \mathbb{R}$	Unobserved preference	$\bar{v}_p \in \mathbb{R}$	Micro part sample value
$\nu_{it} \in \mathbb{R}^{C \times 1}$	All unobserved preferences	$\bar{v} \in \mathbb{R}^{P_{\mathrm{M}} \times 1}$	All micro part sample values
		$f_m(\bar{v}) \in \mathbb{R}$	Micro moment sample value
$u_{ijt} \in \mathbb{R}$	Indirect utility		
$\delta_{jt} \in \mathbb{R}$	Mean utility	$v_p(\theta) \in \mathbb{R}$	Micro part expected value
$\mu_{ijt} \in \mathbb{R}$	Heterogeneous utility	$v(\theta) \in \mathbb{R}^{P_{\mathrm{M}} \times 1}$	All micro part expected values
$\varepsilon_{ijt} \in \mathbb{R}$	Idiosyncratic preference	$f_m(v(\theta)) \in \mathbb{R}$	Micro moment expected value
$s_{ijt} \in (0,1)$	Choice probability		
$s_{jt} \in (0,1)$	Market share	$s_{ijkt} \in (0,1)$	Joint choice probability
$\mathcal{S}_{jt} \in (0,1)$	Observed market share	$s_{ik(-j)t} \in (0,1)$	Probability of choosing k without j
		$s_{ik(-h(j))t} \in (0,1)$	The same, but without a group $h(j)$
$\beta \in \mathbb{R}^{C \times 1}$	Linear parameters		
$\Pi \in \mathbb{R}^{C \times R}$	Consumer demographic parameters	$M = M_{\mathrm{A}} + M_{\mathrm{M}}$	Number of combined moments
$\Sigma \in \mathbb{R}^{C \times C}$	Unobserved preference parameters	$\hat{g}(\theta) \in \mathbb{R}^{M \times 1}$	Combined sample moments
$\theta = (\beta, \Pi, \Sigma)$	All parameters	$\hat{W} \in \mathbb{R}^{M \times M}$	Combined weighting matrix
$N_{\mathrm{A}} = \sum_{t \in \mathcal{T}} |\mathcal{J}_t|$	Number of aggregate observations	$N_d = |\mathcal{N}_d|$	Number of micro observations
$\hat{g}_{\mathrm{A}}(\theta) \in \mathbb{R}^{M_{\mathrm{A}} \times 1}$	Aggregate sample moments	$\hat{g}_{\mathrm{M}}(\theta) \in \mathbb{R}^{M_{\mathrm{M}} \times 1}$	Micro sample moments
$\hat{W}_{\mathrm{A}} \in \mathbb{R}^{M_{\mathrm{A}} \times M_{\mathrm{A}}}$	Aggregate weighting matrix	$\hat{W}_{\mathrm{M}} \in \mathbb{R}^{M_{\mathrm{M}} \times M_{\mathrm{M}}}$	Micro weighting matrix

This table summarizes the notation we introduce in Sections 2 and 4. Subscripts on parameters such as θ_0 refer to true values. Subscripts on operators such as $\mathbb{P}_{\mathrm{A}}$ indicate conditioning on all aggregate data.

Table 3: Micro Moment Examples

Shorthand	Papers
"$\mathbb{P}(j \in \mathcal{J}_m \mid i \in \mathcal{I}_m)$"	Petrin (2002); Thomadsen (2005); Goeree (2008); Nakamura and Zerom (2010); Beresteanu and Li (2011); Li (2012); Starc (2014); Ching, Hayashi, and Wang (2015); Li, Xiao, and Liu (2015); Barwick, Cao, and Li (2017); Li (2018); Li, Gordon, and Netzer (2018); Bodéré (2023)
"$\mathbb{E}[y_{rit} \mid j \in \mathcal{J}_m]$"	Petrin (2002); Ciliberto and Kuminoff (2010); Li (2012); Copeland (2014); Nurski and Verboven (2016); Murry (2017); Wollmann (2018); Backus, Conlon, and Sinkinson (2021); Armitage and Pinter (2022); Döpper, MacKay, Miller, and Stiebale (2022); Durrmeyer (2022); Weber (2022); Conlon and Rao (2023)
"$\mathbb{E}[x_{cjt} \mid i \in \mathcal{I}_m, j \neq 0]$"	Starc (2014); Grieco, Murry, and Yurukoglu (2021); Neilson (2021); Weber (2022); Bodéré (2023); Conlon and Rao (2023)
"$\mathbb{C}(x_{cjt}, y_{rit} \mid j \neq 0)$"	Berry, Levinsohn, and Pakes (2004); Nurski and Verboven (2016); Backus, Conlon, and Sinkinson (2021); Durrmeyer (2022); Montag (2023)
"$\mathbb{C}(x_{cjt}, x_{ek(-j)t} \mid j, k \neq 0)$"	Berry, Levinsohn, and Pakes (2004); Grieco, Murry, and Yurukoglu (2021); Montag (2023)

This table lists examples of micro moments that we discuss in Section 5. Each row lists our notation-abusing shorthand and empirical papers from Table 1 that have used essentially the same micro moment.

Table 4: Demographic Variation

|			MAE (%)					Bias (%)				
Variation	Distributions	Markets	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\beta}_1$	$\hat{\beta}_x$	$\hat{\alpha}$	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\beta}_1$	$\hat{\beta}_x$	$\hat{\alpha}$
National	1	40	436.6	133.3	8.3	5.1	1.2	-110.3	-45.6	1.8	2.1	0.1
States	50	40	197.8	60.6	3.9	2.4	1.2	-31.3	-12.6	0.6	0.4	0.1
PUMAs	982	40	97.5	30.0	2.7	1.4	1.2	-5.8	-4.7	0.2	0.2	0.1
National	1	80	327.7	102.8	6.1	4.0	0.9	-98.5	-48.4	2.0	2.1	0.0
States	50	80	139.5	42.7	2.7	1.6	0.9	-7.4	-6.1	0.2	0.2	-0.0
PUMAs	982	80	65.9	21.3	1.9	1.0	0.8	-4.9	-2.4	0.2	0.2	-0.0

This table reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets for different amounts of cross-market demographic variation. We randomly assign each market either to the same national distribution of income, to one of 50 US states, or to one of the 982 Public Use Microdata Areas (PUMAs) used by the American Community Survey (ACS). In the last three rows, we simulate 40 more markets, keeping the same demographic distributions as in the first 40, but with different choice sets.

Table 5: Standard Micro Moments

	MAE (%)		Bias (%)	
Micro Moments Shorthand	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\pi}_1$	$\hat{\pi}_x$
No Micro Moments	197.8	60.6	-31.3	-12.6
"$\mathbb{E}[y_{it} \mid j \neq 0]$"	164.8	44.9	2.9	1.4
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	53.8	11.7	17.5	2.3
"$\mathbb{E}[y_{it} \mid j \neq 0], \mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	34.0	10.8	4.1	1.1
"$\mathbb{E}[y_{it} \mid j \neq 0], \mathbb{E}[x_{2jt} \cdot y_{it} \mid j \neq 0]$"	37.6	12.0	3.2	0.7
"$\mathbb{E}[y_{it} \mid j \neq 0], \mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"	62.7	17.3	0.9	1.0
"$\mathbb{E}[y_{it} \mid j \neq 0], \mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0], \mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	34.2	10.7	4.3	1.1

This table reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets for different combinations of standard micro moments. The cutoff $\bar{y}_t$ is the median income y_{it} in market t.

Figure 1: Standard Micro Moment Correlations
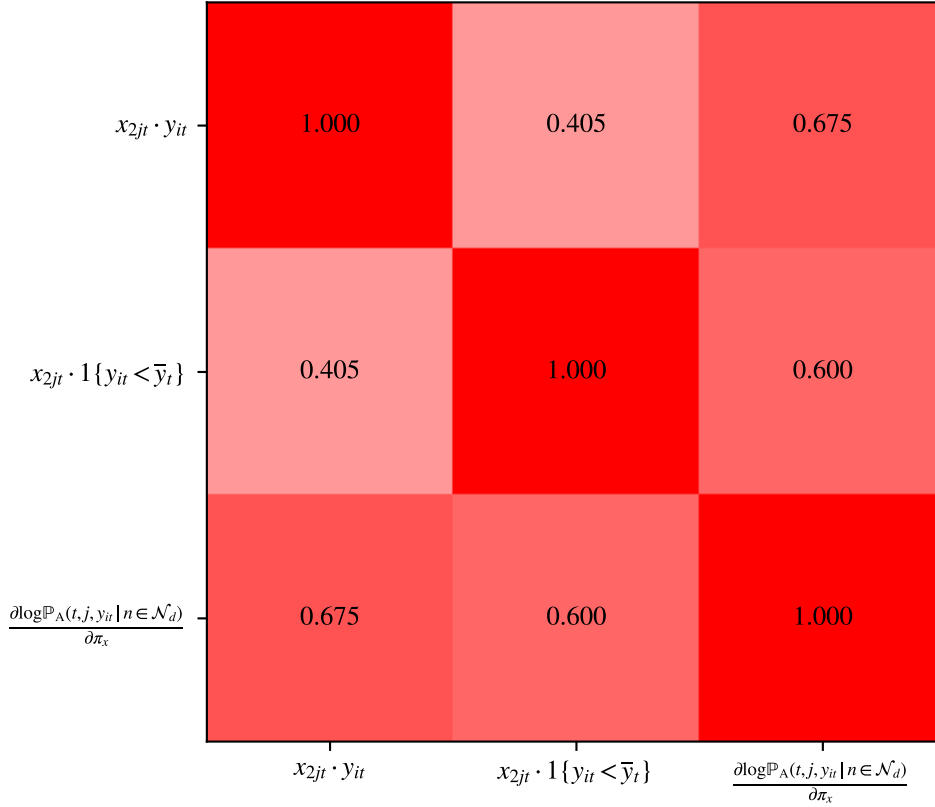
This figure reports median absolute correlations between different micro statistics over 1,000 simulated micro datasets underlying the micro moments in Table 5. For each micro observation n in market $t_n = t$ of type $i_n = i$ with choice $j_n = j$, we compute three statistics: $x_{2jt} \cdot y_{it}$ captures variation in "$\mathbb{E}[x_{2jt} \cdot y_{it} \mid j \neq 0]$" and "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" moments, $x_{2jt} \cdot 1\{y_{it} < \bar{y}_t\}$ captures variation in "$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$" moments, and $\partial \log \mathbb{P}_A(t, j, y_{it} \mid n \in \mathcal{N}_d)/\partial \pi_x$ is the score for π_x at the true θ_0.

Table 6: Optimal Micro Moments and Compatibility

Micro Moments (plus $\mathbb{E}[y_{it} \mid j \neq 0]$)	Incompatible	Optimal	MAE (%)		Bias (%)	
			$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\pi}_1$	$\hat{\pi}_x$
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			34.0	10.8	4.1	1.1
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"		Yes	23.8	6.3	-0.6	-0.1
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"			62.7	17.3	0.9	1.0
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"		Yes	24.1	6.4	-0.5	-0.4
"$\mathbb{E}[x_{2jt} \mid \tilde{y}_{it} < \bar{y}_t, j \neq 0]$"	Yes		64.4	18.0	0.8	0.7
"$\mathbb{E}[x_{2jt} \mid \tilde{y}_{it} < \bar{y}_t, j \neq 0]$"	Yes	Yes	107.1	19.1	104.6	-13.7

This table reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets for standard and optimal micro moments. The first and third rows are the same as the fourth and sixth rows in Table 5. The second and fourth rows use these same standard micro moments in the first GMM step to construct optimal micro moments for the second step. For the last two rows, we simulate a second, independent micro dataset that is configured the same, except we replace y_{it} with $\tilde{y}_{it}$: the 25th percentile of income if below the median or the 75th percentile if above. We use this second dataset for "$\mathbb{E}[x_{2jt} \mid \tilde{y}_{it} < \bar{y}_t, j \neq 0]$" and in the last row, optimal micro moments as well.
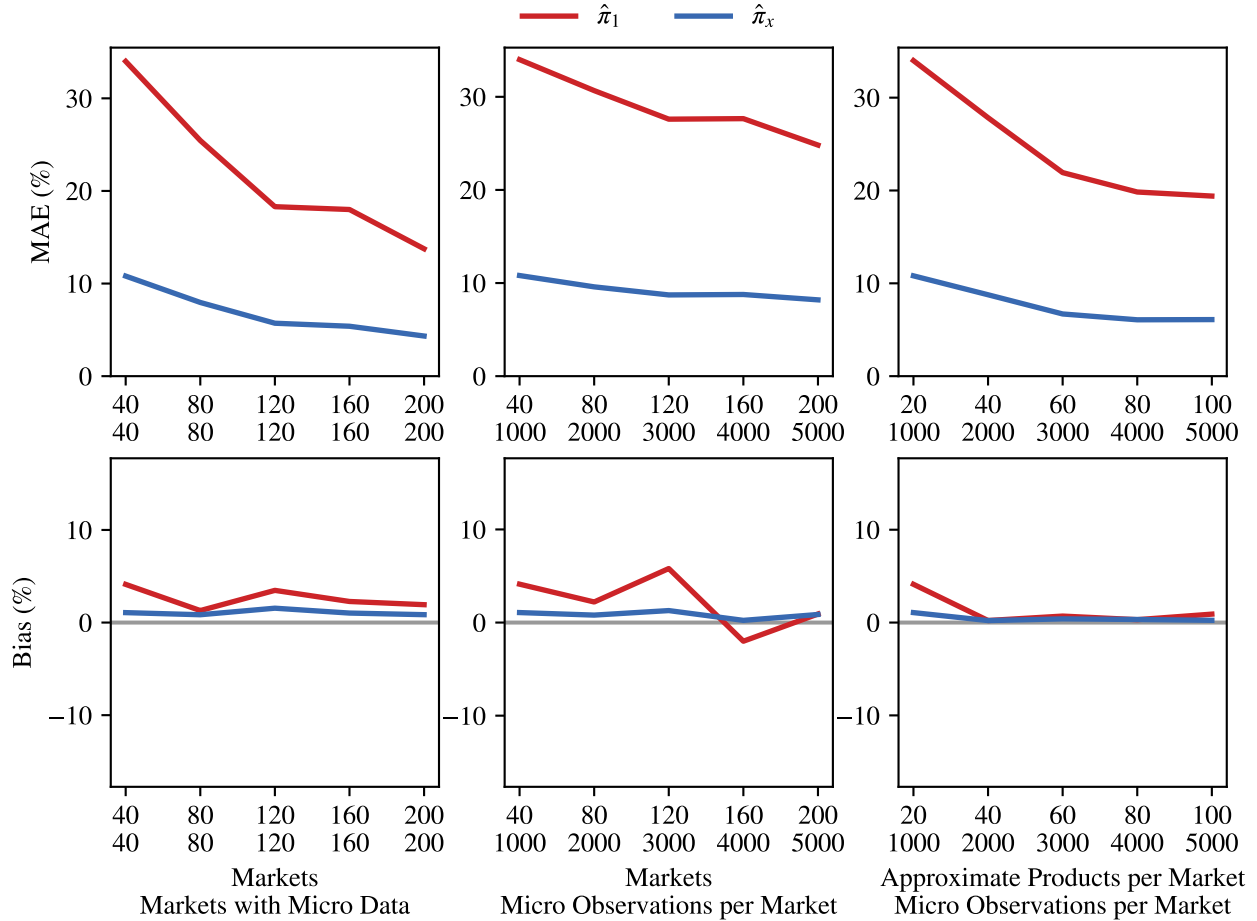
Figure 2: Pooling Markets
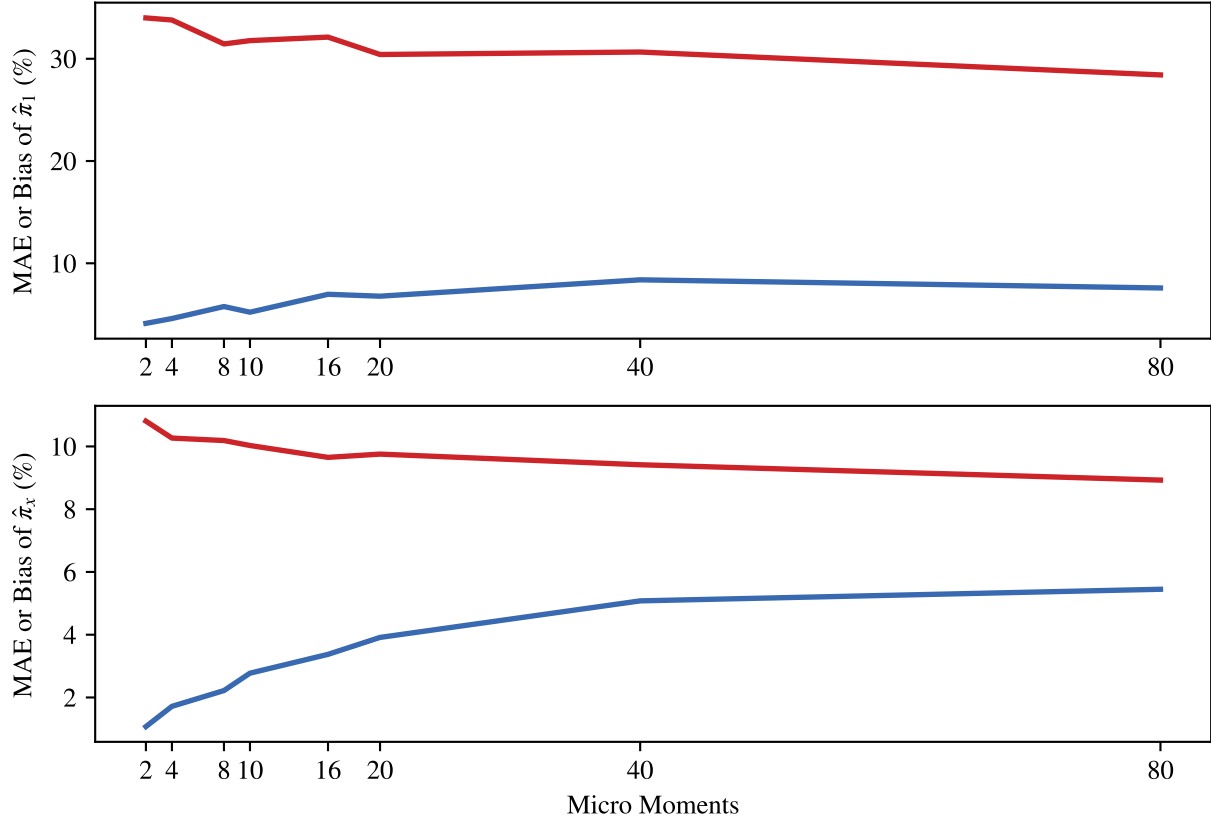
This figure reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets for an increasing number of micro moments that are pooled across a decreasing number of markets. On the left, we match the same $M_M = 2$ micro moments "$\mathbb{E}[y_{it} \mid j \neq 0]$" and "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" in the fourth row of Table 5, which are pooled across all $T = 40$ markets. On the right, we match $M_M = 80$ micro moments "$\mathbb{E}[y_{it} \mid j \neq 0, t]$" and "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0, t)$," one for each market t. In the middle, we pool moments across decreasing numbers of markets (factors of the full 40). We do not use any observables to select which markets to pool for each micro moments. The top panel reports results for $\hat{\pi}_1$; the bottom, for $\hat{\pi}_x$.

Table 7: Numerical Integration

Micro Moments (plus "$\mathbb{E}[y_{it} \mid j \neq 0]$")	Integration	MAE (%)		Bias (%)	
		$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\pi}_1$	$\hat{\pi}_x$
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Quadrature	31.6	9.2	-1.9	-1.1
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Monte Carlo	34.0	10.8	4.1	1.1
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"	Quadrature	251.8	71.0	24.6	5.4
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"	Monte Carlo	62.7	17.3	0.9	1.0

This table reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets for different choices of consumer types $\mathcal{I}_t$ for numerically integrating over the lognormal population distribution of income y_{it}. "Quadrature" refers to $|\mathcal{I}_t| = 7$ Gauss-Hermite nodes and weights that exactly integrate polynomials of degree $2 \times 7 - 1 = 13$ or less. Quadrature nodes are transformed into nodes for income with the mean and standard deviation of log income in each market. "Monte Carlo" refers to $|\mathcal{I}_t| = 1{,}000$ pseudo-Monte Carlo draws from the true distribution of income. The cutoff $\bar{y}_t$ is the median income y_{it} in market t.

Figure 3: Problem Scaling
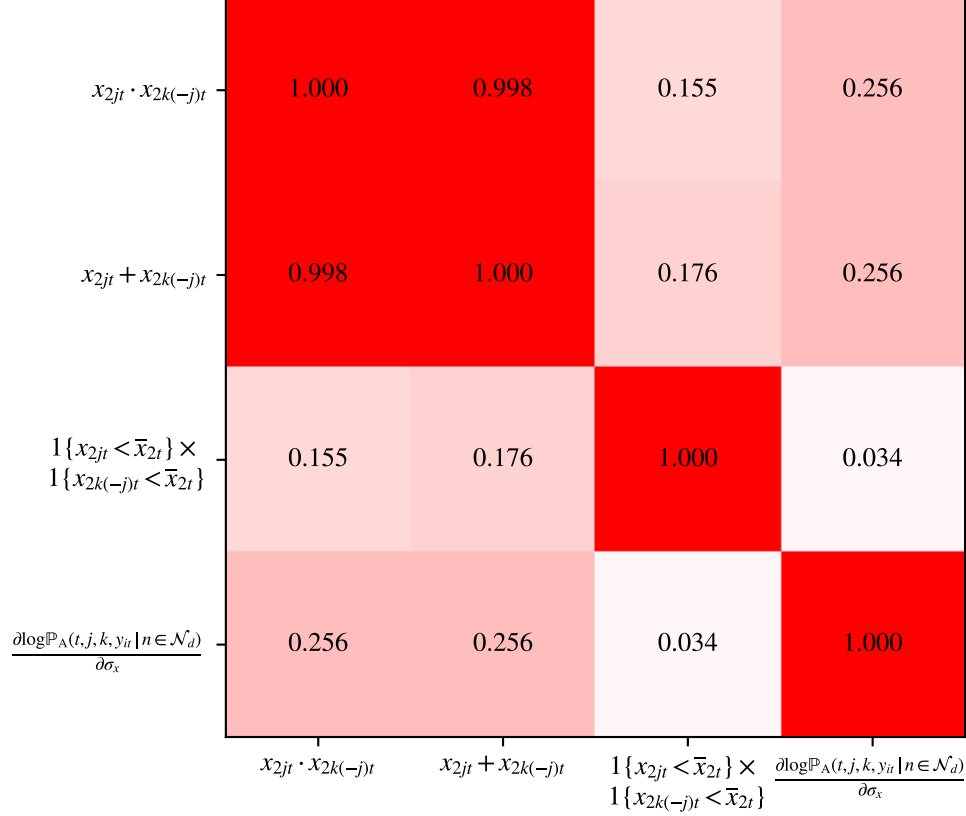
This figure reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets as finite sample sizes approach the three asymptotic thought experiments discussed in Appendix D. In all panels we match the same "$\mathbb{E}[y_{it} \mid j \neq 0]$" and "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" moments in the fourth row of Table 5. The leftmost panel fixes the number of products and micro observations per market and scales up the number of markets, including those with micro data. The middle panel fixes the number of products per market and the number of markets with micro data and scales up the number of aggregate markets and the number of micro observations in each of the fixed number of markets. The rightmost panel fixes the number of markets and scales up the number of products and micro observations per market.

Table 8: Unobserved Heterogeneity

Micro Moments Shorthand	$\mathcal{J}_t = \mathcal{J}$	Optimal	MAE (%)			Bias (%)		
			$\hat\pi_1$	$\hat\pi_x$	$\hat\sigma_x$	$\hat\pi_1$	$\hat\pi_x$	$\hat\sigma_x$
No Micro Moments			225.7	76.5	3.4	-43.4	-14.6	-0.3
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			39.2	12.1	3.2	3.3	0.1	-0.3
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"		Yes	29.1	8.3	3.3	-2.8	-0.7	-0.3
No Micro Moments	Yes		153.6	79.3	99.5	2.8	21.9	31.8
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Yes		33.6	23.1	94.0	-0.8	-13.8	-82.3
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Yes	Yes	31.8	24.0	99.2	-5.5	-17.9	-86.4

This table reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets with unobserved preferences for different amounts of choice set variation and different micro moments. We draw unobserved preferences ν_{2it} from the standard normal distribution and add $\sigma_x x_{2jt}\nu_{2it}$ to μ_{ijt} with the true $\sigma_{0x} = 0.5$. In the bottom three rows, we use the same choice set $\mathcal{J}_t = \mathcal{J}$ in each market, cluster our estimates of the asymptotic covariance matrix for ξ_{jt} by product j, and use the number of markets T as the number of aggregate observations N_{A}.

Table 9: Second Choices

Micro Moments (plus "$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$")	Optimal	MAE (%)			Bias (%)		
		$\hat\pi_1$	$\hat\pi_x$	$\hat\sigma_x$	$\hat\pi_1$	$\hat\pi_x$	$\hat\sigma_x$
No Second Choice Moments		33.2	23.2	94.4	-0.9	-14.0	-82.7
"$\mathbb{C}(x_{2jt}, x_{2k(-j)t} \mid j, k \neq 0)$"		33.9	12.0	16.4	3.7	0.3	-2.3
"$\mathbb{E}[x_{2jt} + x_{2k(-j)t} \mid j, k \neq 0]$"		34.0	10.4	5.3	4.4	1.7	-0.5
"$\mathbb{P}(x_{2k(-j)t} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t}, j, k \neq 0)$"		34.7	11.0	12.5	3.5	1.9	-2.9
"$\mathbb{P}(x_{2k(-j)t} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t}, j, k \neq 0)$"	Yes	16.9	4.8	4.3	-0.3	-0.8	-1.0

This table reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets with unobserved preferences for different micro moments based on second choice data. We draw unobserved preferences ν_{2it} from the standard normal distribution and add $\sigma_x x_{2jt}\nu_{2it}$ to μ_{ijt} with the true $\sigma_{0x} = 0.5$. To eliminate cross-market choice set variation, we use the same choice set $\mathcal{J}_t = \mathcal{J}$ in each market, cluster our estimates of the asymptotic covariance matrix for ξ_{jt} by product j, and use the number of markets T as the number of aggregate observations N_{A}. In addition to the main micro dataset, we simulate a second, independent micro dataset that is configured the same, except that it also reports second choices. The shorthand "$\mathbb{P}(x_{2k(-j)t} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t}, j, k \neq 0)$" refers to two moments that match the share of individuals who divert from a below- or above-median x_{2jt} first choice j to a below-median x_{2kt} second choice k.

Figure 4: Second Choice Micro Moment Correlations

This figure reports median absolute correlations between different micro statistics over 1,000 simulated micro datasets underlying the second choice moments in Table 9. For each micro observation n in market $t_n = t$ of type $i_n = i$ with choices $j_n = j$ and $k_n = k$, we compute four statistics: $x_{2jt} \cdot x_{2k(-j)t}$ captures variation in "$\mathbb{C}(x_{2jt}, x_{2k(-j)t} \mid j, k \neq 0)$" moments, $x_{2jt} + x_{2k(-j)t}$ captures variation in "$\mathbb{E}[x_{2jt} + x_{2k(-j)t} \mid j, k \neq 0]$" moments, $1\{x_{2jt} < \bar{x}_{2t}\} \cdot 1\{x_{2k(-j)t} < \bar{x}_{2t}\}$ captures variation in "$\mathbb{P}(x_{2k(-j)t} < \bar{x}_{2t} \mid x_{2jt} < \bar{x}_{2t}, j, k \neq 0)$" moments, and $\partial \log \mathbb{P}_A(t, j, k, y_{it} \mid n \in \mathcal{N}_d)/\partial \sigma_x$ is the score for σ_x at the true θ_0.

Figure 5: Petrin (2002) Replication Code

```python
import numpy as np
import pandas as pd
from pyblp import data, Problem, Formulation, MicroDataset, MicroPart, MicroMoment, Optimization, Iteration

# Configure the aggregate problem: linear demand ("X1"), nonlinear demand ("X2"), marginal costs ("X3"), and demographics
problem = Problem(
    product_formulations=[
        Formulation('1 + hpwt + space + air + mpd + fwd + mi + sw + su + pv + pgnp + trend + trend2'),
        Formulation('1 + I(-prices) + hpwt + space + air + mpd + fwd + mi + sw + su + pv'),
        Formulation('1 + log(hpwt) + log(wt) + log(mpg) + air + fwd + trend * (jp + eu) + log(q)'),
    ],
    costs_type='log',
    agent_formulation=Formulation('1 + I(low / income) + I(mid / income) + I(high / income) + I(log(fs) * fv) + age + fs + mid + high'),
    product_data=pd.read_csv(data.PETRIN_PRODUCTS_LOCATION),
    agent_data=pd.read_csv(data.PETRIN_AGENTS_LOCATION),
)

# Configure the micro dataset: name, number of observations, and a function that computes sampling weights
micro_dataset = MicroDataset("CEX", 29125, lambda t, p, a: np.ones((a.size, 1 + p.size)))

# Configure micro moment parts: names, datasets, and functions that compute micro values
age_mi_part      = MicroPart("E[age_i * mi_j]",      micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 5], np.r_[0, p.X2[:,  7]]))
age_sw_part      = MicroPart("E[age_i * sw_j]",      micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 5], np.r_[0, p.X2[:,  8]]))
age_su_part      = MicroPart("E[age_i * su_j]",      micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 5], np.r_[0, p.X2[:,  9]]))
age_pv_part      = MicroPart("E[age_i * pv_j]",      micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 5], np.r_[0, p.X2[:, 10]]))
fs_mi_part       = MicroPart("E[fs_i * mi_j]",       micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 6], np.r_[0, p.X2[:,  7]]))
fs_sw_part       = MicroPart("E[fs_i * sw_j]",       micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 6], np.r_[0, p.X2[:,  8]]))
fs_su_part       = MicroPart("E[fs_i * su_j]",       micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 6], np.r_[0, p.X2[:,  9]]))
fs_pv_part       = MicroPart("E[fs_i * pv_j]",       micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 6], np.r_[0, p.X2[:, 10]]))
inside_mid_part  = MicroPart("E[1{j > 0} * mid_i]",  micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 7], np.r_[0, p.X2[:,  0]]))
inside_high_part = MicroPart("E[1{j > 0} * high_i]", micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 8], np.r_[0, p.X2[:,  0]]))
mi_part          = MicroPart("E[mi_j]",              micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 0], np.r_[0, p.X2[:,  7]]))
sw_part          = MicroPart("E[sw_j]",              micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 0], np.r_[0, p.X2[:,  8]]))
su_part          = MicroPart("E[su_j]",              micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 0], np.r_[0, p.X2[:,  9]]))
pv_part          = MicroPart("E[pv_j]",              micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 0], np.r_[0, p.X2[:, 10]]))
mid_part         = MicroPart("E[mid_i]",             micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 7], np.r_[1, p.X2[:,  0]]))
high_part        = MicroPart("E[high_i]",            micro_dataset, lambda t, p, a: np.outer(a.demographics[:, 8], np.r_[1, p.X2[:,  0]]))

# Configure micro moments: names, observed values, parts, and functions that combine parts
compute_ratio = lambda v: v[0] / v[1]
compute_ratio_gradient = lambda v: [1 / v[1], -v[0] / v[1]**2]
micro_moments = [
    MicroMoment("E[age_i | mi_j]",      0.783,  [age_mi_part,      mi_part],   compute_ratio, compute_ratio_gradient),
    MicroMoment("E[age_i | sw_j]",      0.730,  [age_sw_part,      sw_part],   compute_ratio, compute_ratio_gradient),
    MicroMoment("E[age_i | su_j]",      0.740,  [age_su_part,      su_part],   compute_ratio, compute_ratio_gradient),
    MicroMoment("E[age_i | pv_j]",      0.652,  [age_pv_part,      pv_part],   compute_ratio, compute_ratio_gradient),
    MicroMoment("E[fs_i | mi_j]",       3.86,   [fs_mi_part,       mi_part],   compute_ratio, compute_ratio_gradient),
    MicroMoment("E[fs_i | sw_j]",       3.17,   [fs_sw_part,       sw_part],   compute_ratio, compute_ratio_gradient),
    MicroMoment("E[fs_i | su_j]",       2.97,   [fs_su_part,       su_part],   compute_ratio, compute_ratio_gradient),
    MicroMoment("E[fs_i | pv_j]",       3.47,   [fs_pv_part,       pv_part],   compute_ratio, compute_ratio_gradient),
    MicroMoment("E[1{j > 0} | mid_i]",  0.0794, [inside_mid_part,  mid_part],  compute_ratio, compute_ratio_gradient),
    MicroMoment("E[1{j > 0} | high_i]", 0.1581, [inside_high_part, high_part], compute_ratio, compute_ratio_gradient),
]

# Configure two-step minimum distance: starting values, numerical optimization, clustered aggregate moments, and micro moments
problem_results = problem.solve(
    sigma=np.diag([3.23, 0, 4.43, 0.46, 0.01, 2.58, 4.42, 0, 0, 0, 0]),
    pi=np.array([
        [0, 0,    0,     0,     0,    0, 0, 0, 0],
        [0, 7.52, 31.13, 34.49, 0,    0, 0, 0, 0],
        [0, 0,    0,     0,     0,    0, 0, 0, 0],
        [0, 0,    0,     0,     0,    0, 0, 0, 0],
        [0, 0,    0,     0,     0,    0, 0, 0, 0],
        [0, 0,    0,     0,     0,    0, 0, 0, 0],
        [0, 0,    0,     0,     0,    0, 0, 0, 0],
        [0, 0,    0,     0,     0.57, 0, 0, 0, 0],
        [0, 0,    0,     0,     0.28, 0, 0, 0, 0],
        [0, 0,    0,     0,     0.31, 0, 0, 0, 0],
        [0, 0,    0,     0,     0.42, 0, 0, 0, 0],
    ]),
    optimization=Optimization('bfgs', {'gtol': 1e-4}),
    iteration=Iteration('squarem', {'atol': 1e-13}),
    se_type='clustered',
    W_type='clustered',
    micro_moments=micro_moments,
)
```

This Python code demonstrates how to construct and solve the problem from Petrin (2002) with PyBLP. Names in the formulation objects correspond to variable names in the datasets, which are packaged with PyBLP. Micro moment values are from Table 6a in the working paper version of Petrin (2002). We report replication results from running this code in the rightmost column of Table 10.

Table 10: Petrin (2002) Replication

| | | | | Replicated with Different Micro Covariances | |
		Logit	Published	Micro Data	Estimated
Price Coefficients	Low Income	0.13	7.52	3.81	3.86
		(0.01)	(1.24)	(0.36)	(0.36)
	Middle Income	0.13	31.13	11.93	12.06
		(0.01)	(4.07)	(1.00)	(1.01)
	High Income	0.13	34.49	23.56	23.79
		(0.01)	(2.56)	(2.43)	(2.40)
Base Coefficients	Constant	-10.05	-15.67	-8.81	-8.91
		(0.34)	(4.39)	(1.39)	(1.42)
	Horsepower/Weight	3.79	-2.83	8.42	8.34
		(0.47)	(8.16)	(2.27)	(2.40)
	Size	3.25	4.80	4.93	4.89
		(0.27)	(3.57)	(1.71)	(1.61)
	Air Conditioning Standard	0.22	3.88	3.59	3.81
		(0.08)	(2.21)	(1.24)	(1.22)
	Miles/Dollar	0.05	-15.79	-0.13	-0.14
		(0.06)	(0.87)	(0.33)	(0.32)
	Front Wheel Drive	0.15	-12.32	-6.48	-6.45
		(0.06)	(2.36)	(1.83)	(1.81)
	Minivan	-0.10	-5.65	-1.98	-2.10
		(0.15)	(0.68)	(0.46)	(0.48)
	Station Wagon	-1.12	-1.31	-1.31	-1.33
		(0.06)	(0.36)	(0.25)	(0.20)
	Sport-utility	-0.62	-4.38	-1.08	-1.08
		(0.11)	(0.41)	(0.29)	(0.28)
	Full-size Van	-1.89	-5.26	-3.34	-3.31
		(0.13)	(1.30)	(0.57)	(0.52)
	Percent Change in GNP	0.04	0.24	0.03	0.03
		(0.01)	(0.02)	(0.01)	(0.01)
Random Coefficients	Constant		3.23	-0.00	0.03
			(0.72)	(0.54)	(0.53)
	Horsepower/Weight		4.43	0.03	0.12
			(1.60)	(0.83)	(0.81)
	Size		0.46	-0.12	-0.09
			(1.07)	(0.68)	(0.61)
	Air Conditioning Standard		0.01	-1.16	-1.33
			(0.78)	(1.03)	(1.09)
	Miles/Dollar		2.58	-0.16	-0.16
			(0.14)	(0.22)	(0.22)
	Front Wheel Drive		4.42	1.62	1.62
			(0.79)	(0.37)	(0.37)
	Minivan		0.57	0.40	0.42
			(0.10)	(0.05)	(0.05)
	Station Wagon		0.28	0.16	0.17
			(0.09)	(0.06)	(0.04)

Continued on the next page.

		Logit	Published	Micro Data	Estimated
	Sport-utility		0.31	0.10	0.10
			(0.09)	(0.06)	(0.05)
	Full-size Van		0.42	0.25	0.25
			(0.21)	(0.10)	(0.08)
Cost	Constant		1.50	1.38	1.40
Coefficients			(0.08)	(0.14)	(0.14)
	log(Horsepower/Weight)		0.84	0.88	0.88
			(0.03)	(0.05)	(0.05)
	log(Weight)		1.28	1.42	1.41
			(0.04)	(0.08)	(0.08)
	log(Miles/Gallon)		0.23	0.13	0.12
			(0.04)	(0.06)	(0.06)
	Air Conditioning Standard		0.24	0.27	0.27
			(0.01)	(0.02)	(0.02)
	Front Wheel Drive		0.01	0.07	0.07
			(0.01)	(0.02)	(0.02)
	Trend		-0.01	-0.01	-0.01
			(0.01)	(0.00)	(0.00)
	Japan		0.12	0.10	0.10
			(0.01)	(0.03)	(0.03)
	Japan $\times$ Trend		-0.01	0.00	0.00
			(0.01)	(0.00)	(0.00)
	Europe		0.47	0.46	0.46
			(0.03)	(0.04)	(0.04)
	Europe $\times$ Trend		-0.01	-0.01	-0.01
			(0.01)	(0.00)	(0.00)
	log(Quantity)		-0.05	-0.07	-0.07
			(0.01)	(0.01)	(0.01)
Micro	"$\mathbb{P}(\text{Middle Age}_{it} \mid \text{Minivan}_{jt}) = 0.783$"		0.750	0.749	0.754
Moments	"$\mathbb{P}(\text{Middle Age}_{it} \mid \text{Station Wagon}_{jt}) = 0.730$"		0.675	0.677	0.683
	"$\mathbb{P}(\text{Middle Age}_{it} \mid \text{Sport-utility}_{jt}) = 0.740$"		0.663	0.680	0.681
	"$\mathbb{P}(\text{Middle Age}_{it} \mid \text{Full-size Van}_{jt}) = 0.652$"		0.725	0.730	0.729
	"$\mathbb{E}[\text{Family Size}_{it} \mid \text{Minivan}_{jt}] = 3.86$"		3.85	3.83	3.87
	"$\mathbb{E}[\text{Family Size}_{it} \mid \text{Station Wagon}_{jt}] = 3.17$"		3.19	3.15	3.18
	"$\mathbb{E}[\text{Family Size}_{it} \mid \text{Sport-utility}_{jt}] = 2.97$"		3.02	2.98	2.98
	"$\mathbb{E}[\text{Family Size}_{it} \mid \text{Full-size Van}_{jt}] = 3.47$"		3.44	3.51	3.49
	"$\mathbb{P}(j \neq 0 \mid \text{Middle Income}_{it}) = 0.0794$"		0.0807	0.0799	0.0799
	"$\mathbb{P}(j \neq 0 \mid \text{High Income}_{it}) = 0.1581$"		0.1596	0.1598	0.1602
Minivan	1984 Compensating Variation (Dollars, Millions)	1,240.34	367.29	429.89	425.91
Innovation		(242.46)		(250.10)	(224.57)

This table reports replication results for Petrin (2002) described in Section 8. From left to right, we report our exactly replicated IV logit estimates, micro BLP estimates from the original paper, replication results with micro moment covariances estimated from the micro data, and results with covariances estimated by PyBLP so the only micro statistics needed are the values in Figure 5. Standard errors are in parentheses; we compute those for the minivan innovation counterfactual with a parametric bootstrap.

			Micro Moments		
		Aggregate	Standard	Diversion	Optimal
Price/Ounce Coefficients	Constant	-52.645	-52.343	-38.538	-37.902
		(4.660)	(4.694)	(4.034)	(4.217)
	High Income Household		3.549	3.178	4.062
			(0.940)	(1.046)	(0.992)
	Child in Household		-6.915	-8.119	-11.105
			(1.274)	(1.406)	(1.458)
	Unobserved Preference	19.631	19.229	15.256	14.941
		(1.802)	(1.805)	(2.569)	(2.749)
Inside Goods Coefficients	High Income Household		-0.053	0.348	-0.278
			(0.040)	(0.130)	(0.120)
	Child in Household		0.498	1.210	1.884
			(0.050)	(0.239)	(0.355)
	Unobserved Preference			4.964	5.178
				(0.387)	(0.410)
Diet Formula Coefficients	High Income Household		0.708	0.999	0.684
			(0.043)	(0.142)	(0.124)
	Child in Household		-0.852	-1.463	-1.037
			(0.056)	(0.271)	(0.216)
	Unobserved Preference			2.606	2.671
				(0.868)	(0.950)
Small Sized Coefficients	High Income Household		-0.690	-0.710	-0.662
			(0.060)	(0.061)	(0.058)
	Child in Household		0.689	0.716	0.641
			(0.069)	(0.071)	(0.066)
Standard Micro Statistics	"$\mathbb{P}(\text{High}_{it} \mid j \neq 0) = 0.597$"		0.597	0.597	0.561
	"$\mathbb{P}(\text{Child}_{it} \mid j \neq 0) = 0.203$"		0.203	0.203	0.228
	"$\mathbb{C}(\text{Price}_{jt}, \text{High}_{it} \mid j \neq 0) = -0.0004$"		-0.0004	-0.0004	-0.0002
	"$\mathbb{C}(\text{Price}_{jt}, \text{Child}_{it} \mid j \neq 0) = -0.0001$"		-0.0001	-0.0001	-0.0004
	"$\mathbb{C}(\text{Diet}_{jt}, \text{High}_{it} \mid j \neq 0) = 0.0355$"		0.0355	0.0355	0.0220
	"$\mathbb{C}(\text{Diet}_{jt}, \text{Child}_{it} \mid j \neq 0) = -0.0264$"		-0.0264	-0.0264	-0.0172
	"$\mathbb{C}(\text{Small}_{jt}, \text{High}_{it} \mid j \neq 0) = -0.0207$"		-0.0207	-0.0207	-0.0192
	"$\mathbb{C}(\text{Small}_{jt}, \text{Child}_{it} \mid j \neq 0) = 0.0116$"		0.0116	0.0116	0.0089
Diversion Micro Statistics	"$\mathbb{P}(k(\text{-}b(j)) = 0 \mid \text{Surveyed Non-diet}_{jt}) = 0.16$"	0.92	0.93	0.16	0.14
	"$\mathbb{P}(\text{Diet}_{k(\text{-}b(j))t} \mid \text{Surveyed Non-diet}_{jt}) = 0.17$"	0.03	0.03	0.17	0.17

		Aggregate	Standard	Diversion	Optimal
Aggregate	Product-Retailer-Quarters	78,161	78,161	78,161	78,161
Observations	↪ Products	2,672	2,672	2,672	2,672
	↪ Retailers	5	5	5	5
	↪ Quarters (Markets)	40	40	40	40
	↪ Brands (Clusters)	425	425	425	425
Fixed	Product-Retailers	5,815	5,815	5,815	5,815
Effects	Retailer-Quarters	200	200	200	200
Micro	Grocery Trips		10,455	10,455	10,455
Observations	↪ Household-Years		1,130	1,130	1,130
	↪ Survey Years		10	10	10
	Second Choice Responses			100	100
Tax	Weighted Average Taxed Elasticity	-1.354	-1.349	-1.327	-1.320
Counterfactual		(0.064)	(0.065)	(0.090)	(0.095)
	Taxed Volume Change (%)	-30.095	-29.973	-15.870	-15.652
		(1.439)	(1.452)	(1.676)	(1.487)
	↪ Low Income Households		-31.134	-16.472	-16.289
			(1.435)	(1.941)	(1.635)
	↪ High − Low Income		1.967	1.026	1.104
			(0.778)	(0.895)	(0.615)
	↪ Households without Children		-28.790	-15.282	-14.508
			(1.471)	(1.760)	(1.626)
	↪ With − without Children		-4.891	-2.475	-4.793
			(0.915)	(1.492)	(2.218)
	Untaxed Volume Change (%)	0.872	0.835	9.238	9.383
		(0.036)	(0.039)	(3.430)	(3.490)
	↪ Low Income Households		0.948	12.782	11.643
			(0.040)	(4.000)	(3.919)
	↪ High − Low Income		-0.163	-5.114	-3.522
			(0.039)	(1.370)	(0.850)
	↪ Households without Children		0.736	8.096	8.453
			(0.035)	(3.256)	(3.246)
	↪ With − without Children		0.620	7.234	4.740
			(0.064)	(1.540)	(1.955)

This table reports results for the empirical example described in Section 8. From left to right, we report estimates using aggregate moments, adding standard micro moments, adding second choice moments, and replacing standard micro moments with the optimal micro moments described in Section 6. Standard errors are clustered by brand for the aggregate moments and are in parentheses; we compute those for tax counterfactual with a parametric bootstrap.

References

Ackerberg, D. A. and M. Rysman (2005). Unobserved product differentiation in discrete-choice models: Estimating price elasticities and welfare effects. *RAND Journal of Economics 36*(4), 771–789.

Agarwal, N. and P. Somaini (2020). Revealed preference analysis of school choice models. *Annual Review of Economics 12*, 471–501.

Allcott, H., B. B. Lockwood, and D. Taubinsky (2019). Regressive sin taxes, with an application to the optimal soda tax. *The Quarterly Journal of Economics 134*(3), 1557–1626.

Allenby, G. M., N. Hardt, and P. E. Rossi (2019). Economic foundations of conjoint analysis. In *Handbook of the Economics of Marketing*, Volume 1, pp. 151–192. Elsevier.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics 14*(1), 57–67.

Armitage, S. and F. Pinter (2022). Regulatory mandates and electric vehicle product variety. Working paper.

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

Backus, M., C. Conlon, and M. Sinkinson (2021). Common ownership and competition in the ready-to-eat cereal industry. Working paper.

Barker, A. R., S. Mazzucca, and R. An (2022). The impact of sugar-sweetened beverage taxes by household income: A multi-city comparison of Nielsen purchasing data. *Nutrients 14*(5), 922.

Barwick, P. J., S. Cao, and S. Li (2017). Local protectionism, market structure, and social welfare: China's automobile market. Working paper.

Beresteanu, A. and S. Li (2011). Gasoline prices, government support, and the demand for hybrid vehicles in the United States. *International Economic Review 52*(1), 161–182.

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

Berry, S. T. and P. A. Haile (2014). Identification in differentiated products markets using market level data. *Econometrica 82*(5), 1749–1797.

Berry, S. T. and P. A. Haile (2021). Foundations of demand estimation. In *Handbook of Industrial Organization*, Volume 4, Chapter 1, pp. 1–62. Elsevier.

Berry, S. T. and P. A. Haile (2022). Nonparametric identification of differentiated products demand using micro data. Working paper.

Berry, S. T., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.

Berry, S. T., J. Levinsohn, and A. Pakes (1999). Voluntary export restraints on automobiles: Evaluating a trade policy. *American Economic Review 89*(3), 400–430.

Berry, S. T., J. Levinsohn, and A. Pakes (2004). Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy 112*(1), 68–105.

Berry, S. T., O. B. Linton, and A. Pakes (2004). Limit theorems for estimating the parameters of differentiated product demand systems. *The Review of Economic Studies 71*(3), 613–654.

Berry, S. T. and A. Pakes (2007). The pure characteristics demand model. *International Economic Review 48*(4), 1193–1225.

Billingsley, P. (1995). *Probability and measure* (3 ed.). John Wiley & Sons.

Bodéré, P. (2023). Dynamic spatial competition in early education: An equilibrium analysis of the preschool market in Pennsylvania. Working paper.

Brenkers, R. and F. Verboven (2006). Liberalizing a distribution system: The European car market. *Journal of the European Economic Association 4*(1), 216–251.

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

Cawley, J., D. Frisvold, A. Hill, and D. Jones (2019). The impact of the philadelphia beverage tax on purchases and consumption by adults and children. *Journal of Health Economics 67*, 102225.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Chen, J., C. Reinhardt, and S. A. Syed Shah (2022). Substitution patterns and welfare implications of local taxation: Empirical analysis of a soda tax. Working paper.

Ching, A. T., F. Hayashi, and H. Wang (2015). Quantifying the impacts of limited supply: The case of nursing homes. *International Economic Review 56*(4), 1291–1322.

Chintagunta, P. K. and J.-P. Dubé (2005). Estimating a stockkeeping-unit-level brand choice model that combines household panel data and store data. *Journal of Marketing Research 42*(3), 368–379.

Ciliberto, F. and N. V. Kuminoff (2010). Public policy and market competition: how the master settlement agreement changed the cigarette industry. *The B.E. Journal of Economic Analysis & Policy 10*(1), Article 63.

Conlon, C. (2013). The empirical likelihood MPEC approach to demand estimation. Working paper.

Conlon, C. and J. Gortmaker (2020). Best practices for differentiated products demand estimation with PyBLP. *RAND Journal of Economics 51*(4), 1108–1161.

Conlon, C. and J. H. Mortimer (2021). Empirical properties of diversion ratios. *The RAND Journal of Economics 52*(4), 693–726.

Conlon, C., J. H. Mortimer, and P. Sarkis (2023). Estimating preferences and substitution patterns from second choice data alone. Working paper.

Conlon, C. and N. Rao (2023). Market power as second-worst regulation: Welfare consequences of post and hold pricing in distilled spirits. Working paper.

Conlon, C., N. Rao, and Y. Wang (2022). Who pays sin taxes? understanding the overlapping burdens of corrective taxes. *The Review of Economics and Statistics*, 1–27.

Copeland, A. (2014). Intertemporal substitution and new car purchases. *RAND Journal of Economics 45*(3), 624–644.

Deaton, A. and J. Muellbauer (1980). An almost ideal demand system. *American Economic Review 70*(3), 312–326.

DellaVigna, S. and M. Gentzkow (2019). Uniform pricing in us retail chains. *The Quarterly Journal of Economics 134*(4), 2011–2084.

Döpper, H., A. MacKay, N. Miller, and J. Stiebale (2022). Rising markups and the role of consumer preferences. Working paper.

Dubé, J.-P., J. T. Fox, and C.-L. Su (2012). Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica 80*(5), 2231–2267.

Durrmeyer, I. (2022). Winners and losers: The distributional effects of the French feebate on the automobile market. *The Economic Journal 132*(644), 1414–1448.

Eyal, P., R. David, G. Andrew, E. Zak, and D. Ekaterina (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods 54*, 1–20.

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

Gandhi, A. and J.-F. Houde (2020). Measuring substitution patterns in differentiated-products industries. Working paper.

Goeree, M. S. (2008). Limited information and advertising in the us personal computer industry. *Econometrica 76*(5), 1017–1074.

Goolsbee, A. and A. Petrin (2004). The consumer gains from direct broadcast satellites and the competition with cable TV. *Econometrica 72*(2), 351–381.

Grieco, P. L. E., C. Murry, J. Pinkse, and S. Sagl (2023). Conformant and efficient estimation of discrete choice demand models. Working paper.

Grieco, P. L. E., C. Murry, and A. Yurukoglu (2021). The evolution of market power in the US auto industry. Working paper.

Han, C. and P. C. Phillips (2006). GMM with many moment conditions. *Econometrica 74*(1), 147–192.

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

Hendel, I. and A. Nevo (2006). Measuring the implications of sales and consumer inventory behavior. *Econometrica 74*(6), 1637–1673.

Hong, H., H. Li, and J. Li (2021). BLP estimation using Laplace transformation and overlapping simulation draws. *Journal of Econometrics 222*(1), 56–72.

Howell, J. R., S. Lee, and G. M. Allenby (2016). Price promotions in choice models. *Marketing Science 35*(2), 319–334.

Imbens, G. W. and T. Lancaster (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies 61*(4), 655–680.

Ketz, P. (2019). On asymptotic size distortions in the random coefficients logit model. *Journal of Econometrics 212*(2), 413–432.

Kleibergen, F. and R. Paap (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics 133*(1), 97–126.

Lee, J. and K. Seo (2015). A computationally fast estimator for random coefficients logit demand models using aggregate data. *RAND Journal of Economics 46*(1), 86–102.

Lewbel, A. and K. Pendakur (2009). Tricks with Hicks: The EASI demand system. *American Economic Review 99*(3), 827–63.

Li, S. (2012). Traffic safety and vehicle choice: quantifying the effects of the 'arms race' on American roads. *Journal of Applied Econometrics 27*(1), 34–62.

Li, S. (2018). Better lucky than rich? Welfare analysis of automobile licence allocations in Beijing and Shanghai. *The Review of Economic Studies 85*(4), 2389–2428.

Li, S., J. Xiao, and Y. Liu (2015). The price evolution in China's automobile market. *Journal of Economics & Management Strategy 24*(4), 786–810.

Li, Y., B. R. Gordon, and O. Netzer (2018). An empirical study of national vs. local pricing by chain stores under competition. *Marketing Science 37*(5), 812–837.

McFadden, D. (1978). Modeling the choice of residential location. In A. Karlqvist, L. Lundqvist, F. Snickars, and J. Wiebull (Eds.), *Spatial Interaction Theory and Planning Models*, pp. 75–96. Amsterdam: North Holland.

Montag, F. (2023). Mergers, foreign competition, and jobs: Evidence from the US appliance industry. Working paper.

Morrow, W. R. and S. J. Skerlos (2011). Fixed-point approaches to computing Bertrand-Nash equilibrium prices under mixed-logit demand. *Operations Research 59*(2), 328–345.

Murry, C. (2017). Advertising in vertical relationships: An equilibrium model of the automobile industry. Working paper.

Myojo, S. and Y. Kanazawa (2012). On asymptotic properties of the parameters of differentiated product demand and supply systems when demographically categorized purchasing pattern data are available. *International Economic Review 53*(3), 887–938.

Nakamura, E. and D. Zerom (2010). Accounting for incomplete pass-through. *The Review of Economic Studies 77*(3), 1192–1230.

Neilson, C. (2021). Targeted vouchers, competition among schools, and the academic achievement of poor students. Working paper.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Newey, W. K. and F. Windmeijer (2009). Generalized method of moments with many weak moment conditions. *Econometrica 77*(3), 687–719.

Nurski, L. and F. Verboven (2016). Exclusive dealing as a barrier to entry? Evidence from automobiles. *The Review of Economic Studies 83*(3), 1156–1188.

Oddo, V. M., J. Leider, and L. M. Powell (2021). The impact of Seattle's sugar-sweetened beverage tax on substitution to sweets and salty snacks. *The Journal of Nutrition 151*(10), 3232–3239.

Owen, A. B. (2017). A randomized Halton algorithm in R. Working paper.

Peer, E., L. Brandimarte, S. Samat, and A. Acquisti (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology 70*, 153–163.

Petrin, A. (2002). Quantifying the benefits of new products: The case of the minivan. *Journal of Political Economy 110*(4), 705–729.

Powell, L. M., J. F. Chriqui, T. Khan, R. Wada, and F. J. Chaloupka (2013). Assessing the potential effectiveness of food and beverage taxes and subsidies for improving public health: A systematic review of prices, demand and body weight outcomes. *Obesity reviews 14*(2), 110–128.

Powell, L. M. and J. Leider (2020). The impact of seattle's sweetened beverage tax on beverage prices and volume sold. *Economics & Human Biology 37*, 100856.

Powell, L. M. and J. Leider (2022). Impact of the Seattle sweetened beverage tax on substitution to alcoholic beverages. *PLoS One 17*(1), e0262578.

Reynaert, M. and F. Verboven (2014). Improving the performance of random coefficients demand models: The role of optimal instruments. *Journal of Econometrics 179*(1), 83–98.

Reynolds, G. and C. Walters (2008). The use of customer surveys for market definition and the competitive assessment of horizontal mergers. *Journal of Competition Law and Economics 4*(2), 411–431.

Roberto, C. A., H. G. Lawman, M. T. LeVasseur, N. Mitra, A. Peterhans, B. Herring, and S. N. Bleich (2019). Association of a beverage tax on sugar-sweetened and artificially sweetened beverages with changes in beverage prices and sales at chain retailers in a large urban setting. *JAMA 321*(18), 1799–1810.

Rogers, N. T., S. Cummins, H. Forde, C. P. Jones, O. Mytton, H. Rutter, S. J. Sharp, D. Theis, M. White, and J. Adams (2023). Associations between trajectories of obesity prevalence in English primary school children and the UK soft drinks industry levy: An interrupted time series analysis of surveillance data. *PLoS Medicine 20*(1), e1004160.

Salanié, B. and F. A. Wolak (2022). Fast, detail-free, and approximately correct: Estimating mixed demand systems. Working paper.

Seiler, S., A. Tuchman, and S. Yao (2021). The impact of soda taxes: Pass-through, tax avoidance, and nutritional effects. *Journal of Marketing Research 58*(1), 22–49.

Stantcheva, S. (2023). How to run surveys: A guide to creating your own identifying variation and revealing the invisible. *Annual Review of Economics 15*, 205–234.

Starc, A. (2014). Insurer pricing and consumer welfare: Evidence from Medigap. *RAND Journal of Economics 45*(1), 198–220.

Su, C.-L. and K. L. Judd (2012). Constrained optimization approaches to estimation of structural models. *Econometrica 80*(5), 2213–2230.

Thomadsen, R. (2005). The effect of ownership structure on prices in geographically differentiated industries. *RAND Journal of Economics 36*(4), 908–929.

Tiago de Oliveira, J. (1959). Extremal distributions. *Revista de Faculdada du Ciencia, Lisboa, Serie A 7*, 215—227.

Varadhan, R. and C. Roland (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics 35*(2), 335–353.

Weber, S. (2022). Undervaluation of future fuel savings and efficiency standards for heavy-duty trucks. Working paper.

Wedel, M. and R. Pieters (2000). Eye fixations on advertisements and memory for brands: A model and findings. *Marketing Science 19*(4), 297–312.

Wollmann, T. G. (2018). Trucks without bailouts: Equilibrium product characteristics for commercial vehicles. *American Economic Review 108*(6), 1364–1406.

Zhen, C., E. A. Finkelstein, J. M. Nonnemaker, S. A. Karns, and J. E. Todd (2014). Predicting the effects of sugar-sweetened beverage taxes on food and beverage demand in a large demand system. *American Journal of Agricultural Economics 96*(1), 1–25.

A. Lognormal Price Coefficient

Perhaps the most popular variant of the BLP-style model in Section 2 is one in which we replace the random coefficient β_{cit} on price $x_{cjt} = p_{jt}$ with a negative lognormal random coefficient:

$$\alpha_{cit} = -\exp(\Pi_c y_{it} + \Sigma_c \nu_{it}) < 0, \quad \nu_{it} \sim N(0, I). \tag{A1}$$

One reason this variant is popular is that it guarantees downward-sloping demand for all consumers. This is in contrast to a normally distributed α_{cjt}, which would give upward-sloping demand for consumer types with very positive unobserved preferences ν_{cit} for price.

When using this variant, which is fully supported by PyBLP, one drops the linear coefficient in β on price and instead estimates a nonlinear coefficient in Π on a constant demographic to shift the level of α_{cit}. Typically, one may also wish to estimate another nonlinear coefficient in Π on income to reflect wealth effects, and potentially a third nonlinear coefficient in Σ to reflect unobserved heterogeneity in price sensitivity that is not accounted for by measured income. In addition to an instrument for endogenous prices such as a cost shifter, one will also want to interact this with mean income and build a differentiation IV from it to have three instruments for the three parameters that govern the distribution of α_{cit}.

It is also worth noting that although the FRAC approach of Salanié and Wolak (2022) discussed in Section 3 and Appendix C is "robust" to the distribution of random coefficients in the sense that it delivers an approximate estimator of the mean and variance of a lognormal random coefficient,[96] it will not deliver approximate estimating equation that is linear in $\Sigma_c \Sigma_c'$ and Π_c because these will enter nonlinearly into the lognormal mean and variance.[97] A practical recommendation is to use the non-lognormal FRAC expressions to get a sense for non-lognormal parameter values, and to convert approximate estimators into "lognormal space" using the expressions for the mean and variance of the lognormal distribution.

Intuition from Scores

To build intuition about which micro moments may help estimate the parameters that govern a lognormal random price coefficient, consider the simplest case with $C = 1$ observed characteristic, price p_{jt}; $R = 1$ demographic, income y_{it}; no unobserved heterogeneity, $\Sigma = 0$;

---

[96]Salanié and Wolak (2022) discuss this extension in their Monte Carlo simulations.

[97]Replacing x_{jt} with p_{jt} in the simple example in (8) gives $\log(s_{jt}/s_{0t}) \approx m_t^\alpha p_{jt} + v_t^\alpha a_{jt} + \xi_{jt}$ where the mean of the lognormal random coefficient is $m_t^\alpha = -\exp[\beta + \pi m_t^y + (\sigma^2 + \pi^2 v_t^y)/2]$ and its variance is $v_t^\alpha = [\exp(\sigma^2 + \pi^2 v_t^y) - 1]\exp[2(\beta + \pi m_t^y) + (\sigma^2 + \pi^2 v_t^y)]$.

and a micro dataset d with no selection, $w_{dijt} = 1$. The score for π in $\alpha_{it} = \exp(\pi y_{it})$ is the same as in (24), but the derivative of indirect utility for $j \neq 0$ with respect to π is now

$$\frac{\partial u_{ijt}}{\partial \pi} = \frac{\partial \mu_{ijt}}{\partial \pi} + \frac{\partial \delta_{jt}}{\partial \pi} = p_{jt} \cdot \alpha_{it} \cdot y_{it} + \frac{\partial \delta_{jt}}{\partial \pi}. \tag{A2}$$

Compared to the linear random coefficient case in (25), the first term is now scaled by α_{it}.

A first-order Taylor approximation around $\pi = 0$ gives $p_{jt} \cdot \alpha_{it} \cdot y_{it} \approx p_{jt} \cdot y_{it} + \pi \cdot p_{jt} \cdot y_{it}^2$. This suggests that the most important moment to match is again "$\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$." To the extent that higher-order terms explain more variation in the score, it may also help to match a "$\mathbb{C}(p_{jt}, y_{it}^2 \mid j \neq 0)$" moment, if available. Moments of the form "$\mathbb{E}[p_{jt} \mid y_{it} < \bar{y}_t, j \neq 0]$" are more likely to be collected by surveys, and may also help to better span the curvature in the score introduced by the lognormal coefficient.

Monte Carlo Results

To illustrate the performance of the micro BLP estimator with a lognormal price coefficient, we modify our Monte Carlo configuration described in Section 7 with unobserved heterogeneity for x_{2jt} to instead have a lognormal random coefficient on price:

$$\mu_{ijt} = \pi_1 y_{it} + \alpha_{it} p_{jt}, \quad \alpha_{it} = -\exp(\alpha + \pi_p y_{it} + \sigma_p \nu_{3it}), \tag{A3}$$

in which we draw ν_{3it} from the standard normal distribution, and use 1,000 scrambled Halton draws (Owen, 2017) to approximate this distribution during estimation.

In Table A1 we illustrate the impact of the micro moments discussed above. Without any micro moments, $\hat{\pi}_1$ and $\hat{\pi}_p$ have substantial variance because they are identified only from limited cross-market variation in the distribution of income. This only slightly contaminates the estimator of σ_p, which is otherwise well-estimated because our default configuration has a great deal of cross-market choice set variation.

Like with a normally distributed coefficient, matching "$\mathbb{E}[y_{it} \mid j \neq 0]$" and "$\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$" substantially reduces the variance of the estimators. Incorporating a "$\mathbb{C}(p_{jt}, y_{it}^2 \mid j \neq 0)$" moment does not seem to be particularly helpful, suggesting that at least in this simulation, the first term in a Taylor approximation to the score explains most of the variation. Finally, optimal micro moments that require the full micro dataset reduce the bias and variance of the estimator even more.

Table A1: Lognormal Price Coefficient

Micro Moments Shorthand	Optimal	MAE (%)			Bias (%)		
		$\hat{\pi}_1$	$\hat{\pi}_p$	$\hat{\sigma}_p$	$\hat{\pi}_1$	$\hat{\pi}_p$	$\hat{\sigma}_p$
No Micro Moments		809.8	195.4	3.6	26.2	-24.8	-0.8
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$"		47.7	12.8	3.0	-4.4	2.0	-0.0
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$, $\mathbb{C}(p_{jt}, y_{it}^2 \mid j \neq 0)$"		47.4	13.0	3.0	-3.8	1.9	-0.1
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$, $\mathbb{C}(p_{jt}, y_{it}^2 \mid j \neq 0)$"	Yes	37.0	11.0	3.1	1.1	0.8	-0.1

This table reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets for different combinations of micro moments with a lognormal price coefficient.

B. Nested Logit and RCNL

Another popular variant of the BLP-style model in Section 2 is one in which we replace type I extreme value idiosyncratic preferences ε_{ijt} with those that follow the assumptions of a two-level nested logit (McFadden, 1978; Cardell, 1997). The resulting random coefficients nested logit (RCNL) model of Brenkers and Verboven (2006), which is fully supported by PyBLP, is popular in applications where the most important characteristic governing substitution is categorical, with categories or nests $h \in \mathcal{H}_t$ in each market $t \in \mathcal{T}$. Each product $j \in \mathcal{J}_t$ is in nest $h(j) \in \mathcal{H}_t$ and each nest $h \in \mathcal{H}_t$ contains products $\mathcal{J}_{ht} \subset \mathcal{J}_t$.

Within-category correlation of ε_{ijt} is governed by a new parameter ρ.[98] PyBLP also supports assigning a different parameter ρ_h to each category h, but for notational simplicity, we focus on the case here with a common nesting parameter. The nested logit probability that a consumer of type $i \in \mathcal{I}_t$ chooses a product $j \in \mathcal{J}_t$ is[99]

$$s_{ijt} = s_{ih(j)t} \cdot s_{ijt|h(j)} = \frac{\exp[(1 - \rho) \cdot IV_{ih(j)t}]}{1 + \sum_{h \in \mathcal{H}_t} \exp[(1 - \rho) \cdot IV_{iht}]} \cdot \frac{\exp V_{ijt}}{\exp IV_{ih(j)t}}, \tag{B1}$$

in which $V_{ijt} = (\delta_{jt} + \mu_{ijt})/(1 - \rho)$ is the scaled non-idiosyncratic indirect utility from j and $IV_{iht} = \log \sum_{j \in \mathcal{J}_{ht}} \exp V_{ijt}$ is McFadden's (1978) "inclusive value" of h.

In Conlon and Gortmaker (2020) we discuss best practices for this extension to the aggregate estimator.[100] Salanié and Wolak (2022) extend their FRAC approximation discussed in Section 3 and Appendix C to the RCNL case. Although these results could be used to extend our full FRAC expression in Appendix C, the resulting estimating equation would be less interpretable and nonlinear in ρ, so we do not derive this extension ourselves.

Ackerberg and Rysman (2005) study identification of the nested logit model and illustrate how the nested logit structure gives rise to two key sources of aggregate variation that can each identify ρ: cross-category switching due to changes in product characteristics x_{jt} and cross-category switching due to changes in the number of products $J_{ht} = |\mathcal{J}_{ht}|$. Although it is convenient that ρ can be identified from different sources of aggregate variation, Ackerberg and Rysman (2005) also argue that the strong restrictions imposed by the nested logit structure that lead to this behavior can be undesirable. Their solution is to include the number of products in x_{jt} so that it is clear what variation identifies ρ. An alternative

---

[98]Within nest $h(j) = h(k) = h$, the correlation between ε_{ijt} and ε_{ikt} was originally derived by Tiago de Oliveira (1959) to be $1 - (1 - \rho)^2$.

[99]We impose the same normalization $\delta_{0t} = \mu_{i0t} = 0$ and put the outside alternative in its own nest $h(0) = 0$. Its inclusive value is then $IV_{i0t} = 0$, giving the one in the denominator of $s_{ih(j)t}$.

[100]We use slightly different notation here to make score expressions less cumbersome.

approach would be to match only micro statistics, discussed below, that the researcher believes are intuitive sources of identifying variation.

Intuition from Scores

One concern is that with a nesting structure, intuition about the score for Π from Section 6 may be different. It turns out that this is not the case. The score for a scalar $\Pi = \pi$ is

$$
\frac{\partial \log \mathbb{P}_{\mathrm{A}}(t_n = t,\, j_n = j,\, y_{i_n t_n} = y_{it} \mid n \in \mathcal{N}_d)}{\partial \pi}
$$
$$
= \frac{\partial V_{ijt}}{\partial \pi} - \left(\rho \sum_{k \in \mathcal{J}_{h(j)t}} s_{ikt|h(j)} \cdot \frac{\partial V_{ikt}}{\partial \pi} + (1 - \rho) \sum_{k \in \mathcal{J}_t} s_{ikt} \cdot \frac{\partial V_{ikt}}{\partial \pi} \right), \tag{B2}
$$

in which the derivative of the scaled value with respect to π is

$$
\frac{\partial V_{ijt}}{\partial \pi} = \left(x_{jt} \cdot y_{it} + \frac{\partial \delta_{jt}}{\partial \pi} \right) / (1 - \rho). \tag{B3}
$$

These expressions are very similar to their non-nested counterparts in (24) and (25). The only term directly observed in the micro data is $x_{jt} \cdot y_{it}$, suggesting that the standard "$\mathbb{C}(x_{jt}, y_{it} \mid j \neq 0)$" moment should still be very informative about π. In the next subsection, we confirm this intuition in a small Monte Carlo experiment.

To build intuition about which micro moments will be most useful for estimating the new parameter ρ, consider the case without any unobserved heterogeneity, $\Sigma = 0$, and a micro dataset d with no selection, $w_{dijt} = 1$. The score for ρ is

$$
\frac{\partial \log \mathbb{P}_{\mathrm{A}}(t_n = t,\, j_n = j,\, y_{i_n t_n} = y_{it} \mid n \in \mathcal{N}_d)}{\partial \rho} \tag{B4}
$$
$$
= \frac{\partial V_{ijt}}{\partial \rho} - IV_{ih(j)t} - \left(\rho \sum_{k \in \mathcal{J}_{h(j)t}} s_{ikt|h(j)} \cdot \frac{\partial V_{ikt}}{\partial \rho} + (1 - \rho) \sum_{k \in \mathcal{J}_t} s_{ikt} \cdot \frac{\partial V_{ikt}}{\partial \rho} - \sum_{h \in \mathcal{H}_t} s_{iht} \cdot IV_{iht} \right).
$$

Noting the similarity to the score for π in (24), we again focus on the first couple of terms. The derivative of the scaled value with respect to ρ is

$$
\frac{\partial V_{ijt}}{\partial \rho} = \left(V_{ijt} + \frac{\partial \delta_{jt}}{\partial \rho} \right) / (1 - \rho). \tag{B5}
$$

Intuitively, V_{ijt} will correlate with particularly important parts of utility, suggesting that most types of micro moments discussed so far may help at least a little to estimate ρ. The derivative $\frac{\partial \delta_{jt}}{\partial \rho}$ can in general be quite complicated, but without any heterogeneity,

$\Pi = \Sigma = 0$, it simplifies to $\frac{\partial \delta_{jt}}{\partial \rho} = -\log s_{jt|h(j)}$ where $s_{jt|h(j)}$ is the share of j within its category $h(j)$.[101] Finally, $IV_{ih(j)t}$ is the inclusive or expected value from all products in $h(j)$. Both $s_{jt|h(j)}$ and $IV_{ih(j)t}$ will correlate with the category size $J_{h(j)t}$. However, in the spirit of Ackerberg and Rysman (2005), we may not believe that such moments are particularly credible sources of identification.

Unlike Π, micro data that only links demographics to first choices does not appear to be particularly useful for estimating ρ because such demographics do not explicitly show up in its score. Like Σ, however, second choice data is potentially more promising. This is not surprising—a nesting structure is the same as including random coefficients on category indicator variables with a particular distribution.

Recalling that $s_{ijkt} = s_{ik(-j)t} - s_{ikt}$, the score for ρ on a micro dataset with second choices is $s_{ik(-j)t}/s_{ijkt}$ times (B4) replacing j with k and with contributions from j removed, minus s_{ikt}/s_{ijkt} times (B4) replacing j with k and still with contributions from j. The terms that we have been focusing on are

$$\frac{\partial V_{ikt}}{\partial \rho} - \left(\frac{s_{ik(-j)t}}{s_{ijkt}} \cdot \log \sum_{\ell \in \mathcal{J}_{h(k)t} \backslash \{j\}} \exp V_{i\ell t} - \frac{s_{ikt}}{s_{ijkt}} \cdot \log \sum_{\ell \in \mathcal{J}_{h(k)t}} \exp V_{i\ell t} \right). \tag{B6}$$

The second choice score looks very different when the choices are in the same category compared to when they are in different categories. For example, when in different categories, the above expression simplifies to $\frac{\partial V_{ikt}}{\partial \rho} - IV_{ih(k)t}$. This motivates matching the share "$\mathbb{P}(h(j) = h(k) \mid j, k \neq 0)$" of individuals who would not divert to a different category, which is a fairly intuitive source of identifying variation for ρ. If using different nesting parameters for different categories, one could match a separate share for diversion from each category. Further inspection of the second choice score could yield additional moments, although they may provide less credible sources of identification.

Monte Carlo Results

To illustrate the performance of the micro BLP estimator with a nesting structure, we modify our baseline Monte Carlo configuration described in Section 7. We randomly assign each product j to one of $|\mathcal{H}_t| = 3$ categories or nests $h(j)$ with probabilities 0.1, 0.3, and 0.6, and re-draw nests until each $|\mathcal{J}_{ht}| \geq 2$. The true nesting parameter is $\rho_0 = 0.2$.

In Table B1 we illustrate the impact of the micro moments discussed above. Just like for the non-nested model, matching "$\mathbb{E}[y_{it} \mid j \neq 0]$" and "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" substantially

---

[101] For this simple nested logit model, Berry (1994) derived $\delta_{jt} = \log(s_{jt}/s_{0t}) - \rho \log s_{jt|h(j)}$.

reduces the variance of the estimators for π_1 and π_x. With substantial cross-market choice set variation in the first two rows, $\hat{\rho}$ has very low bias and variance, just like $\hat{\sigma}_x$ in our configuration with unobserved preferences for x_{2jt} in Section 7.

Also like the configuration with unobserved preferences, using the same choice set $\mathcal{J}_t = \mathcal{J}$ in the third row eliminates cross-market choice set variation, substantially increasing the bias and variance of $\hat{\rho}$. Intuitively, just like how cross-market choice set variation is needed to credibly identify Σ, it is also needed to credibly identify ρ with just aggregate data.

In the following rows, we illustrate the benefits of second choice data. First, we match the same moments that we considered for targeting σ_x: the covariance between the exogenous product characteristic for first and second choices, the sum of these, and the share of consumers who divert from a low- or high-x_{2jt} first choice to a low-x_{2kt} second choice. Each substantially improves the performance of $\hat{\rho}$, even though these moments are not specifically "targeted" to the nested case.

We also consider matching the share of individuals who would not divert to a different category, which, as discussed above, is a fairly intuitive source of identifying variation for ρ. Unsurprisingly, this targeted moment performs the best, delivering very low variance. Finally, optimal micro moments that require the full micro dataset reduce the bias and variance of the estimator even more.

Table B1: Nesting Parameter

Micro Moments Shorthand	$\mathcal{J}_t = \mathcal{J}$	Optimal	MAE (%)			Bias (%)		
			$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\rho}$	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\rho}$
No Micro Moments			210.9	67.8	3.7	-10.0	-1.1	0.2
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			36.8	11.1	3.3	-1.1	-0.4	0.1
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Yes		44.7	14.2	65.3	20.9	5.4	-28.8
and "$\mathbb{C}(x_{2jt}, x_{2kt} \mid j, k \neq 0)$"	Yes		28.2	8.7	6.5	2.3	0.5	-0.3
and "$\mathbb{E}[x_{2jt} + x_{2kt} \mid j, k \neq 0]$"	Yes		28.6	9.3	6.5	5.3	0.9	-0.5
and "$\mathbb{P}(x_{2kt} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t}, j, k \neq 0)$"	Yes		26.4	8.5	4.8	4.6	1.0	-0.2
and "$\mathbb{P}(h(j) = h(k) \mid j, k \neq 0)$"	Yes		27.6	8.9	1.2	4.0	0.8	-0.2
and "$\mathbb{P}(h(j) = h(k) \mid j, k \neq 0)$"	Yes	Yes	10.6	2.8	1.1	-1.6	-0.6	-0.0

This table reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets for different combinations of micro moments with a nesting parameter. After the first two rows, we use the same choice set $\mathcal{J}_t = \mathcal{J}$ in each market, cluster our estimates of the asymptotic covariance matrix for ξ_{jt} by product j, and use the number of markets T as the number of aggregate observations N_A. In addition to the main micro dataset, we simulate a second, independent micro dataset that is configured the same, except that it also reports second choices. The shorthand "$\mathbb{P}(x_{2kt} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t}, j, k \neq 0)$" refers to two moments that match the share of individuals who divert from a below- or above-median x_{2jt} first choice j to a below-median x_{2kt} second choice k.

C. FRAC Approximation

Salanié and Wolak (2022) approximate the aggregate BLP estimator with a "small-σ" expansion, which gives a linear two stage least squares estimator that is fast to compute, "robust" to different high moments of random coefficients, and approximately correct ("FRAC"). In (8) we wrote down the FRAC estimator for the simplest scalar case with $C = 1$ product characteristic and $R = 1$ demographic.

Here, using the notation in this paper, we write down the FRAC estimator for the full model with $c = 1, \ldots, C$ characteristics, $r = 1, \ldots, R$ demographics, and indirect utility given by (1) to (3). This estimator is derived by Salanié and Wolak (2022), whose paper contains many more details about the estimator and its performance. We rewrite expressions in the original paper to assist readers who are more familiar with our notation.

A second-order Taylor expansion of the Berry, Levinsohn, and Pakes (1995) inversion around $(\Pi, \Sigma) = 0$ yields

$$
\begin{aligned}
\log \frac{s_{jt}}{s_{0t}} = {} & \sum_c \beta_c \cdot x_{cjt} + \sum_{c' \leq c} (\Sigma\Sigma')_{cc'} \cdot a_{cc'jt} \\
& + \sum_c \sum_r \Pi_{cr} \cdot m_{rt}^y \cdot x_{cjt} + \sum_{c' \leq c} \sum_{r,r'} \Pi_{cr} \cdot \Pi_{c'r'} \cdot v_{rr't}^y \cdot a_{cc'jt} \\
& + \xi_{jt} + O\big(\|\tfrac{\Pi}{\Sigma}\|^{3/2}\big),
\end{aligned}
\tag{C1}
$$

in which the within-market mean of demographic r is $m_{rt}^y = \sum_{i \in \mathcal{I}_t} w_{it} \cdot y_{rit}$, the within-market covariance between demographics r and r' is $v_{rr't}^y = \sum_{i \in \mathcal{I}_t} w_{it} \cdot (y_{rit} - m_{rt}^y) \cdot (y_{r'it} - m_{r't}^y)$, and "artificial regressors" are

$$
a_{cc'jt} = \begin{cases} \left(\frac{x_{cjt}}{2} - \sum_{k \in \mathcal{J}_t} s_{kt} \cdot x_{ckt}\right) \cdot x_{cjt} & \text{if } c' = c, \\ x_{cjt} \cdot x_{c'jt} - x_{cjt} \sum_{k \in \mathcal{J}_t} s_{kt} \cdot x_{c'kt} - x_{c'jt} \sum_{k \in \mathcal{J}_t} s_{kt} \cdot x_{ckt} & \text{if } c' \neq c. \end{cases}
\tag{C2}
$$

In practice, one would ignore the approximation term in (C1) and estimate the regression using standard instruments for the endogenous regressors (including the artificial regressors). A small complication is that Π enters both linearly and quadratically. When the number of characteristics and demographics are reasonably small, it is perhaps simplest to treat each $\Pi_{cr} \cdot \Pi_{c'r'}$ as an additional unconstrained parameter, and to estimate Π only from cross-market variation in demographic means m_{rt}^y, while "controlling" for each of the $C^2 \times R^2$ covariates $v_{rr't}^y \cdot a_{cc'jt}$.

D. Asymptotics

The micro BLP model has a number of quantities that could be interpreted as "sample sizes." The total number of aggregate observations $N_{\mathrm{A}} = \sum_{t \in \mathcal{T}} |\mathcal{J}_t|$ can be decomposed into the number of markets $T = |\mathcal{T}|$ and the number of products per market $J_t = |\mathcal{J}_t|$. Similarly, in micro dataset $d \in \mathcal{D}$, the total number of observations $N_d = |\mathcal{N}_d|$ can be decomposed into the number of micro markets $T_d = |\mathcal{T}_d| = |\{t \in \mathcal{T} : \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t \cup \{0\}} w_{dijt} > 0\}|$ with nonzero sampling weights that it is possible to sample from and the number of micro observations per market $N_{dt} = |\mathcal{N}_{dt}| = |\{n \in \mathcal{N}_d : t_n = t\}|$.

Depending on the relative sizes of these quantities, different asymptotic thought experiments are more or less appropriate. The following cases should cover most situations. For simplicity, we focus on cases for which the asymptotic behavior is the same for all micro datasets.

(a) Fixing J_t, let $T \to \infty$ and $T_d \to \infty$ with $T/T_d \to \lambda_d^a$. This case is most appropriate when there are many markets, including those covered by surveys.

　(a1) A large fraction of markets are covered by surveys: $\lambda_d^a \in [1, \infty)$.

　(a2) A small fraction is covered, but this still represents a large number: $\lambda_d^a = \infty$.

(b) Fixing J_t and T_d, let $T \to \infty$ and $N_d \to \infty$ with $T/N_d \to \lambda_d^b$. This case requires at least one large micro dataset d with $N_{dt} \to \infty$ and is most appropriate when there are many markets, few with surveys, but the surveys are large.

　(b1) The number of markets and the sizes of surveys are comparable: $\lambda_d^b \in (0, \infty)$.

　(b2) The number of markets is appreciably larger: $\lambda_d^b = \infty$.

　(b3) The sizes of surveys are appreciably larger: $\lambda_d^b = 0$.

(c) Fixing T and T_d, let $N_{\mathrm{A}} \to \infty$ and $N_d \to \infty$ with $N_{\mathrm{A}}/N_d \to \lambda_d^c$. This case requires at least one large market t with $J_t \to \infty$ and is most appropriate when there are few markets, but markets and surveys are both large.

　(c1) The number of products and the sizes of surveys are comparable: $\lambda_d^c \in (0, \infty)$.

　(c2) The number of products is appreciably larger: $\lambda_d^c = \infty$.

　(c3) The sizes of surveys are appreciably larger: $\lambda_d^c = 0$.

Each case considers asymptotics for a sequence of data generating processes (DGPs) indexed by N_A. Each case (a), (b), and (c) fixes sequences of (T, T_d), (T, N_d), and (N_A, N_d), respectively, indexed by N_A. This delivers different sequences of DGPs for each case. Each DGP prescribes how aggregate data are generated according to the process described in Section 2, and conditional on the aggregate data, how micro data are generated according to the process described in Section 4. We will describe the behavior of the micro BLP estimator under each of these fixed sequences.

Our goal is not to provide a fully formal set of conditions under which the micro BLP estimator is consistent and asymptotically normal, but rather to provide just enough formalities to give insight into when the estimator is expected to work well. We primarily focus on asymptotic normality but briefly discuss consistency. The punchline is that we should expect the estimator to work well in most situations, although we point to a few when it can perform poorly. Our Monte Carlo experiments in Section 7 and in particular those in Figure 3 suggest that these desirable asymptotic properties translate to finite samples.

For the aggregate estimator, Freyberger (2015) and Hong et al. (2021) provide a more formal treatment of the many-markets case with $T \to \infty$. Berry, Linton, and Pakes (2004) provide a more formal treatment of the many-products case with $J_t \to \infty$. Myojo and Kanazawa (2012) extend this many-products treatment to case (c1) with micro moments of the form used by Petrin (2002).

PyBLP also supports clustered aggregate moments. Case (a) or (b) is most appropriate when the number of clusters scales with the number of markets; we will consider the case when we expect ξ_{jt} to be correlated within market, so we cluster by market t. Case (c) is most appropriate when the number of clusters scales with the number of products per market; we will also consider the case when we expect ξ_{jt} to be correlated within a product identifier j that is common across markets, so we cluster by j.

One "sample size" that we do not consider in our asymptotic analysis is the number of simulation draws underlying each finite set of consumer types $\mathcal{I}_t$, if simulation was indeed used to form these sets. Myojo and Kanazawa (2012) extend Berry, Linton, and Pakes's (2004) analysis of simulation error to case (c1), which substantially complicates asymptotics. Incorporating simulation error into the asymptotic analysis presented in this appendix may be an interesting direction for future research. Our sense is that resulting practical guidance will be similar to best practices outlined in Conlon and Gortmaker (2020), which boil down to using many scrambled Halton draws (Owen, 2017) or a suitable quadrature rule. For those particularly concerned about sampling error, PyBLP also supports resampling from

consumer types to estimate the contribution of simulation error to the estimator's asymptotic covariance matrix.

Consistency

Let $\hat{Q}(\theta) = \hat{g}(\theta)'\hat{W}\hat{g}(\theta)$ be the sample objective in (18) and let Θ be the parameter space. Theorem 2.1 of Newey and McFadden (1994) states that if there exists a function $Q(\theta)$ such that (i) $Q(\theta)$ is uniquely maximized at θ_0; (ii) Θ is compact; (iii) $Q(\theta)$ is continuous; (iv) $\hat{Q}(\theta)$ converges uniformly in probability to $Q(\theta)$, then $\hat{\theta} \xrightarrow{P} \theta_0$.

For nonlinear models, it is notoriously difficult to come up with primitive conditions that deliver consistency. We follow the literature by simply assuming that all conditions all satisfied. For the micro BLP model, the substantive conditions are (i) and (ii). In practice, we recommend following the best practices discussed in Conlon and Gortmaker (2020)— using multiple starting values and checking both first- and second-order conditions—to be more confident that the sample objective is globally minimized at $\hat{\theta}$. We also recommend imposing the assumed bounds on Θ during optimization.

The remaining conditions (iii) and (iv) are standard continuity and uniform convergence assumptions, which are fairly mild in models for which we have every reason to expect moments to exist and be continuous. Coming up with primitive conditions that deliver uniform convergence for case (c) may be particularly difficult because of the non-standard limit theorems required to handle that case, which we discuss below when we can develop more useful intuition.

One specific case that these assumptions rule out is non-smooth micro moment functions f_m. Specifically, conditions for consistency and asymptotic normality will typically include twice continuous differentiability. However, all micro BLP applications with which we are familiar match averages or covariances, both of which can be implemented with smooth micro moment functions.

Asymptotic Normality

Let $\hat{G}(\theta) = \frac{\partial \hat{g}(\theta)}{\partial \theta'}$. Assume $\hat{W} \xrightarrow{P} W$ is positive semi-definite, as is the case for all weighting matrices we consider. Given consistency $\hat{\theta} \xrightarrow{P} \theta_0$, Theorem 3.2 of Newey and McFadden (1994) states that if (i) θ_0 is in the interior of Θ; (ii) $\hat{g}(\theta)$ is continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0; (iii) $N_A^{1/2}\hat{g}(\theta_0) \xrightarrow{D} N(0, S)$; (iv) there is $G(\theta)$ that is continuous at θ_0 and $\sup_{\theta \in \mathcal{N}} \|\hat{G}(\theta) - G(\theta)\| \xrightarrow{P} 0$; (v) for $G = G(\theta_0)$, $G'WG$ is nonsingular, then $N_A^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, (G'WG)^{-1}G'WSWG(G'WG)^{-1})$. We scale by $N_A^{1/2}$ because the number

of aggregate observations is the common "sample size" and goes to infinity in all the above cases. We will point to cases when some parts of $\hat{\theta}$ converge at different rates.

Again, we simply assume that all conditions except (iii) are satisfied. Conditions (ii), (iv), and (v) are mild and follow from similar primitives that deliver consistency. In practice, (i) can be a concern when standard deviation estimators in $\hat{\Sigma}$ are near zero. Although beyond the scope of this paper, we point to Ketz (2019) who discusses how this boundary issue can result in asymptotic size distortions for the aggregate BLP estimator.

We focus on condition (iii), that $N_{\mathrm{A}}^{1/2} g(\theta_0) \xrightarrow{D} N(0, S)$. Let $\hat{g}_{\mathrm{P}}(\theta_0) = \bar{v} - v(\theta_0)$ be micro part sample moments with terms that when passed through micro moment functions $f = (f_1, \ldots, f_{M_{\mathrm{M}}})'$ deliver micro sample "moments" $\hat{g}_{\mathrm{M}}(\theta_0) = f(\bar{v}) - f(v(\theta_0))$. It will suffice to show that $N_{\mathrm{A}}^{1/2}(\hat{g}_{\mathrm{A}}(\theta_0)', \hat{g}_{\mathrm{P}}(\theta_0)')' \xrightarrow{D} N(0, \Omega)$ since the multivariate delta method then delivers (iii) with $S = \mathrm{diag}(I_{M_{\mathrm{A}}}, F)\, \Omega\, \mathrm{diag}(I_{M_{\mathrm{A}}}, F)'$ where $F = \frac{\partial f(v(\theta_0))}{\partial v'}$. Again, use of the delta method rules out non-smooth micro moment functions. Specifically, it requires continuous differentiability, although other primitive conditions that we have simply assumed will likely require twice continuous differentiability.

Towards deriving Ω for each case (a), (b), and (c), which we denote by Ω^a, Ω^b, and Ω^c, we decompose the number of aggregate observations into $N_{\mathrm{A}} = T \cdot \bar{J}$ where the average number of products per market is $\bar{J} = \frac{1}{T} \sum_{t \in \mathcal{T}} J_t$. For each case, we rewrite

$$
N_{\mathrm{A}}^{1/2} \begin{bmatrix} \hat{g}_{\mathrm{A}}(\theta_0) \\ \hat{g}_{\mathrm{P}}(\theta_0) \end{bmatrix} = \begin{bmatrix} \bar{J}^{1/2} \cdot \frac{1}{T^{1/2}} \sum_{t \in \mathcal{T}} \frac{1}{J_t} \sum_{j \in \mathcal{J}_t} \xi_{jt} \cdot z_{jt} \\ \bar{J}^{1/2} \cdot \frac{T^{1/2}}{T_{d_1}^{1/2}} \cdot \frac{1}{T_{d_1}^{1/2}} \sum_{t \in \mathcal{T}_{d_1}} \frac{1}{N_{d_1 t}} \sum_{n \in \mathcal{N}_{d_1 t}} (v_{1 i_n j_n t} - v_1(\theta_0)) \\ \vdots \\ \bar{J}^{1/2} \cdot \frac{T^{1/2}}{T_{d_{P_{\mathrm{M}}}}^{1/2}} \cdot \frac{1}{T_{d_{P_{\mathrm{M}}}}^{1/2}} \sum_{t \in \mathcal{T}_{d_{P_{\mathrm{M}}}}} \frac{1}{N_{d_{P_{\mathrm{M}}} t}} \sum_{n \in \mathcal{N}_{d_{P_{\mathrm{M}}} t}} (v_{P_{\mathrm{M}} i_n j_n t} - v_{P_{\mathrm{M}}}(\theta_0)) \end{bmatrix} \tag{D1a}
$$

$$
= \begin{bmatrix} \bar{J}^{1/2} \cdot \frac{1}{T^{1/2}} \sum_{t \in \mathcal{T}} \frac{1}{J_t} \sum_{j \in \mathcal{J}_t} \xi_{jt} \cdot z_{jt} \\ \bar{J}^{1/2} \cdot \frac{T^{1/2}}{N_{d_1}^{1/2}} \cdot \frac{1}{N_{d_1}^{1/2}} \sum_{n \in \mathcal{N}_{d_1}} (v_{1 i_n j_n t_n} - v_1(\theta_0)) \\ \vdots \\ \bar{J}^{1/2} \cdot \frac{T^{1/2}}{N_{d_{P_{\mathrm{M}}}}^{1/2}} \cdot \frac{1}{N_{d_{P_{\mathrm{M}}}}^{1/2}} \sum_{n \in \mathcal{N}_{d_{P_{\mathrm{M}}}}} (v_{P_{\mathrm{M}} i_n j_n t_n} - v_{P_{\mathrm{M}}}(\theta_0)) \end{bmatrix} \tag{D1b}
$$

$$
= \begin{bmatrix} \frac{1}{N_{\mathrm{A}}^{1/2}} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}_t} \xi_{jt} \cdot z_{jt} \\ \frac{N_{\mathrm{A}}^{1/2}}{N_{d_1}^{1/2}} \cdot \frac{1}{N_{d_1}^{1/2}} \sum_{n \in \mathcal{N}_{d_1}} (v_{1 i_n j_n t_n} - v_1(\theta_0)) \\ \vdots \\ \frac{N_{\mathrm{A}}^{1/2}}{N_{d_{P_{\mathrm{M}}}}^{1/2}} \cdot \frac{1}{N_{d_{P_{\mathrm{M}}}}^{1/2}} \sum_{n \in \mathcal{N}_{d_{P_{\mathrm{M}}}}} (v_{P_{\mathrm{M}} i_n j_n t_n} - v_{P_{\mathrm{M}}}(\theta_0)) \end{bmatrix}. \tag{D1c}
$$

Depending on the asymptotic thought experiment, variants of the Central Limit Theorem will apply to the different sums in the above expressions. Regardless, we will need the covariance matrix Ω^0 for the mean-zero vector $(\xi_{jt} \cdot z_{jt}, v_{1i_n j_n t_n} - v_1(\theta_0), \ldots, v_{P_M i_n j_n t_n} - v_{P_M}(\theta_0))'$. The upper-left block is simply

$$\Omega^0_{00} = \mathbb{V}(\xi_{jt} \cdot z_{jt}). \tag{D2}$$

Since micro data are generated conditional on the aggregate data, there is zero covariance between aggregate moments and each micro part p:

$$\Omega^0_{0p} = \mathbb{E}[\xi_{jt} \cdot z_{jt} \cdot \mathbb{E}_A[v_{p i_n j_n t_n} - v_p(\theta_0)]] = 0. \tag{D3}$$

Since micro datasets are independent conditional on the aggregate data, there is also zero covariance between parts p and q based on different datasets $d_p \neq d_q$:

$$\Omega^0_{pq} = \mathbb{E}[(v_{p i_n j_n t_n} - v_p(\theta_0)) \cdot \mathbb{E}_A[v_{q i_n j_n t_n} - v_q(\theta_0) \mid \mathcal{N}_{d_p}]] = 0. \tag{D4}$$

The covariance between parts p and q based on the same dataset $d_p = d_q = d$ is

$$\Omega^0_{pq} = \mathbb{E}[\mathbb{C}_A(v_{p i_n j_n t_n}, v_{q i_n j_n t_n})], \tag{D5}$$

in which conditional on aggregate data, the covariance between parts p and q is

$$\begin{aligned}
&\mathbb{C}_A(v_{p i_n j_n t_n}, v_{q i_n j_n t_n}) \\
&= \frac{\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t \cup \{0\}} w_{it} \cdot s_{ijt}(\theta_0) \cdot w_{dijt} \cdot (v_{pijt} - v_p(\theta_0)) \cdot (v_{qijt} - v_q(\theta_0))}{\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t \cup \{0\}} w_{it} \cdot s_{ijt}(\theta_0) \cdot w_{dijt}}.
\end{aligned} \tag{D6}$$

Case (a) with fixed J_t and $T/T_d \to \lambda_d^a$ is when there are many markets, including those covered by surveys. Since markets are i.i.d., terms in the sums over $t \in \mathcal{T}$ and $t \in \mathcal{T}_d$ in (D1a) are as well. The classical Central Limit Theorem and the Cramér–Wold device delivers convergence in distribution to $N(0, \Omega^a)$ where Ω^a has the same zeros as Ω^0. Without any

clustering,[102] its upper-left block is

$$\Omega_{00}^a = \mathbb{E}[J_t] \cdot \mathbb{V}\left(\frac{1}{J_t} \sum_{j \in \mathcal{J}_t} \xi_{jt} \cdot z_{jt}\right) \tag{D7}$$

$$= \mathbb{E}[J_t] \cdot \mathbb{E}\left[\mathbb{V}\left(\frac{1}{J_t} \sum_{j \in \mathcal{J}_t} \xi_{jt} \cdot z_{jt} \ \middle|\ J_t\right)\right] + 0$$

$$= \mathbb{E}[J_t] \cdot \mathbb{E}[1/J_t] \cdot \Omega_{00}^0.$$

Since $T_d \to \infty$, we treat survey selection probabilities w_{dijt} as random variables in the aggregate data, and hence i.i.d. across markets. Intuitively, as the number of markets increases, each survey is extended similarly across these new markets. This means that we can equivalently write (D5) as $\Omega_{pq} = \mathbb{E}[\mathbb{C}_A(v_{p i_n j_n t_n}, v_{q i_n j_n t_n} \mid n \in \mathcal{N}_{dt})]$, conditioning on an arbitrary market t within the outer expectation over i.i.d. markets. For micro parts p and q based on the same dataset $d_p = d_q = d$,

$$\Omega_{pq}^a = \mathbb{E}[J_t] \cdot \lambda_d^a \cdot \mathbb{C}\left(\frac{1}{N_{dt}} \sum_{n \in \mathcal{N}_{dt}} (v_{p i_n j_n t_n} - v_p(\theta_0)), \frac{1}{N_{dt}} \sum_{n \in \mathcal{N}_{dt}} (v_{q i_n j_n t_n} - v_q(\theta_0))\right) \tag{D8}$$

$$= \mathbb{E}[J_t] \cdot \lambda_d^a \cdot \mathbb{E}\left[\mathbb{C}_A\left(\frac{1}{N_{dt}} \sum_{n \in \mathcal{N}_{dt}} (v_{p i_n j_n t_n} - v_p(\theta_0)), \frac{1}{N_{dt}} \sum_{n \in \mathcal{N}_{dt}} (v_{q i_n j_n t_n} - v_q(\theta_0)) \ \middle|\ N_{dt}\right)\right] + 0$$

$$= \mathbb{E}[J_t] \cdot \lambda_d^a \cdot \mathbb{E}[1/N_{dt}] \cdot \Omega_{pq}^0.$$

Sub-case (a1) with $T/T_d \to \lambda_d^a \in [1, \infty)$ is when the number of survey markets T_d diverges at the same rate as the total number of markets T. All elements of $\hat{\theta}$ converge to θ_0 at the common rate $T^{1/2}$, which for this case is asymptotically equivalent to $N_A^{1/2}$, $\max_{d \in \mathcal{D}} T_d^{1/2}$, and $\max_{d \in \mathcal{D}} N_d^{1/2}$.

Sub-case (a2) with $T/T_d \to \lambda_d^a = \infty$ is when the number of markets T diverges faster than the number of survey markets T_d. If aggregate variation is sufficient to strongly identify all parameters in θ, all of $\hat{\theta}$ will converge at the faster rate of $T^{1/2}$ and the micro data will be asymptotically negligible.[103] If some parameters are not identified or only weakly identified by aggregate variation but are identified by variation in some micro datasets d, then their estimators in $\hat{\theta}$ will at least converge at the slower rate of $\max_d T_d^{1/2}$. For example, if there is no cross-market variation in the distribution of demographics whatsoever, Π will not be

---

[102]With clustering, for example, by market t, $\Omega_{00}^a = \mathbb{E}[J_t] \cdot \mathbb{V}(\frac{1}{J_t} \sum_{j \in \mathcal{J}_t} \xi_{jt} \cdot z_{jt})$, the second term of which is estimable with a cluster-robust covariance estimator. The other terms in Ω^a are the same.

[103]In this case, the aggregate moments' asymptotic covariance matrix will be finite and the micro moments' asymptotic covariance matrix will be infinite because it is scaled by $\lambda_d^a = \infty$.

identified from aggregate variation, and its estimator will converge at the rate of the micro data.

Case (b) with fixed J_t, fixed T_d, and $T/N_d \to \lambda_d^b$ is when there are many markets, few with surveys, but the surveys are large. For this thought experiment to make sense, we need to treat survey sampling probabilities as non-random; otherwise, each new market would have a chance of being included in a micro dataset, and surveys would extend to infinitely many markets. Unlike case (a), the Central Limit Theorem would not directly apply to terms in the sums over $t \in \mathcal{T}$ and $n \in \mathcal{N}_d$ in (D1b) because there are overlapping markets that introduce correlation between the terms. However, conditional on all aggregate data in these overlapping markets $t \in \cup_{d \in \mathcal{D}} \mathcal{T}_d$, which are asymptotically negligible as $T \to \infty$, the classical Central Limit Theorem and the Cramér–Wold device guarantee conditional convergence in distribution to $N(0, \Omega^b)$ where Ω^b now depends on the aggregate data in these markets. Again, Ω^b has the same zeros as Ω^0 and the same upper-left block case as for case (a),

$$\Omega_{00}^b = \Omega_{00}^a. \tag{D9}$$

For micro parts p and q based on the same dataset $d_p = d_q = d$,

$$\Omega_{pq}^b = \mathbb{E}[J_t] \cdot \lambda_d^b \cdot \Omega_{pq}^0, \tag{D10}$$

in which we can write (D5) as $\Omega_{pq} = \mathbb{C}_A(v_{p i_n j_n t_n}, v_{q i_n j_n t_n})$ without the outer expectation because we are already conditioning on all aggregate data in markets $t \in \mathcal{T}_d$ on which micro dataset d could possibly depend.

Sub-case (b1) with $T/N_d \to \lambda_d^b \in (0, \infty)$ is when the number of markets T diverges at the same rate as each survey size N_d. Rates of convergence are the same as for sub-case (a1).

Sub-case (b2) with $T/N_d \to \lambda_d^b = \infty$ is when the number of markets T diverges faster than each survey size N_d. Rates of convergence are the same as for sub-case (a2).

Sub-case (b3) with $T/N_d \to \lambda_d^b = 0$ is when the number of markets T diverges slower than each survey size N_d. Since micro data are uninformative about linear parameters β (see Footnote 57), $\hat{\beta}$ converges to β_0 at the slower rate of $T^{1/2}$. If variation in some micro datasets d is sufficient to strongly identify all nonlinear parameters in (Π, Σ), all of $(\hat{\Pi}, \hat{\Sigma})$ will converge at the faster rate of $\max_d N_d^{1/2}$ and the aggregate data will only have an asymptotic influence on the linear estimators $\hat{\beta}$. If some nonlinear parameters are not identified or only weakly identified by micro variation but are identified by aggregate variation, then their estimators will at least converge at the slower rate of $T^{1/2}$.

Case (c) with fixed T, fixed T_d, and $N_A/N_d \to \lambda_d^c$ is when there are few markets, but markets and surveys are both large. Unlike case (b), fixing T means that overlapping markets $t \in \cup_{d \in \mathcal{D}} \mathcal{T}_d$ are no longer asymptotically negligible. The simplest approach is to still condition on all aggregate data in these overlapping markets; if there are any markets remaining without micro data, the classical Central Limit Theorem and the Cramér–Wold device applied to (D1c) would again guarantee conditional convergence in distribution to $N(0, \Omega^c)$ where Ω^c again depends on the aggregate data in these markets. Again, Ω^c has the same zeros as Ω^0. Without any clustering,[104] its upper-left block is simply

$$\Omega_{00}^c = \Omega_{00}^0. \tag{D11}$$

For micro parts p and q based on the same dataset $d_p = d_q = d$,

$$\Omega_{pq}^c = \lambda_d^c \cdot \Omega_{pq}^0, \tag{D12}$$

in which like for case (b), we can simply write (D5) as $\Omega_{pq}^0 = \mathbb{C}_A(v_{p i_n j_n t_n}, v_{q i_n j_n t_n})$ because we are already conditioning on the relevant aggregate data.

In practice, conditioning on overlapping markets means dropping them from the aggregate moments by setting instruments $z_{jt} = 0$ in all $t \in \cup_{d \in \mathcal{D}} \mathcal{T}_d$. If a sizeable proportion of markets are covered by surveys, this means discarding a great deal of aggregate variation, which may be useful in finite samples for estimating linear parameters β, or for providing backup variation if micro data does not identify or only weakly identifies some of the nonlinear parameters (Π, Σ).

We can retain this aggregate variation by applying the Lyapunov Central Limit Theorem for triangular arrays (e.g., Theorem 27.3 in Billingsley, 1995) to (D1c). This requires making two additional assumptions that correspond to B4(d) and B4(h) in Myojo and Kanazawa (2012). First, we assume that as $N_A \to \infty$, the limit of $\Omega_{pq} = \mathbb{C}_A(v_{p i_n j_n t_n}, v_{q i_n j_n t_n})$ becomes non-random. For example, we could apply a law of large numbers to sums over products in (D6). This would be problematic if, for example, our micro moments match statistics for only a single or a few products (e.g., "$\mathbb{E}[y_{rit} \mid j = j_m]$" for a single product j_m), rather than for groups of products that expand asymptotically. Dropping markets with such statistics could be a safer approach. Second, for each dataset $d \in \mathcal{D}$, we assume there exists a $\delta_d > 0$ such that $N_d \cdot \mathbb{E}[\| [v_{p i_n j_n t_n} - v_p(\theta_0)]_{d_p = d}/N_d \|^{2+\delta_d}] \to 0$ as $N_A \to \infty$. This Lyapunov condition is similar

---

[104]With clustering, for example, by product $j \in \mathcal{J}_t = \mathcal{J}$, $\Omega_{00}^c = T \cdot \mathbb{V}(\frac{1}{T} \sum_{t \in \mathcal{T}} \xi_{jt} \cdot z_{jt})$, the second term of which is estimable with a cluster-robust covariance estimator. The other terms in Ω^c are the same.

to the regularity conditions we have assumed for consistency and asymptotic normality, and in general, we expect it to hold.

Sub-case (c1) with $N_A/N_d \to \lambda_d^c \in (0, \infty)$ is when the number of products N_A diverges at the same rate as each survey size N_d. Rates of convergence are the same as for sub-cases (a1) and (b1).

Sub-case (c2) with $N_A/N_d \to \lambda_d^c = \infty$ is when the number of products N_A diverges faster than each survey size N_d. Rates of convergence are the same as for sub-cases (a2) and (b2).

Sub-case (c3) with $N_A/N_d \to \lambda_d^c = 0$ is when the number of products N_A diverges slower than each survey size N_d. Rates of convergence are the same as for sub-case (b3).

Asymptotic Variance Estimation with PyBLP

To obtain a consistent estimator of the asymptotic variance matrix for $\hat{\theta}$, it suffices to find a consistent estimator $\hat{S} \xrightarrow{P} S$. Given the assumptions used to show asymptotic normality, Theorem 4.2 of Newey and McFadden (1994) states that $(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{S}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1} \xrightarrow{P} (G'WG)^{-1}G'WSWG(G'WG)^{-1}$.

With micro moments, there is no "canonical" choice for the initial weighting matrix, like the 2SLS weighting matrix for the aggregate estimator. Instead, we prefer to use $\hat{W} = \hat{S}^{-1}$ at some initial guess for the true θ_0, which could be informed by estimates that only use aggregate variation. After one round of optimization, we compute a consistent estimator of an efficient weighting matrix $\hat{W} = \hat{S}^{-1}$ at the initial consistent estimator $\hat{\theta}$.

There are at least three approaches to computing $\hat{G} = \frac{\partial \hat{g}(\hat{\theta})}{\partial \theta'}$: numerical, analytic, or automatic differentiation. In practice, $\hat{G}(\theta)$ is also needed during numerical optimization to provide the optimizer with the objective's gradient $\frac{\partial \hat{Q}(\theta)}{\partial \theta} = 2\hat{G}(\theta)'\hat{W}\hat{g}(\theta)$. In PyBLP we support both numerical and analytic differentiation, but use analytic derivatives by default to minimize numerical error.[105] In Appendix A2 of Conlon and Gortmaker (2020) we derive expressions for derivatives of aggregate sample moments, including for the case with supply-side moments. For micro sample moments,

$$\frac{\partial \hat{g}_M(\theta)}{\partial \theta'} = \begin{bmatrix} \frac{\partial f_1(v(\theta))}{\partial v'}\frac{\partial v(\theta)}{\partial \theta'} \\ \vdots \\ \frac{\partial f_{M_M}(v(\theta))}{\partial v'}\frac{\partial v(\theta)}{\partial \theta'} \end{bmatrix}, \tag{D13}$$

---

[105]We chose to implement analytic gradients because ceding control to an automatic differentiation library can limit one's ability to handle numerical errors. However, automatic differentiation is a promising technique for structural estimation and we are optimistic going forward.

in which when PyBLP users specify a micro moment function f_m they also specify its derivative function $\frac{\partial f_m}{\partial v}$, and each $\frac{\partial v_p(\theta)}{\partial \theta'}$ is simply the derivative of (17). The one difficulty is that choice probabilities $s_{ijt}(\theta)$ depend on θ both directly and through their dependence on $\hat{\delta}(\theta)$, which needs to be accounted for during differentiation.

Finally, we compute $\hat{S} = \text{diag}(\hat{S}_A, \hat{S}_M)$. The block-diagonal structure comes from the lack of correlation between aggregate and micro moments. Without any clustering,[106]

$$\hat{S}_A = \frac{1}{N_A} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}_t} \hat{g}_{jt} \hat{g}'_{jt}, \quad \hat{g}_{jt} = \hat{\xi}_{jt}(\hat{\theta}) \cdot z_{jt}. \tag{D14}$$

The micro block is

$$\hat{S}_M = \hat{F} \hat{S}_P \hat{F}', \quad \hat{F} = \frac{\partial f(v(\hat{\theta}))}{\partial v'}, \tag{D15}$$

in which element (p, q) of $\hat{S}_P$ is zero if $d_p \neq d_q$, and otherwise equal to $\frac{N_A}{N_d} \cdot \hat{\Omega}^0_{pq}$ where $\hat{\Omega}^0_{pq}$ is the sample analogue of (D6):

$$\hat{\Omega}^0_{pq} = \frac{\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t \cup \{0\}} w_{it} \cdot s_{ijt}(\hat{\theta}) \cdot w_{dijt} \cdot (v_{pijt} - v_p(\hat{\theta})) \cdot (v_{qijt} - v_q(\hat{\theta}))}{\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t \cup \{0\}} w_{it} \cdot s_{ijt}(\hat{\theta}) \cdot w_{dijt}}. \tag{D16}$$

For each case (a), (b), and (c), $\text{diag}(\hat{S}_A, \hat{S}_P)$ converges in probability to Ω^a, Ω^b, and Ω^c.

---

[106] For clusters $\ell = 1, \dots, L$ of products $\mathcal{J}_{\ell t} \subset \mathcal{J}_t$, we compute $\hat{S}_A = \frac{1}{N_A} \sum_{\ell=1}^L \hat{g}_\ell \hat{g}'_\ell$ for $\hat{g}_\ell = \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}_{\ell t}} \hat{g}_{jt}$.

E. Efficiency

In Algorithm 2 we describe an algorithm for computing an optimal micro BLP estimator. Building on Appendix D where we discuss under which conditions the estimator is consistent and asymptotically normal, here we show that the optimal estimator we compute is in fact asymptotically efficient within the class of possible micro BLP estimators. We also discuss why it should in general be unnecessary to adjust for first-step estimation error.

Similar to in Appendix D, our goal is not to provide a fully formal set of conditions under which the estimator we compute is asymptotically efficient within the class of estimators that we consider. Instead of fully working out the tangent spaces and regularity conditions that a precise econometric treatment would require, we limit our formality to that of the heuristic efficiency framework in Section 5.3 of Newey and McFadden (1994).

An optimal micro BLP estimator has three parts: an optimal weighting matrix, optimal instruments, and optimal micro moments that match conditional scores. Efficiency should be intuitive. The form of an optimal weighting matrix for minimum distance estimation is well-known. Since aggregate and micro moments are uncorrelated, it is intuitive that it should be efficient to replace each with their efficient counterparts when alone. With aggregate moments alone, Chamberlain's (1987) optimal instruments are efficient. With micro moments alone, maximum likelihood would be efficient, so matching scores should be as well.

Optimal Weighting Matrix

Since the micro BLP estimator in (18) is a minimum distance estimator, we can apply Theorem 5.2 in Newey and McFadden (1994), which guarantees that $\hat{W} = \hat{S}^{-1}$ from Appendix D is asymptotically efficient in the class of minimum distance estimators matching the same statistics. Going forward, we will assume the use of the optimal weighting matrix $W = S^{-1}$ to simplify expressions.

Optimal Instruments and Matching Scores

For simplicity, we will focus on only demand-side aggregate moments and micro moments based on only a single micro dataset d. Extending the argument for efficiency to stacked supply-side moments and multiple micro datasets is straightforward, although notationally cumbersome.

Given aggregate moment conditions $\mathbb{E}[\xi_{jt} \mid z_{jt}] = 0$, any function b of z_{jt} yields valid unconditional moments $\mathbb{E}[g_A(j, t; \theta_0)] = \mathbb{E}[\xi_{jt}(\theta_0) \cdot b(z_{jt})] = 0$ based on the aggregate data.

Similarly, given conditional likelihoods $\mathbb{P}_{\mathrm{A}}(t_n, j_n, y_{i_n t_n} \mid n \in \mathcal{N}_d)$, any function v of the aggregate data and $(t_n, j_n, y_{i_n t_n})$ yields valid moments $\mathbb{E}[g_{\mathrm{P}}(n; \theta_0)] = \mathbb{E}[v(n; \theta_0) - v(\theta_0)] = 0$ where $v(\theta) = \mathbb{E}_{\mathrm{A}}^{\theta}[v(n; \theta_0)]$. These can be passed through any smooth function f to form valid "micro moments" $f(\bar{v}) - f(v(\theta_0)) \xrightarrow{P} 0$ where $\bar{v} = \frac{1}{N_d} \sum_{n \in \mathcal{N}_d} v(n; \theta_0)$.

Using the efficiency framework in Section 5.3 of Newey and McFadden (1994), our goal is to find the "index" $\tau = (b, f, v)$ that minimizes the asymptotic variance of a minimum distance estimator based on these "moments." Restricting our search to indices τ with $f(v) = v$ and $v(\theta_0) = 0$ will deliver asymptotic efficiency within all GMM estimators that stack aggregate and micro moments g_{A} and $g_{\mathrm{P}} = v$. We consider the slightly more general case, which requires applying the delta method.

We incorporate the delta method into a combination of two of Newey and McFadden's (1994) applications that show the efficiency of Chamberlain's (1987) optimal instruments and of maximum likelihood. First, we rewrite the asymptotic variance of a generic micro BLP estimator from Appendix D. Let

$$G = \begin{bmatrix} G_{\mathrm{A}} \\ G_{\mathrm{M}} \end{bmatrix}, \quad G_{\mathrm{A}} = \mathbb{E}\left[\frac{\partial g_{\mathrm{A}}(j, t; \theta_0)}{\partial \theta'}\right], \quad G_{\mathrm{M}} = \mathbb{E}\left[F \cdot \frac{\partial g_{\mathrm{P}}(n; \theta_0))}{\partial \theta'}\right], \quad F = \frac{\partial f(v(\theta_0))}{\partial v'}. \quad \text{(E1)}$$

The asymptotic variance of a micro BLP estimator based on τ and weighting matrix W is

$$V_\tau = D_\tau^{-1} \mathbb{E}[Y_\tau(j, t, n) Y_\tau(j, t, n)'] D_\tau^{-1'},$$

$$D_\tau = G'WG, \quad Y_\tau(j, t, n) = G'W \begin{bmatrix} g_{\mathrm{A}}(j, t; \theta_0) \\ F \cdot g_{\mathrm{P}}(n; \theta_0) \end{bmatrix}. \quad \text{(E2)}$$

Using the above notation, Theorem 5.3 in Newey and McFadden (1994) states that if τ^* satisfies $D_\tau = \mathbb{E}[Y_\tau(j, t, n) \cdot Y_{\tau^*}(j, t, n)']$ for all τ, then any estimator with variance V_{τ^*} is efficient in the class of estimators indexed by τ. A standard intuition for this result is that efficient estimators are uncorrelated with the difference with any other consistent estimator in the same class. We will show that this result holds for $\tau^* = (b^*, f^*, v^*)$ where

$$b^*(z_{jt}) = \mathbb{E}\left[\frac{\partial \xi_{jt}(\theta_0)}{\partial \theta} \,\bigg|\, z_{jt}\right] \mathbb{V}(\xi_{jt} \mid z_{jt})^{-1} \quad \text{(E3)}$$

are Chamberlain's (1987) optimal instruments for the case with a scalar error term,

$$f^*(v) = v \quad \text{(E4)}$$

is simply the identity function so that minimum distance collapses to GMM under τ^*, and

$$v^*(n;\theta) = v^*(n) = \mathbb{S}_{\mathrm{A}}(n \mid d)' \tag{E5}$$

implements Section 6's optimal micro moments by matching the average $1 \times \dim(\theta)$ score $\mathbb{S}_{\mathrm{A}}(n \mid d)$ function from (F1) evaluated at the true θ_0 and micro observation $n \in \mathcal{N}_d$. Note that this choice of $v^*(n;\theta) = v^*(n)$ does not depend on θ.

Stacking scores directly with $v^*(n;\theta) = \mathbb{S}_{\mathrm{A}}^\theta(n \mid d)'$ would also be efficient, and would require very little modification to the below proof. We prefer our approach because, as we discuss in Section 6, the lack of dependence on θ makes it more computationally tractable. In particular, we do not need to differentiate the score, and we only need to compute scores for every micro observation a single time. These computational benefits are typical for estimators based on the unfortunately-named "one step" method discussed, for example, in Section 3.4 of Newey and McFadden (1994).

First, we re-establish the result from Appendix D that aggregate and micro moments are uncorrelated. Iterated expectations give

$$\mathbb{E}[g_{\mathrm{A}}(j,t;\theta_0) \cdot g_{\mathrm{P}}(n;\theta_0)'] = \mathbb{E}[g_{\mathrm{A}}(j,t;\theta_0) \cdot \mathbb{E}_{\mathrm{A}}[g_{\mathrm{P}}(n;\theta_0)']] = 0. \tag{E6}$$

In particular, this implies that the optimal weighting matrix is block-diagonal. Under τ^*,

$$W^* = \mathbb{V}\begin{pmatrix} \xi_{jt} \cdot b^*(z_{jt}) \\ v^*(n) \end{pmatrix}^{-1} = \begin{bmatrix} (\xi_{jt} \cdot b^*(z_{jt}) \cdot b^*(z_{jt})' \cdot \xi_{jt})^{-1} & 0 \\ 0 & (v^*(n) \cdot v^*(n)')^{-1} \end{bmatrix}, \tag{E7}$$

in which we have used the fact that the score is mean-zero, $\mathbb{E}_{\mathrm{A}}[v^*(n)] = \mathbb{E}_{\mathrm{A}}[\mathbb{S}_{\mathrm{A}}(n \mid d)'] = 0$. Next, using the definition for b^* and iterated expectations, we can rewrite

$$\mathbb{E}\left[\frac{\partial g_{\mathrm{A}}(j,t;\theta_0)}{\partial\theta'}\right] = \mathbb{E}\left[\frac{\partial\xi_{jt}(\theta_0)}{\partial\theta'} \cdot b(z_{jt})\right] = \mathbb{E}[\xi_{jt} \cdot b(z_{jt}) \cdot b^*(z_{jt})' \cdot \xi_{jt}]. \tag{E8}$$

To derive a similar expression for the micro moments, we differentiate $\mathbb{E}_\theta[g_{\mathrm{P}}(n;\theta)] = 0$ with respect to θ. Differentiating under the expectation requires regularity conditions similar to those we have assumed for consistency and asymptotic normality in Appendix D. Evaluating at θ_0 and using the definition for v^*, we can also rewrite

$$\mathbb{E}\left[\frac{\partial g_{\mathrm{P}}(n;\theta_0)}{\partial\theta'}\right] = -\mathbb{E}[g_{\mathrm{P}}(n;\theta_0) \cdot \mathbb{S}_{\mathrm{A}}(n \mid d)] = -\mathbb{E}[g_{\mathrm{P}}(n;\theta_0) \cdot v^*(n)']. \tag{E9}$$

Using all of the above results, we can simplify

$$Y_{\tau^*}(j, t, n) = \mathbb{E}\begin{bmatrix} \xi_{jt} \cdot b^*(z_{jt}) \cdot b^*(z_{jt})' \cdot \xi_{jt} \\ -v^*(n) \cdot v^*(n)' \end{bmatrix}' W^* \begin{bmatrix} \xi_{jt} \cdot b^*(z_{jt}) \\ v^*(n) \end{bmatrix} = \begin{bmatrix} \xi_{jt} \cdot b^*(z_{jt}) \\ -v^*(n) \end{bmatrix}. \qquad \text{(E10)}$$

Again using the lack of correlation between aggregate and micro moments,

$$D_\tau = G'W\mathbb{E}\begin{bmatrix} \xi_{jt} \cdot b(z_{jt}) \cdot b^*(z_{jt})' \cdot \xi_{jt} \\ -F \cdot g_{\mathrm{P}}(n; \theta_0) \cdot v^*(n)' \end{bmatrix} = \mathbb{E}[Y_\tau(j, t, n) \cdot Y_{\tau^*}(j, t, n)']. \qquad \text{(E11)}$$

Two-step Estimation

In the discussion so far, we have replaced the optimal weighting matrix, instruments, and micro moments with consistent estimators obtained in a first step. Typically, we would have to correct for first-step estimation error when calculating second-step standard errors. However, according to the general principle developed, for example, in Section 6 of Newey and McFadden (1994), adjusting for first-step estimation error is unnecessary if the consistency of the first-step estimator does not affect the consistency of the second-step estimator.

In Appendix D we simply follow the literature by assuming consistency of $\hat{\theta}$ for general weighting matrices, instruments, and micro moments, while also discussing when this assumption may fail. Typically, however, we expect both steps to deliver consistent estimators so that this principle holds and we do not have to adjust for first-step estimation error. Indeed, standard errors calculated for our Monte Carlo experiments in Appendix H have fairly low bias and good coverage in finite samples when using optimal instruments and optimal micro moments.

F. Micro Data Scores

In micro dataset $d \in \mathcal{D}$, the $1 \times \dim(\theta)$ score function evaluated at parameters θ and micro observation $n \in \mathcal{N}_d$ is

$$\mathbb{S}_{\mathrm{A}}^{\theta}(n \mid d) = \frac{\partial \log \mathbb{P}_{\mathrm{A}}^{\theta}(t_n, j_n, y_{i_n t_n} \mid n \in \mathcal{N}_d)}{\partial \theta'}, \tag{F1}$$

or with an additional k_n if dataset d contains second choices. The probability of $n \in \mathcal{N}_d$ evaluated at θ is

$$\mathbb{P}_{\mathrm{A}}^{\theta}(t_n, j_n, y_{i_n t_n} \mid n \in \mathcal{N}_d) = \frac{\sum_{i \in \mathcal{I}_{t_n}} 1\{y_{it_n} = y_{i_n t_n}\} \cdot w_{it_n} \cdot s_{ij_n t_n}(\theta) \cdot w_{dij_n t_n}}{\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t \cup \{0\}} w_{it} \cdot s_{ijt}(\theta) \cdot w_{dijt}}, \tag{F2}$$

or with second choices,

$$\mathbb{P}_{\mathrm{A}}^{\theta}(t_n, j_n, k_n, y_{i_n t_n} \mid n \in \mathcal{N}_d)$$
$$= \frac{\sum_{i \in \mathcal{I}_{t_n}} 1\{y_{it_n} = y_{i_n t_n}\} \cdot w_{it_n} \cdot s_{ij_n k_n t_n}(\theta) \cdot w_{dij_n k_n t_n}}{\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t \cup \{0\}} \sum_{k \in \mathcal{J}_t \cup \{0\} \backslash \{j\}} w_{it} \cdot s_{ijkt}(\theta) \cdot w_{dijkt}}. \tag{F3}$$

The numerator of (F2) is an integral over unobserved heterogeneity ν_{it_n}. Without unobserved heterogeneity (i.e., $\Sigma = 0$), observing a consumer's demographic $y_{i_n t_n}$ is equivalent to observing the consumer's type i_n, so the numerator in (F2) is simply $s_{i_n j_n t_n}(\theta) \cdot w_{di_n j_n t_n}$. If there is unobserved heterogeneity, it is often simplest to approximate the integral with quadrature.

The denominator of (F2) is $\mathbb{P}_{\mathrm{A}}(n \in \mathcal{N}_d)$, the probability of an arbitrary consumer being selected for the survey. If this probability does not depend on θ because, for example, selection probabilities w_{dijt} do not depend on choice j, then the denominator drops out of the score.

Computing Scores with PyBLP

It is straightforward to compute scores with PyBLP. In Figure F1 we start with the results from a problem (for example, from the replication in Figure 5 of Petrin, 2002) and some micro data. For each observation in the micro data, we compute the score $v_{m i_n j_n t_n}$ in (23) and (F1). Then, for each possible (i, j, t), we compute the same v_{mijt}. When constructing optimal micro moments, we use the average of the former as $f_m(\bar{v}) = \bar{v}_m$, and we use the latter to form $f_m(v(\theta)) = v_m(\theta)$.

One trick that can speed up computing scores in the micro data arises because the score is the same for each distinct set of demographic values, product choice, and market. To use this, we can call each of these a different "observation," and when computing the average $\bar{v}_m$, overweight these "observations" by how many actual micro observations are underlying them. This trick will particularly speed up computation time when demographics take on only a few discrete values (e.g., when they are binary-valued) and purchases are concentrated within a small number of products.

Figure F1: Computing Optimal Micro Moments with PyBLP

```python
from pyblp import Integration, MicroPart, MicroMoment

# Compute scores, integrating over unobserved preferences with quadrature
score_integration = Integration('product', size=7)
micro_scores = problem_results.compute_micro_scores(micro_dataset, micro_data, score_integration)
agent_scores = problem_results.compute_agent_scores(micro_dataset, integration=score_integration)

# Construct optimal micro moments
optimal_micro_moments = []
for m, (micro_scores_m, agent_scores_m) in enumerate(zip(micro_scores, agent_scores)):
    optimal_micro_moments.append(MicroMoment(
        name=f"Score for parameter #{m}",
        value=micro_scores_m.mean(),
        parts=MicroPart(
            name=f"Score for parameter #{m}",
            dataset=micro_dataset,
            compute_values=lambda t, p, a, v=agent_scores_m: v[t],
        ),
    ))
```

This Python code demonstrates how to construct optimal micro moments with PyBLP. After obtaining problem results as in the Petrin (2002) replication code in Figure 5, and given another dataset of micro observations, we compute scores for the micro dataset and all possible consumer type-choices, and then use these to construct a list of optimal micro moments.

G. Selection Procedure

To supplement our discussion of optimal micro moments in Section 6, we provide a more formal approach for determining which summary statistics are most informative about the parameters in the model. The goal is to identify a small number of maximally informative summary statistics that are more likely to be collected by survey administrators than model-specific average scores.

Like our two-step approach for computing optimal micro moments, we begin with a specific model and a first-stage estimator $\hat{\theta}$ or educated guess. The researcher may not have full access to a micro dataset d, either because of data limitations or because the survey has yet to be administered. However, given knowledge about sampling probabilities w_{dijt}, we can easily simulate many observations from the dataset $\{(t_n, j_n, y_{i_n t_n})\}_{n \in \mathcal{N}_d}$ under $\hat{\theta}$. We can then compute scores values $v_{m i_n j_n t_n}$ in (23) for each simulated observation $n \in \mathcal{N}_d$ and add a column of score values for each nonlinear parameter to the simulated micro dataset.

In addition to scores, we can also add columns for many candidate micro values $v_{p i_n j_n t_n}$, averages of which could reasonably be collected by a survey administrator. For example, we could restrict ourselves to second-order polynomial interactions of demographics $y_{r i_n t_n}$ and merged-in aggregate data, such as product characteristics $x_{c j_n t_n}$ and market shares $\mathcal{S}_{j_n t_n}$. We can exclude trivial candidates such as a constant, $v_{p i_n j_n t_n} = 1$, or others with expectations that do not depend on θ.[107]

Given a simulated dataset of scores and candidate micro values, we can run a number of procedures to determine which sets of micro values are most informative about the scores. Perhaps the simplest is to regress each subset of candidate micro values on the scores and keep those sets with the highest R^2. If there are many candidate micro values, a lasso regression could be more practical. To compare these different procedures, it will help to formally define the problem and how it motivates a selection procedure.

Formal Problem

We wish to select $m = 1, \ldots, M_{\mathrm{M}}$ micro moments to minimize the asymptotic variance of the micro BLP estimator in (18). Since aggregate and micro moments are asymptotically uncorrelated (see Appendix D), it will suffice to minimize the micro moments' contribution to the asymptotic variance. Since micro moments from different micro datasets are also

---

[107]For example, with no conditioning, $w_{dijt} = 1$, a candidate micro value simply equal to a demographic, $v_{p i_n j_n t_n} = y_{r i_n t_n}$, gives an expected micro value equal to the mean of that demographic, $v_m(\theta) = \sum_{i \in \mathcal{I}_{t_n}} w_{i t_n} \cdot y_{r i t}$, which is uninformative about θ.

asymptotically uncorrelated, we will focus on a single dataset d. For simplicity, we focus on means: $f_m(\bar{v}) = \bar{v}_m$. The following discussion could be amended to focus on smooth functions of means by using the multivariate delta method to obtain asymptotic covariance matrices.

Let $\mathbb{E}[g_{\mathrm{M}}(n; \theta_0)] = \mathbb{E}[v(n) - v(\theta_0)] = 0$ be the $M_{\mathrm{M}} \times 1$ vector of unconditional micro moments where $v(\theta) = \mathbb{E}_{\mathrm{A}}^{\theta}[v(n)]$ and $v(n) = (v_{1i_n j_n t_n}, \ldots, v_{M_{\mathrm{M}} i_n j_n t_n})'$. Under the optimal weighting matrix, the contribution of these micro moments to the asymptotic variance of $\hat{\theta}$ is the familiar expression for the asymptotic variance of an efficient GMM estimator. Denoting by $\mathcal{V}$ a finite set of candidate micro values, the problem is

$$\min_{v(\cdot) \in \mathcal{V}} \left\| \left(G_{\mathrm{M}}' \mathbb{V}(g_{\mathrm{M}}(n; \theta_0))^{-1} G_{\mathrm{M}} \right)^{-1} \right\|, \quad G_{\mathrm{M}} = \mathbb{E}\left[\frac{\partial g_{\mathrm{M}}(n; \theta_0)}{\partial \theta'} \right], \tag{G1}$$

in which $\|\cdot\|$ is a matrix norm that governs relative weights on parameters in θ. To simplify (G1), note that $\mathbb{V}(g_{\mathrm{M}}(n; \theta_0)) = \mathbb{V}(v(n))$ and

$$\frac{\partial g_{\mathrm{M}}(n; \theta_0)}{\partial \theta'} = \frac{\partial v(\theta_0)}{\partial \theta'} = \frac{\partial \mathbb{E}_{\mathrm{A}}^{\theta_0}[v(n)]}{\partial \theta'} = \mathbb{E}_{\mathrm{A}}[v(n) \mathbb{S}_{\mathrm{A}}(n \mid d)], \tag{G2}$$

in which $\mathbb{S}_{\mathrm{A}}(n \mid d)$ is the $1 \times \dim(\theta)$ score function from (F1) evaluated at the true θ_0 and micro observation $n \in \mathcal{N}_d$. Using iterated expectations, we can rewrite (G1) as

$$\min_{v(\cdot) \in \mathcal{V}} \left\| \left(\mathbb{E}[v(n) \mathbb{S}_{\mathrm{A}}(n \mid d)]' \mathbb{V}(v(n))^{-1} \mathbb{E}[v(n) \mathbb{S}_{\mathrm{A}}(n \mid d)] \right)^{-1} \right\|. \tag{G3}$$

Candidate Micro Values

For this problem to make sense, we need to restrict the set of candidate micro values $\mathcal{V}$. First, as discussed above, $\mathcal{V}$ should only contain micro moment values that could reasonably be collected by a survey administrator, such as those with low complexity and high interpretability. For example, we could restrict $\mathcal{V}$ to second-order polynomial interactions of demographics and merged-in aggregate data.

Second, for any given $v_{m i_n j_n t_n}$ we can define another that delivers the same asymptotic variance by multiplying it by a constant. Without loss, we can normalize $\mathrm{diag}(\mathbb{V}(v(n))) = 1$. In the context of the feasible procedure described in Section 6, this amounts to standardizing candidate micro values before including them in a lasso regression.

Lastly, for (G1) and (G3) to be well-defined, we require $G \neq 0$. In practice, this amounts to excluding trivial candidates with expectations $v_m(\theta) = \mathbb{E}_{\mathrm{A}}[v_{m i_n j_n t_n}]$ that do not depend

on θ. For example, if weights w_{dijt} do not depend on choices j, we can exclude all candidates v_{mijt} that do not depend on both i and j, such as $v_{mijt} = x_{cjt}$ or $v_{mijt} = y_{rit}$.

Intuition from the Scalar Case

For intuition about the rewritten problem in (G3), consider the case with $\dim(\theta) = 1$ parameter and $M_{\mathrm{M}} = 1$ micro moment with $\mathbb{E}[v(n)\mathbb{S}_{\mathrm{A}}(n \mid d)] = \mathbb{C}(v(n), \mathbb{S}_{\mathrm{A}}(n \mid d))$. Since we normalized $\mathbb{V}(v(n)) = 1$ and $\mathbb{V}(\mathbb{S}_{\mathrm{A}}(n \mid d))$ does not depend on $v(\cdot)$, we can rewrite (G3) as

$$\max_{v(\cdot) \in \mathcal{V}} |\mathrm{Corr}(v(n), \mathbb{S}_{\mathrm{A}}(n \mid d))| . \tag{G4}$$

Intuitively, our goal is to find the micro values in $\mathcal{V}$ that are maximally correlated with the true scores of the model. This is precisely true for the simplest case in (G4), and approximately true for the general case, for which we have to choose a norm $\|\cdot\|$ to weight different parameters and it matters how micro values correlate.

Relationship to R^2

Instead of minimizing the exact objective in (G3), it is easier to regress parameters' scores on candidate sets of micro values and to keep those that maximize the sum of R^2 values across parameters. For intuition about why this simpler procedure approximates the exact objective, consider the case with $\dim(\theta) = 1$ but with possibly more than one micro moment.

Let $Y = (\mathbb{S}_{\mathrm{A}}(1 \mid d), \ldots, \mathbb{S}_{\mathrm{A}}(N_d \mid d))$ be a $N_d \times 1$ vector and let $X = (v(1), \ldots, v(N_d))'$ be a $N_d \times M_{\mathrm{M}}$ matrix. In vector-matrix form, the R^2 of a regression of Y on X is

$$R^2 = 1 - \frac{Y'(I - X(X'X)^{-1}X')Y}{(Y - \overline{Y})'(Y - \overline{Y})} . \tag{G5}$$

The X that maximizes the R^2 of this regression maximizes $(X'Y)'(X'X)^{-1}X'Y$. Equivalently, we could minimize the inverse of this expression, which is precisely the sample analog of the objective in (G3).

Implementing the Procedure with PyBLP

In practice, we replace the true θ_0 with a consistent estimate $\hat{\theta}$ and replace expectations with averages over many simulated micro observations $n \in \mathcal{N}_d$. It is straightforward to simulate micro data and compute scores with PyBLP. In Figure G1 we start with results from a problem (for example, from the Petrin, 2002, replication in Figure 5), simulate micro data,

and compute scores for it.

Values above micro moment counts in Figure G2 report Monte Carlo results for our procedure where we regress each parameter's score on candidate micro values,[108] sum the R^2 values across parameters and keep the set of candidate micro values that maximize this sum of R^2 values. Using more interactions requires collecting more summary statistics, but they help span the optimal score, improving the performance of the estimator. Which interactions are most correlated with the simulated scores depends on the simulation, but "$\mathbb{E}[y_{it} \mid j \neq 0]$" and "$\mathbb{E}[x_{jt} \cdot y_{it} \mid j \neq 0]$" do tend to be some of the first few micro moments retained by the procedure. For comparison, we also include values above "Standard" and "Optimal," which correspond to the first two rows in Table 6.

In Figure G3 we report analogous results for minimizing the exact objective in (G3), in which $\|\cdot\|$ is the Frobenius norm. Results are fairly similar once there are at least three micro moments. Finally, in Figure G4 we run a multivariate lasso regression of scores on all candidate micro values, tune the regularization parameter so that there is only the desired number of nonzero coefficients, and keep the corresponding micro values. The results are a bit worse than for the best subset approaches. However, lasso regressions are far more computationally tractable when there are many micro values, and lasso seems to give a good sense of which may be useful for an exercise that is in any case somewhat heuristic.

Figure G1: Computing Simulated Scores with PyBLP

```
# Simulate micro data
simulated_data = problem_results.simulate_micro_data(micro_dataset, seed=0)

# Compute scores, integrating over unobserved preferences with quadrature
score_integration = pyblp.Integration('product', size=7)
simulated_scores = problem_results.compute_micro_scores(micro_dataset, simulated_data, score_integration)
```

This Python code demonstrates how to simulate scores with PyBLP. After obtaining problem results as in the Petrin (2002) replication code in Figure 5, we simulate micro data and compute scores for each simulated observation.

---

[108]We interact $y_{i_n t_n}$, $1\{y_{i_n t_n} < \bar{y}_{t_n}\}$, $1\{y_{i_n t_n} \geq \bar{y}_{t_n}\}$, $1\{j_n \neq 0\}$, $x_{2 j_n t_n}$, $1\{x_{2 j_n t_n} < \bar{x}_{2 t_n}\}$, $1\{x_{2 j_n t_n} \geq \bar{x}_{2 t_n}\}$, and $\mathcal{S}_{j_n t_n}$ where $\bar{y}_t$ and $\bar{x}_{2t}$ are medians of y_{it} and x_{2jt} in market t.

Figure G2: R^2 Approach
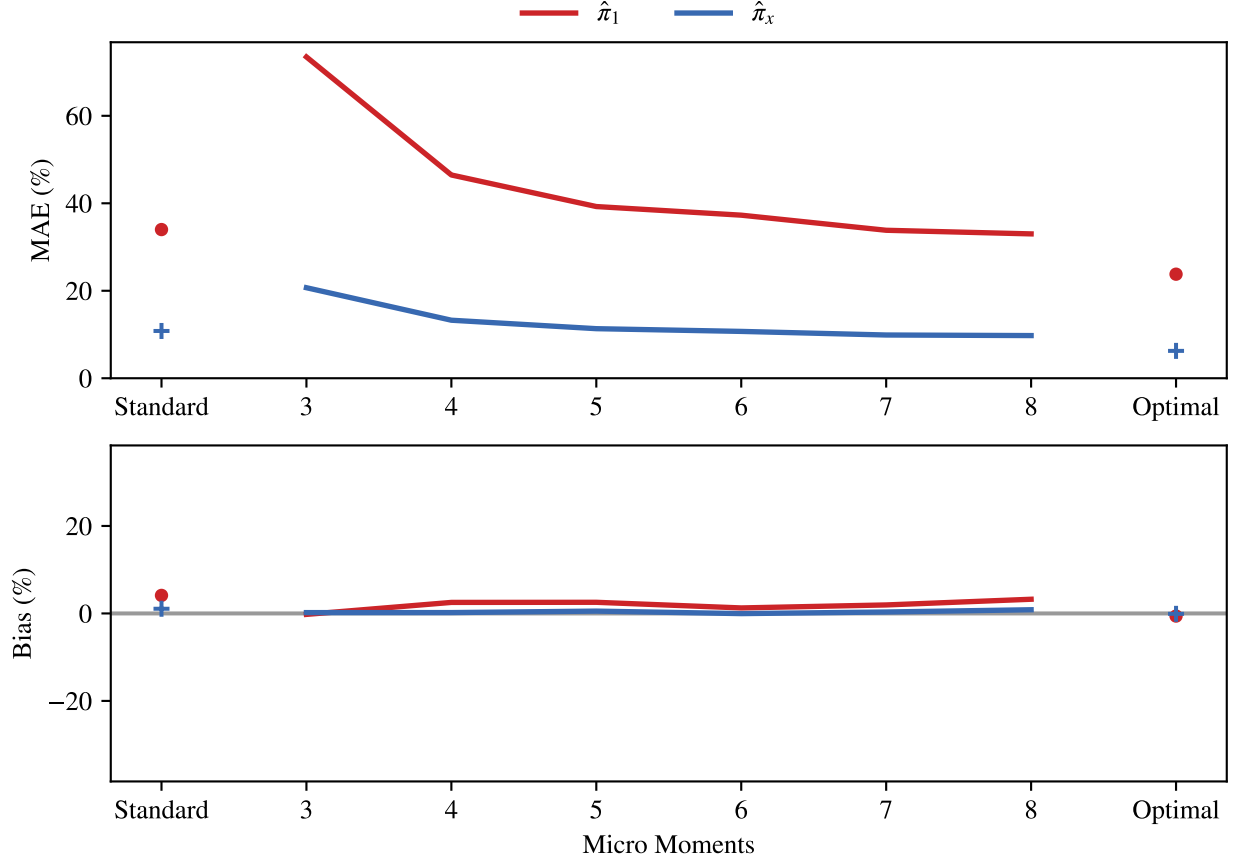
This figure reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets for approximations to the optimal micro moments. Values above "Standard" correspond to the same "$\mathbb{E}[y_{it} \mid j \neq 0]$" and "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" moments in the fourth row of Table 5. For other values, we use these same standard moments to obtain a first-stage estimator. Values above "Optimal" correspond to optimal micro moments. Values above micro moment counts correspond to the following procedure. First, we construct second-order polynomial interactions from $y_{i_n t_n}$, $1\{y_{i_n t_n} < \bar{y}_{t_n}\}$, $1\{y_{i_n t_n} \geq \bar{y}_{t_n}\}$, $1\{j_n \neq 0\}$, $x_{2j_n t_n}$, $1\{x_{2j_n t_n} < \bar{x}_{2t_n}\}$, $1\{x_{2j_n t_n} \geq \bar{x}_{2t_n}\}$, and $\mathcal{S}_{j_n t_n}$ where $\bar{y}_t$ and $\bar{x}_{2t}$ are medians of y_{it} and x_{2jt} in market t. Second, we exclude trivial interactions that do not depend on both i_n and j_n, since this would generate micro moments that would not depend on θ. Finally, we find the set of micro values that maximize the sum of R^2 values across regressions of each parameter's simulated scores on standardized versions of these micro values and use these micro values to form micro moments for the second GMM step.

Figure G3: Exact Approach
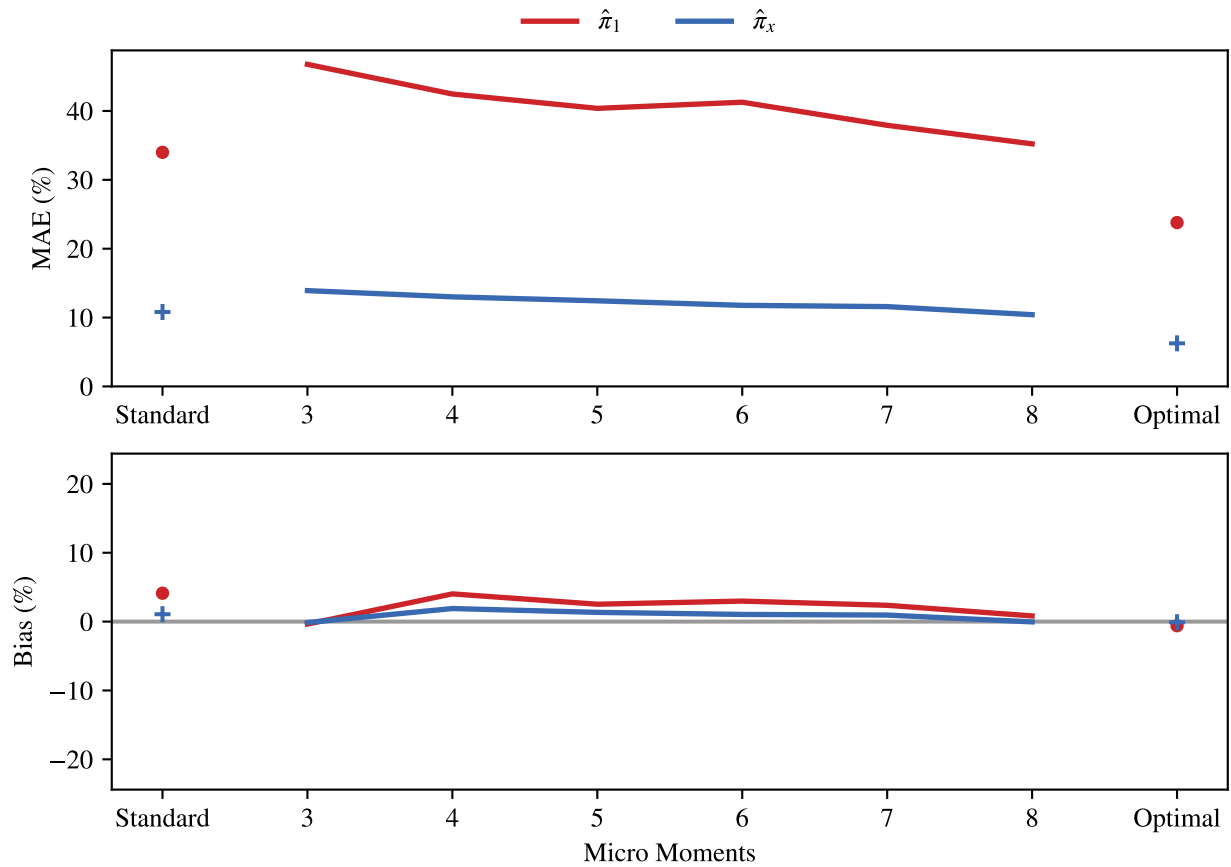
$\hat{\pi}_1$ $\hat{\pi}_x$

This figure is the same as Figure G2, but instead of using a best subset selection procedure that maximizes total R^2, instead minimizes the exact objective in (G3), in which $\|\cdot\|$ is the Frobenius norm.

Figure G4: Lasso Approach
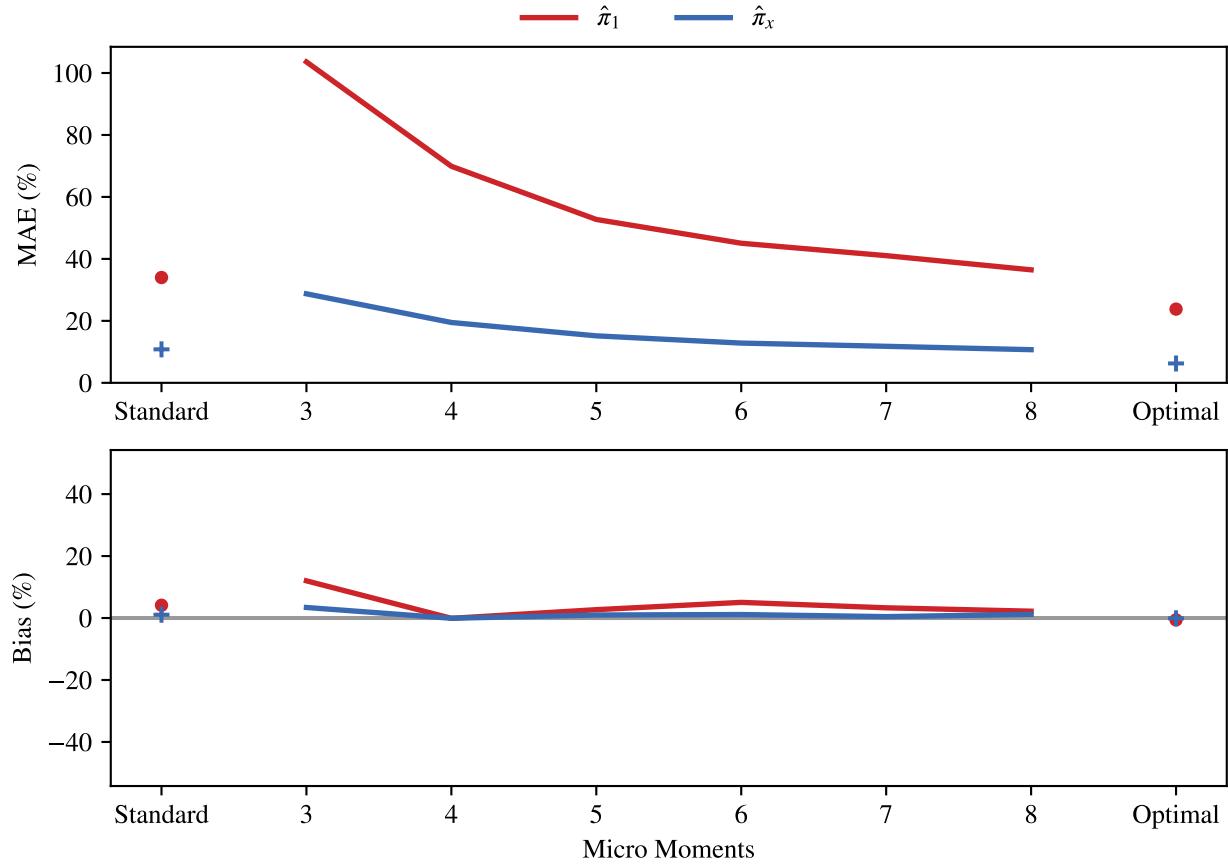
This figure is the same as Figure G2, but instead of using a best subset selection procedure that maximizes total R^2, instead tunes the regularization parameter for a multivariate lasso regression to obtain a specific number of micro values with nonzero coefficients.

H.   Monte Carlo Results for Standard Errors

For each table and figure in Section 7 and Appendices A and B, we also report analogous results for coverage and median bias of standard error estimators based on the expressions in Appendix D for estimating the asymptotic covariance matrix for $\hat{\theta}$:

$$\widehat{SE}(\hat{\theta}) = \sqrt{\mathrm{diag}\big((\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{S}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1}\big)/N_{\mathrm{A}}}. \tag{H1}$$

To evaluate the performance of the standard error estimators, we report coverage and median bias. Coverage is the percent of simulations in which a 95% confidence interval based on a standard error estimate covers the true parameter value.

The bias of standard error estimators is more difficult to evaluate than that of point estimates because comparisons are made with respect to a moving target. To compute a parameter's "true" standard error against which we compare its estimated standard error, we compute the standard deviation of the parameter's point estimate across all 1,000 simulations.

Table H1: Standard Errors, Demographic Variation

			Coverage (%)					Bias (%)				
Variation	Distributions	Markets	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\beta}_1$	$\hat{\beta}_x$	$\hat{\alpha}$	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\beta}_1$	$\hat{\beta}_x$	$\hat{\alpha}$
National	1	40	71.7	77.8	72.5	75.4	95.0	-81.6	-78.4	-48.7	-50.3	-8.6
States	50	40	86.8	87.4	90.1	88.0	94.3	-41.3	-36.1	-19.3	-20.8	-4.0
PUMAs	982	40	90.8	90.5	94.3	92.9	94.4	-62.1	-83.0	-9.7	-45.6	-11.7
National	1	80	76.9	76.0	77.0	75.0	95.4	-58.6	-65.5	-38.2	-51.7	-7.2
States	50	80	90.4	89.3	91.0	89.7	94.7	-49.7	-21.1	-16.7	-8.4	-2.6
PUMAs	982	80	92.0	91.6	93.8	91.9	95.3	-9.8	-10.9	-3.2	-7.8	-2.3

This table reports standard error coverage and median bias for Table 4.

Table H2: Standard Errors, Standard Micro Moments

	Coverage (%)		Bias (%)	
Micro Moments Shorthand	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\pi}_1$	$\hat{\pi}_x$
No Micro Moments	86.8	87.4	-41.3	-36.1
"$\mathbb{E}[y_{it} \mid j \neq 0]$"	92.3	91.7	-11.9	-11.5
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	92.2	84.5	-94.7	-95.8
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	84.4	81.2	-28.9	-34.3
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{E}[x_{2jt} \cdot y_{it} \mid j \neq 0]$"	87.3	85.4	-23.5	-28.5
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"	95.0	94.8	-1.6	-2.3
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	84.9	81.7	-28.2	-33.1

This table reports standard error coverage and median bias for Table 5.

Table H3: Standard Errors, Optimal Micro Moments and Compatibility

Micro Moments (plus $\mathbb{E}[y_{it} \mid j \neq 0]$)	Incompatible	Optimal	Coverage (%)		Bias (%)	
			$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\pi}_1$	$\hat{\pi}_x$
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			84.4	81.2	-28.9	-34.3
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"		Yes	92.5	94.0	-7.2	-5.0
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"			95.0	94.8	-1.6	-2.3
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"		Yes	92.9	93.5	-8.2	-6.1
"$\mathbb{E}[x_{2jt} \mid \tilde{y}_{it} < \bar{y}_t, j \neq 0]$"	Yes		92.4	92.0	-7.8	-8.6
"$\mathbb{E}[x_{2jt} \mid \tilde{y}_{it} < \bar{y}_t, j \neq 0]$"	Yes	Yes	34.6	54.7	-57.0	-56.5

This table reports standard error coverage and median bias for Table 6.

Figure H1: Standard Errors, Pooling Markets
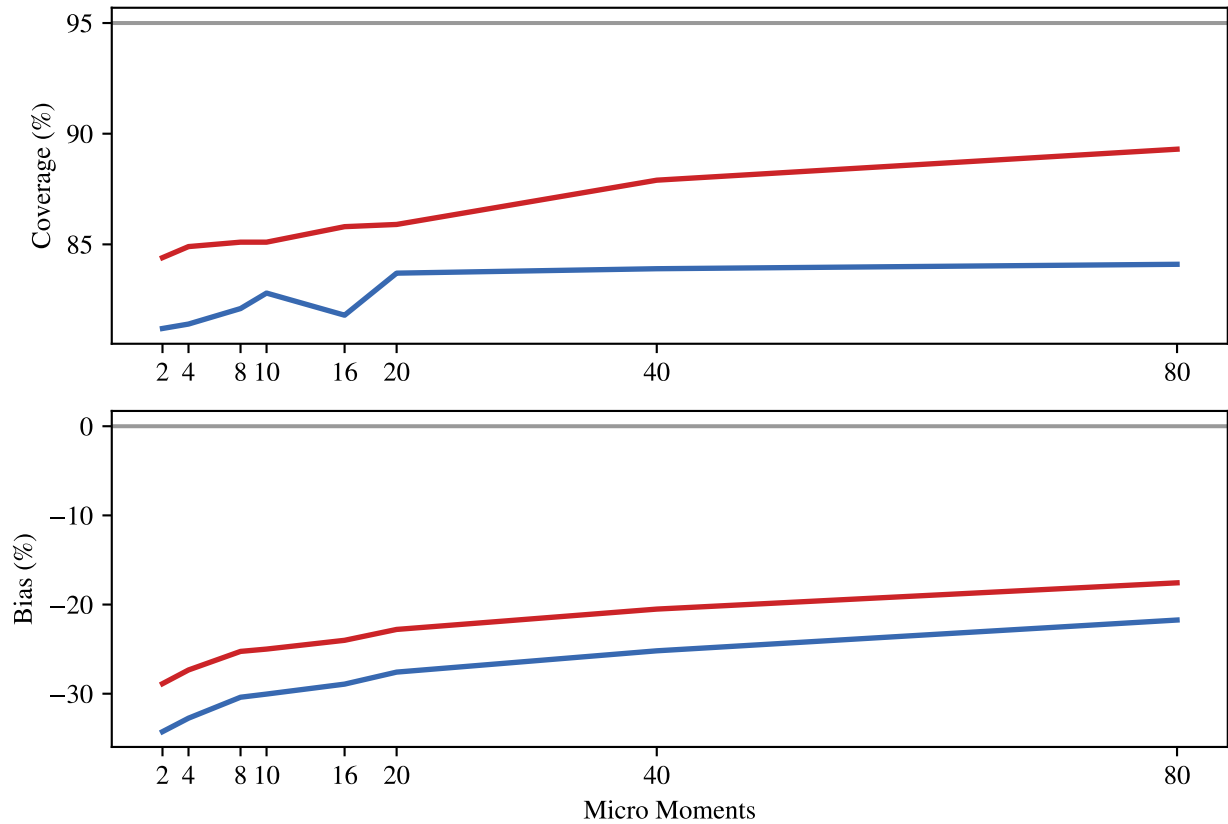
This figure reports standard error coverage and median bias for Figure 2.

Table H4: Standard Errors, Numerical Integration

Micro Moments (plus "$\mathbb{E}[y_{it} \mid j \neq 0]$")	Integration	Coverage (%)		Bias (%)	
		$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\pi}_1$	$\hat{\pi}_x$
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Quadrature	87.6	85.6	-22.0	-26.1
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Monte Carlo	84.4	81.2	-28.9	-34.3
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"	Quadrature	38.3	38.1	-75.9	-75.8
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"	Monte Carlo	95.0	94.8	-1.6	-2.3

This table reports standard error coverage and median bias for Table 7.

Figure H2: Standard Errors, Problem Scaling
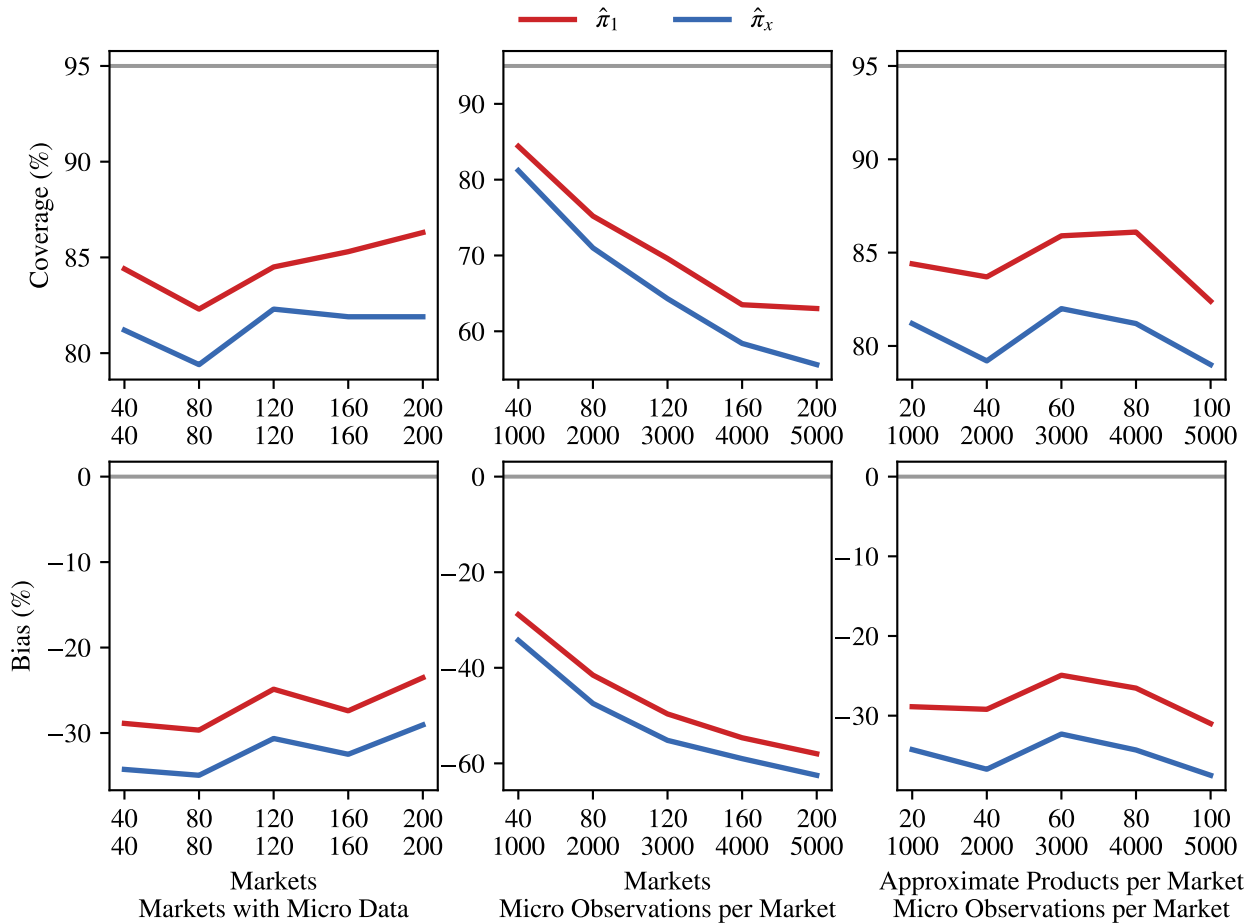
This figure reports standard error coverage and median bias for Figure 3.

Table H5: Standard Errors, Unobserved Heterogeneity

Micro Moments Shorthand	$\mathcal{J}_t = \mathcal{J}$	Optimal	Coverage (%)			Bias (%)		
			$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\sigma}_x$	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\sigma}_x$
No Micro Moments			86.0	85.7	92.0	-13.0	-9.7	-15.3
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			83.7	80.8	91.0	-28.4	-33.3	-16.0
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"		Yes	93.9	91.5	90.9	-9.5	-11.6	-16.6
No Micro Moments	Yes		99.7	99.8	99.5	65.0	70.0	227.2
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Yes		95.9	89.5	98.4	-5.3	-19.2	164.9
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Yes	Yes	93.7	74.3	95.7	-27.7	-73.8	-3.1

This table reports standard error coverage and median bias for Table 8.

Table H6: Standard Errors, Second Choices

Micro Moments (plus "$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$")	Optimal	Coverage (%)			Bias (%)		
		$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\sigma}_x$	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\sigma}_x$
No Second Choice Moments		96.1	89.1	98.5	-4.7	-19.7	160.7
"$\mathbb{C}(x_{2jt}, x_{2k(-j)t} \mid j, k \neq 0)$"		91.0	90.0	94.0	-23.2	-22.1	-23.6
"$\mathbb{E}[x_{2jt} + x_{2k(-j)t} \mid j, k \neq 0]$"		90.2	86.5	93.3	-25.4	-32.7	-55.3
"$\mathbb{P}(x_{2k(-j)t} < \bar{x}_{2t} \mid x_{2jt} \gtreqless \bar{x}_{2t}, j, k \neq 0)$"		91.2	89.9	92.2	-25.9	-34.5	-49.6
"$\mathbb{P}(x_{2k(-j)t} < \bar{x}_{2t} \mid x_{2jt} \gtreqless \bar{x}_{2t}, j, k \neq 0)$"	Yes	87.7	86.1	76.9	-27.8	-26.0	-50.6

This table reports standard error coverage and median bias for Table 9.

Table H7: Standard Errors, Lognormal Price Coefficient

Micro Moments Shorthand	Optimal	Coverage (%)			Bias (%)		
		$\hat{\pi}_1$	$\hat{\pi}_p$	$\hat{\sigma}_p$	$\hat{\pi}_1$	$\hat{\pi}_p$	$\hat{\sigma}_p$
No Micro Moments		63.4	70.0	90.0	-45.3	-42.7	-38.4
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$"		87.7	87.7	88.8	-22.8	-24.3	-19.8
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$, $\mathbb{C}(p_{jt}, y_{it}^2 \mid j \neq 0)$"		87.5	87.3	89.0	-23.0	-24.4	-20.8
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$, $\mathbb{C}(p_{jt}, y_{it}^2 \mid j \neq 0)$"	Yes	92.4	91.1	88.7	-12.1	-16.0	-21.2

This table reports standard error coverage and median bias for Table A1.

Table H8: Standard Errors, Nesting Parameter

Micro Moments Shorthand	$\mathcal{J}_t = \mathcal{J}$	Optimal	Coverage (%)			Bias (%)		
			$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\rho}$	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\rho}$
No Micro Moments			90.3	91.2	95.9	-19.3	-22.2	-0.5
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			80.9	78.9	95.2	-32.6	-37.6	-0.9
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Yes		99.5	99.0	99.7	128.0	96.6	209.8
and "$\mathbb{C}(x_{2jt}, x_{2kt} \mid j, k \neq 0)$"	Yes		90.5	85.3	92.0	-24.9	-31.4	-68.5
and "$\mathbb{E}[x_{2jt} + x_{2kt} \mid j, k \neq 0]$"	Yes		90.7	86.5	90.3	-26.5	-33.5	-69.8
and "$\mathbb{P}(x_{2kt} < \bar{x}_{2t} \mid x_{2jt} \gtreqless \bar{x}_{2t}, j, k \neq 0)$"	Yes		90.7	85.1	92.6	-24.9	-31.4	-53.5
and "$\mathbb{P}(h(j) = h(k) \mid j, k \neq 0)$"	Yes		90.1	84.0	88.5	-26.8	-33.8	-43.5
and "$\mathbb{P}(h(j) = h(k) \mid j, k \neq 0)$"	Yes	Yes	84.1	87.5	79.5	-40.6	-39.7	-70.4

This table reports standard error coverage and median bias for Table B1.

I.   Monte Carlo Results for Counterfactuals

For each table and figure in Section 7 and Appendices A and B, we also report analogous results for median absolute error (MAE) and median bias of estimates from a counterfactual in which we make high-x_{2jt} goods relatively more expensive to produce. For each product j, we increase its marginal cost c_{jt} in Footnote 61 by $2 \times x_{2jt}$ and subtract 6 to keep costs the same on average.

We then re-compute equilibrium prices p_{jt} with the fixed point approach of Morrow and Skerlos (2011) and report the associated change in consumer surplus, separately for high and low-income consumers with income above and below the median $\bar{y}_t$ in each market. With unit market sizes $\mathcal{M}_t = 1$ and up to an arbitrary constant, the consumer surplus of those with income $\underline{y}_t \leq y_{it} < \bar{y}_t$ is

$$CS = \frac{1}{\alpha} \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}_t} \frac{w_{it} \cdot 1\{\underline{y}_t \leq y_{it} < \bar{y}_t\}}{\sum_{\tau \in \mathcal{T}} \sum_{\iota \in \mathcal{I}_\tau} w_{\iota\tau} \cdot 1\{\underline{y}_\tau \leq y_{\iota\tau} < \bar{y}_\tau\}} \log \left(1 + \sum_{j \in \mathcal{J}_t} \exp(\delta_{jt} + \mu_{ijt}) \right). \quad (\text{I1})$$

One complication comes from eliminating cross-market choice set variation in Tables 8, 9, and B1 by using the same choice set $\mathcal{J}_t = \mathcal{J}$ in each market. This drastically reduces the effective number of observations used to estimate linear parameters β to around $|\mathcal{J}| \approx 20$. Although the nonlinear parameters (Π, Σ) can be estimated well with appropriate micro moments, β is estimated poorly, obscuring results from the counterfactual. To still have meaningful results in these three tables, we set $\hat{\beta} = \beta_0$ before computing counterfactuals whenever $\mathcal{J}_t = \mathcal{J}$.

Table I1: Counterfactual, Demographic Variation

			MAE (%)		Bias (%)	
Variation	Distributions	Markets	Low y_{it}	High y_{it}	Low y_{it}	High y_{it}
National	1	40	48.8	111.0	-18.7	9.9
States	50	40	24.4	58.0	-4.2	4.8
PUMAs	982	40	18.2	53.2	-1.4	1.1
National	1	80	37.4	103.6	-17.7	25.0
States	50	80	16.8	46.3	-2.6	0.1
PUMAs	982	80	11.9	46.8	-1.1	2.6

This table reports median absolute error (MAE) and median bias of each CS for Table 4.

Table I2: Counterfactual, Standard Micro Moments

Micro Moments Shorthand	MAE (%)		Bias (%)	
	Low y_{it}	High y_{it}	Low y_{it}	High y_{it}
No Micro Moments	24.4	58.0	-4.2	4.8
"$\mathbb{E}[y_{it} \mid j \neq 0]$"	20.6	52.5	0.7	-2.7
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	14.9	33.8	1.0	-3.1
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	14.2	33.8	0.1	-2.5
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{E}[x_{2jt} \cdot y_{it} \mid j \neq 0]$"	14.5	33.4	-0.2	-2.4
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"	15.6	36.2	-0.1	-3.5
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	14.3	33.7	0.2	-2.7

This table reports median absolute error (MAE) and median bias of each *CS* for Table 5.

Table I3: Counterfactual, Optimal Micro Moments and Compatibility

Micro Moments (plus $\mathbb{E}[y_{it} \mid j \neq 0]$)	Incompatible	Optimal	MAE (%)		Bias (%)	
			Low y_{it}	High y_{it}	Low y_{it}	High y_{it}
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			14.2	33.8	0.1	-2.5
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"		Yes	14.2	32.5	-0.6	-0.2
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"			15.6	36.2	-0.1	-3.5
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"		Yes	14.2	32.9	-0.3	-0.8
"$\mathbb{E}[x_{2jt} \mid \tilde{y}_{it} < \bar{y}_t, j \neq 0]$"	Yes		14.6	37.7	-1.1	-1.5
"$\mathbb{E}[x_{2jt} \mid \tilde{y}_{it} < \bar{y}_t, j \neq 0]$"	Yes	Yes	15.5	35.6	-3.3	-0.3

This table reports median absolute error (MAE) and median bias of each *CS* for Table 6.

Figure I1: Counterfactual, Pooling Markets
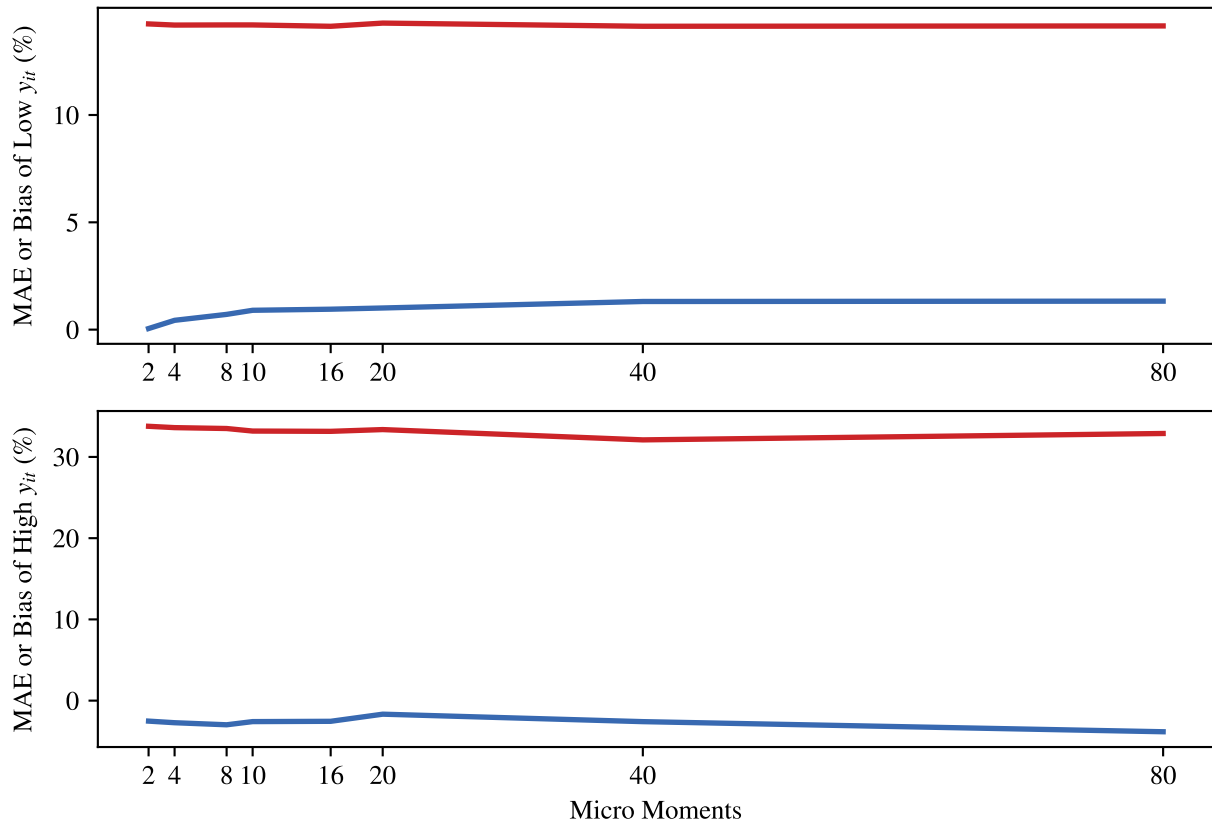
This figure reports median absolute error (MAE) and median bias of each *CS* for Figure 2.

Table I4: Counterfactual, Numerical Integration

Micro Moments (plus "$\mathbb{E}[y_{it} \mid j \neq 0]$")	Integration	MAE (%)		Bias (%)	
		Low y_{it}	High y_{it}	Low y_{it}	High y_{it}
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Quadrature	21.5	46.0	-2.7	-0.8
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Monte Carlo	14.2	33.8	0.1	-2.5
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"	Quadrature	25.7	53.6	-0.9	-7.9
"$\mathbb{E}[x_{2jt} \mid y_{it} < \bar{y}_t, j \neq 0]$"	Monte Carlo	15.6	36.2	-0.1	-3.5

This table reports median absolute error (MAE) and median bias of each *CS* for Table 7.

Figure I2: Counterfactual, Problem Scaling
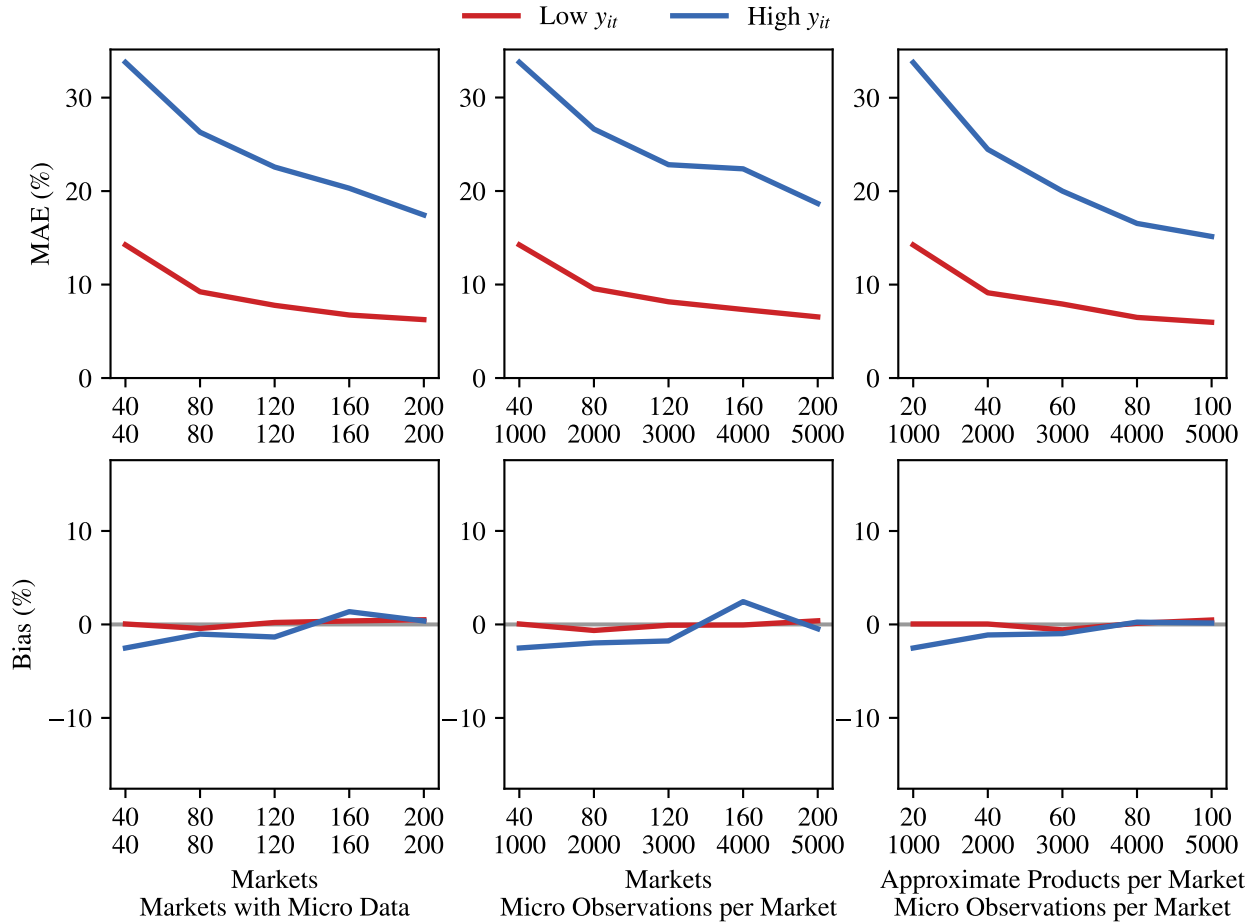
This figure reports median absolute error (MAE) and median bias of each *CS* for Figure 3.

Table I5: Counterfactual, Unobserved Heterogeneity

Micro Moments Shorthand	$\mathcal{J}_t = \mathcal{J}$	Optimal	MAE (%) Low y_{it}	MAE (%) High y_{it}	Bias (%) Low y_{it}	Bias (%) High y_{it}
No Micro Moments			76.8	39.6	0.1	-4.9
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			40.9	22.1	3.7	1.8
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"		Yes	43.6	22.1	2.6	0.9
No Micro Moments	Yes		32.3	30.5	-0.6	0.4
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Yes		20.8	21.6	0.3	-3.4
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Yes	Yes	21.8	22.1	-1.2	-9.0

This table reports median absolute error (MAE) and median bias of each CS for Table 8.

Table I6: Counterfactual, Second Choices

Micro Moments (plus "$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$")	Optimal	MAE (%) Low y_{it}	MAE (%) High y_{it}	Bias (%) Low y_{it}	Bias (%) High y_{it}
No Second Choice Moments		20.8	21.7	0.4	-3.4
"$\mathbb{C}(x_{2jt}, x_{2k(-j)t} \mid j, k \neq 0)$"		6.5	6.0	0.1	0.2
"$\mathbb{E}[x_{2jt} + x_{2k(-j)t} \mid j, k \neq 0]$"		2.8	2.6	-0.0	0.0
"$\mathbb{P}(x_{2k(-j)t} < \bar{x}_{2t} \mid x_{2jt} \gtreqless \bar{x}_{2t}, j, k \neq 0)$"		5.4	5.1	-0.4	-0.4
"$\mathbb{P}(x_{2k(-j)t} < \bar{x}_{2t} \mid x_{2jt} \gtreqless \bar{x}_{2t}, j, k \neq 0)$"	Yes	2.0	1.9	0.1	-0.1

This table reports median absolute error (MAE) and median bias of each CS for Table 9.

Table I7: Counterfactual, Lognormal Price Coefficient

Micro Moments Shorthand	Optimal	MAE (%) Low y_{it}	MAE (%) High y_{it}	Bias (%) Low y_{it}	Bias (%) High y_{it}
No Micro Moments		38.4	83.4	-4.1	4.5
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$"		11.4	17.4	0.6	-0.6
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$, $\mathbb{C}(p_{jt}, y_{it}^2 \mid j \neq 0)$"		11.5	17.5	0.6	-0.6
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(p_{jt}, y_{it} \mid j \neq 0)$, $\mathbb{C}(p_{jt}, y_{it}^2 \mid j \neq 0)$"	Yes	11.5	17.7	0.8	-0.5

This table reports median absolute error (MAE) and median bias of each CS for Table A1.

Table I8: Counterfactual, Nesting Parameter

Micro Moments Shorthand	$\mathcal{J}_t = \mathcal{J}$	Optimal	MAE (%) Low y_{it}	MAE (%) High y_{it}	Bias (%) Low y_{it}	Bias (%) High y_{it}
No Micro Moments			24.9	66.4	-0.5	-2.7
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			13.7	33.6	-1.0	-0.6
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	Yes		44.6	44.6	-0.8	0.3
and "$\mathbb{C}(x_{2jt}, x_{2kt} \mid j, k \neq 0)$"	Yes		6.0	6.7	0.1	-0.2
and "$\mathbb{E}[x_{2jt} + x_{2kt} \mid j, k \neq 0]$"	Yes		6.4	6.3	-0.0	-0.1
and "$\mathbb{P}(x_{2kt} < \bar{x}_{2t} \mid x_{2jt} \gtreqless \bar{x}_{2t}, j, k \neq 0)$"	Yes		4.8	4.8	0.0	-0.1
and "$\mathbb{P}(h(j) = h(k) \mid j, k \neq 0)$"	Yes		1.7	1.8	0.1	-0.1
and "$\mathbb{P}(h(j) = h(k) \mid j, k \neq 0)$"	Yes	Yes	1.1	1.2	0.0	0.1

This table reports median absolute error (MAE) and median bias of each CS for Table B1.

J. Data Details for Estimating Seattle Soft Drink Demand

In this appendix, we discuss all of the decisions we make when constructing the data that we use for our empirical example in Section 8 where we predict substitution from Seattle's 2018 sweetened beverage tax (SBT). We also discuss possible alternative decisions and what these would imply for a micro BLP approach to estimation.

Markets

The first step is to define markets $t \in \mathcal{T}$. We define a market as the entirety of the city of Seattle in each of the 40 quarters from 2007Q1 to 2016Q4.[109] We focus on food and mass merchandiser stores (NielsenIQ channel codes F and M) to keep our sample more manageable, but could have also included convenience stores. The NielsenIQ Retail Scanner dataset starts to have Seattle data in 2007Q1, and after 2016Q4, two large retailers in Seattle drop out of the NielsenIQ sample. Ending far before the January 2018 implementation of the actual tax is in the spirit of this type of prediction exercise, since we do not want to use variation from the tax to estimate demand, but rather to validate our predictions.

Aggregate sales data are weekly, but we aggregate to a quarterly frequency to partially alleviate concerns about stockpiling, which the standard BLP model does not account for.[110] Aggregating to a yearly frequency would further alleviate stockpiling concerns and would reduce computational costs, but it would eliminate a great deal of price variation that is needed estimate the price elasticity of demand.

Adding regions other than Seattle as additional markets would increase the amount of cross-market choice set and demographic variation. But if demand parameters are assumed to be the same across geographic regions, doing so would also require an assumption that preferences for soft drinks are similar across included geographies. We only consider Seattle because doing so does not require making this additional assumption. Thankfully, the combination of aggregate and micro data at our disposal has enough variation for us to predict the effects of the 2018 SBT tax with reasonable precision.

One important concern specific to the tax prediction exercise is that after the tax was implemented, consumers may have switched to stores right outside the city where prices were

---

[109]We use Seattle's 3-digit ZIP code 981 to identify stores and households in the city in NielsenIQ data. We use its five Public Use Microdata Areas (PUMAs) to identify the population of Seattle households in the American Community Survey (ACS).

[110]Existing studies (e.g., Hendel and Nevo, 2006) have documented stockpiling at a weekly frequency. Allcott et al. (2019) find no evidence of stockpiling at the quarterly frequency for soft drinks in NielsenIQ data.

lower. However, unlike other SBTs like in Philadelphia with strong evidence of cross-border shopping (e.g., Roberto et al., 2019; Cawley et al., 2019; Seiler et al., 2021), there did not seem to be substantial cross-border shopping after the Seattle tax (Powell and Leider, 2020).

If cross-border shopping were a concern, one could include stores in the area around Seattle into each quarter's market, define each product j as a product-store combination, and incorporate geographic distance or travel time d_{ijt} between between stores and households into the demand system.[111] To do so, we would need demographic data and store locations at a more detailed level, for example 5-digit ZIP code.[112] When running the tax counterfactual, consumers in the model would trade-off lower prices with distance, giving a more reasonable estimate of cross-border shopping. For example, Chen et al. (2022) use a variant of Grieco et al.'s (2023) CLER estimator to estimate the importance of travel time in generating cross-border effects around the implementation of Philadelphia's SBT tax.

Another approach to geographic market definition would be to split up Seattle into a separate market for each geographic part of Seattle (e.g., north, downtown, and south), each group of stores (e.g., by retailer), or both. Again, there is a tradeoff. Splitting up Seattle into multiple cross-sectional markets would generate useful cross-market variation, but it would require an assumption that consumers' choice sets are truly defined by these cross-sectional segments. If estimating how demand varies with demographics, as in our empirical example, this would also require demographic data at a more detailed level.

Demographics

Since our market definition is the entirety of Seattle at a quarterly frequency, we collect household-level demographic data for the entirety of Seattle from the American Consumer Survey (ACS) at the highest frequency possible, which is annual.[113] In the ACS, Seattle is split up into five Public Use Microdata Areas (PUMAs), which we combine.

An approach that splits Seattle up into multiple geographic regions could use different PUMAs for different market definitions. If instead of geography, Seattle were instead split up by groups of stores, Census data would be less useful because it does not contain information about where households typically shop for groceries. An alternative approach would be to

---

[111]PyBLP fully supports variables like d_{ijt}, which we call "product-specific demographics."

[112]NielsenIQ stores only have 3-digit ZIP codes, but one can impute their 5-digit ZIP codes taking the trip-weighted average centroid of ZIP codes of households in the NielsenIQ Consumer Panel data who shop at the store (DellaVigna and Gentzkow, 2019; Chen et al., 2022).

[113]The time dimension turns out to not be very important for our setting because there is very little time series variation during this period in the demographics we consider: the share of households that are high income and have children.

sample from households in the NielsenIQ Consumer Panel dataset who shop at each group of stores; however, the number of households would be quite small.

In general, constructing demographic distributions from NielsenIQ households can work if estimating demand over larger geographic regions where the number of households being sampled from is quite large.[114] However, one might still want to merge in Census data to construct sampling weights that guarantee sampling from NielsenIQ households is similar to sampling from the actual population.[115] We discuss sampling weight construction below in the micro data subsection.

Since our model only includes two binary demographic variables (y_{it} includes high income and children indicator variables), it is enough to know the share of households in Seattle in each year in each of the four demographic bins. Without any unobserved preference heterogeneity, these shares are integration weights w_{it}. We define a high income household as one with pre-tax household income above \$67,106, the median household income of Washington state in 2016. We deflate all dollar-valued variables, including income, using the 2016 Consumer Price Index (CPI). Households with children are those with at least one member below the age of 18.

To incorporate unobserved heterogeneity ν_{it}, we expand each of our four demographic groups into as many integration nodes are needed for our quadrature rule. We then multiply the demographic shares by the quadrature rule's integration weights to get the final integration weights w_{it}.

For models with continuous demographic variables, such as real-valued income, one approach is to fit a parametric distribution to the demographic,[116] and incorporate it with a quadrature rule just as we incorporate unobserved preference heterogeneity. If there is more than one continuous demographic, it is often easier to use simple Monte Carlo methods, resampling a large number of households from the data, along with Monte Carlo draws for unobserved heterogeneity. In this case, integration weights w_{it} would simply be $1/|\mathcal{I}_t|$.

Lastly, by computing household-level demographic shares (weighting by the household

---

[114]For smaller geographic regions and noisier demographic distributions, one approach is to account for this sampling error by resampling from the small number of households during a nonparametric bootstrap. Alternatively, PyBLP supports estimating the contribution of demographic sampling error to asymptotic moment covariances through resampling demographic at the parameter estimates.

[115]The NielsenIQ data provides projection weights, but our understanding is that these are more useful for constructing a nationally-representative sample, not a sample that is representative for a more specific region.

[116]For example, Backus et al. (2021) use NielsenIQ income bins to fit parametric income distributions at the Designated Market Area (DMA)-chain level, which generally contains many more NielsenIQ households than just Seattle.

weights provided by the ACS), we are assuming that each consumer in the model is a Seattle household. Another approach would be defining consumers as individuals within households, in which case we would want to compute demographic shares at the individual level. We focus on households because households tend to do grocery shopping as a unit and that is how NielsenIQ records purchases.

Products

In the NielsenIQ Retail Scanner dataset, we define a product j as a Uniform Product Code (UPC)-retailer combination.[117] Defining a separate product for each of the five retailers in our sample at which it is sold allows us to exploit cross-retailer price variation, which we will use to construct our price instrument.

When there are a great deal of products in a market,[118] it may be much more computationally tractable to aggregate UPCs up to the brand level. Similarly, in many papers estimating demand for automobiles, products are defined at more aggregated levels than trim. We choose to define products at the more fine UPC level because this allows us to estimate differential preferences for small or individual-sized drinks, which following Powell and Leider (2020) we define as single units no more than one liter in volume.[119] However, we do cluster our standard errors for the aggregate data at the brand level, since we expect marketing and other dimensions of unobserved quality to be strongly correlated within brand.

To construct each set of products $\mathcal{J}_t$, we limit our analysis to NielsenIQ product modules for soft drinks and fruit drinks.[120] There are other modules containing juices with added sugar that the Seattle SBT tax affected, but for simplicity we do not include these other modules and drop juice drinks that are in the above modules.[121]

If we wished to study substitution to other beverage categories, such as juices or other sugary goods, we could have included additional product modules. We choose to focus on substitution to diet drinks for simplicity, and because diet drinks were excluded from the

---

[117]Throughout this appendix, "UPC" will refer to a UPC and version number combination.

[118]For example, this would be the case if we defined product j as each UPC-store to incorporate distances d_{ijt} into estimation.

[119]This product characteristic ends up being important because it is highly correlated with price per ounce and preferences for it differ strongly by demographic group.

[120]The modules that we consider are "Soft Drinks - Carbonated," "Soft Drinks - Low Calorie," "Fruit Drinks - Canned," "Fruit Drinks - Other Container," and "Fruit Drinks & Juices - Cranberry."

[121]We identify juice drinks as those with "JC" surrounded by word boundaries in their UPC descriptions. These constitute only 3.2% of the total ounces purchased in our data because the product modules we consider do not include those that are primarily juice.

Seattle tax. Using the same instrument that we describe shortly, although in a national setting, Allcott et al. (2019) only find evidence of substitution from sugar-sweetened beverages to diet drinks among beverage categories. Comparing with NielsenIQ data from Portland, Oddo et al. (2021) and Powell and Leider (2022) do find evidence of some small substitution to sweets and alcohol, respectively, so a more complete analysis could benefit from including such categories in the demand system as well. For example, Zhen et al. (2014) estimate an Exact Affine Stone Index (EASI) demand model (Lewbel and Pendakur, 2009) that includes 23 different categories related to soft drinks to evaluate the impact of SBTs. The micro BLP approach estimates demand in characteristics space, rather than in complementary product space approaches like EASI and related Almost Ideal Demand System (AIDS) models (Deaton and Muellbauer, 1980).

A common concern with datasets that include many products is that those with very small quantities purchased are in fact unavailable to most consumers, have low-quality data, or have quantities that are noisily estimated. Within each quarter, we combine all products in the bottom 5% of ounces sold with the outside good, effectively dropping these very small-quantity products from our analysis. Among these dropped products, the largest only constituted 0.008% of total ounces purchased during its quarter. We also drop less than 0.03% of product-quarters with size units not denoted in ounces. Across quarters, there are an average of $|\mathcal{J}_t| \approx 1,954$ products with a standard deviation of 154.

In addition to the small-size indicator, we also construct a diet indicator from formula, type, and UPC descriptions.[122] To identify which products were taxed, we use a manual classification that was created and graciously provided to us by the authors and research team of Powell and Leider (2020). Their data are slightly different, but their classification covers 99% of products in our sample in 2016, and we manually match a remaining few products, primarily using the diet indicator.

We define prices p_{jt} as the quantity-weighted price per ounce (deflated by the quarterly 2016 CPI) of UPC-retailer j in quarter t, and quantities q_{jt} as total ounces sold of j in quarter t. An alternative approach would be to define quantities as units sold, and prices as price per unit. We find that using raw units instead of volume leads to a great deal of price variation that is hard to explain, except by constructing many additional product characteristics. Price per unit volume tends to be more uniform and easier to explain with a demand model.

---

[122]We define diet drinks as those with formula or type descriptions equal to "DIET", "LIGHT", "REDUCED CALORIE", "LITE", "LOW CALORIE", or "LOW CALORIE CAFFEINE FREE" or with UPC descriptions that include "DT" or "LT" surrounded by word boundaries.

Instruments

For prices, we construct a Hausman (1996)-type instrument very similar to the one used by Allcott et al. (2019), which exploits the tendency of retailers to vary prices independently of one another, but uniformly across their own stores (DellaVigna and Gentzkow, 2019). Across all NielsenIQ Designated Market Areas (DMAs) that do not include Seattle,[123] we compute the quantity-weighted average price per ounce for each UPC-quarter and UPC-retailer-quarter. Our instrument z_{1jt} for the price p_{jt} of UPC-retailer j in quarter t is the difference of these two numbers,[124] which reflects retailer-specific deviations of prices in non-Seattle markets. Estimating the simple logit model, Kleibergen and Paap's (2006) F-statistic is 1,003.[125]

To identify σ_p, the degree of unobserved preference heterogeneity for prices, we use the "quadratic" version of Gandhi and Houde's (2020) differentiation IVs, which we discuss in Section 3 and also use in our Monte Carlo experiments in Section 7. First, to deal with price endogeneity, we construct predicted prices $\hat{p}_{jt}$ from a linear regression of prices p_{jt} on the Hausman-type instrument, along with product-retailer and retailer-quarter fixed effects. The differentiation IV $z_{2jt} = \sum_{k \neq j} (\hat{p}_{kt} - \hat{p}_{jt})^2$ is the sum of squared deviations of predicted prices from other products in that quarter. This reflects each product's exogenous degree of isolation in product space, as measured by its price.

Since there is a great deal of cross-quarter variation in both prices and the Hausman-type instrument, with this differentiation IV we are able to estimate σ_p fairly precisely with only aggregate variation. Since there is very little cross-quarter variation in the distribution of demographics in Seattle, we do not attempt to identify coefficients in Π with aggregate variation. Indeed, when we try to estimate the model with instruments for these coefficients,[126] we get unsurprisingly noisy estimates that severely corrupt other estimates of interest.

---

[123]Seattle's DMA code 819 includes a number of surrounding counties.

[124]For a small number of UPC-retailers with no observed sales in the NielsenIQ data outside the Seattle DMA, we set $z_{1jt} = 0$. This means that we only use instrument variation from UPC-retailers in other DMAs to identify the price elasticity.

[125]We regress log quantities on prices per ounce, using our Hausman-type instrument. Like in Table 11, we absorb product-retailer and retailer-quarter fixed effects, and cluster standard errors by brand. Retailer-quarter fixed effects absorb variation in market sizes and outside quantities.

[126]If we were working with many different cross-sectional geographic markets, we could attempt to identify Π with only aggregate variation by including instruments that interact demographic means with product characteristics and their differentiation IVs.

Market Sizes

To convert quantities q_{jt}, in ounces, into market shares, $\mathcal{S}_{jt} = q_{jt}/\mathcal{M}_t$, however q_{0t} is unobserved, and therefore we need to make an assumption about each quarter's market size $\mathcal{M}_t$, also in ounces. We assume that each market size $\mathcal{M}_t$ is equal to an estimate of the number of trips at the stores in our data in quarter t, multiplied by 720 ounces per trip. We choose 720 because in a histogram of ounces purchased per trip,[127] 720 ounces (or four 144 ounce packages of 12 cans each) was the last significant spike in the right tail (95th percentile), suggesting that it is the maximum "reasonable" purchase size per trip of beverages in our analysis.

To estimate the number of trips per quarter, we combine the NielsenIQ Retail Scanner data with the Consumer Panel data. Across all Seattle retailers in our product data, we compute each quarter's total revenue across all dry grocery product groups.[128] Across all Seattle households and trips to Seattle stores in our micro data, discussed below, we compute each quarter's weighted average revenue per trip across these same product groups.[129] Our estimate of the number of trips per quarter is the ratio of these two numbers. We restrict our attention to dry grocery rather than all goods (e.g., laundry detergent), because we do not want to include non-grocery trips that are unlikely to include beverage purchases.

Another approach that aims to get a sense of foot traffic in the NielsenIQ data is to regress the total quantity of inside goods on goods that are usually purchased during a typical grocery trip, like milk and eggs (e.g., Backus et al., 2021). Market sizes would then be the predicted number of inside goods purchased as a function of, say, milk and eggs, scaled by a common number to target a certain outside good share. This targeted outside good share is similar to our assumed 720 ounces per trip, and ultimately requires a strong assumption about the size of the market. However, this approach, along with the one that we use in this paper, both capture reasonable cross-market variation in the market size. The benefit of a regression approach is it does not require an estimate of revenue per trip, which can be quite noisy if working with markets that have very few households in the Consumer Panel data.

Perhaps the most common approach is to convert total population of the surrounding

---

[127]We compute this histogram by aggregating ounces purchased in the below micro data to the trip level, and then computing the weighted average of these ounces per trip in different bins. We discuss household sampling weights below.

[128]Perhaps confusingly, the department code of 1 corresponding to dry grocery includes all beverage categories we consider. It contains most standard grocery products.

[129]We discuss household sampling weights below when describing how we construct micro data.

area into quantity units with another conversion factor. This approach seems less reasonable if markets are defined by groups of stores within a region, but could be reasonable for markets defined by clear geographic regions. If we were to estimate the typical number of trips per quarter from the Consumer Panel data, multiply by the total population of Seattle, and scale by 720 ounces, our outside share would be very large because it would include all beverage sales that NielsenIQ does not cover. We could scale down our estimate by some factor of NielsenIQ coverage, but we found estimating trips directly from NielsenIQ to be more straightforward.

For our counterfactual, it would be problematic if the outside good contained a large volume of taxed beverages. It would be unclear by how much to increase the price of the outside good, or equivalently, decrease the price of all inside goods. And not increasing the price of the outside good would result in an unreasonable degree of substitution to the outside good.

In general, a biased market size will also bias demand estimates and counterfactual substitution patterns. One exception is the simple logit model with market fixed effects, which absorb variation from outside quantities and the market size, so that parameter estimates are unaffected by the choice of market size. However, counterfactual substitution will still be affected. The own- and cross-price elasticities of demand for the simple logit model are $\epsilon_{jjt} = \alpha \cdot p_{jt} \cdot (1 - \mathcal{S}_{jt})$ and $\epsilon_{jkt} = -\alpha \cdot p_{jt} \cdot \mathcal{S}_{kt}$, in which $\alpha < 0$ is the coefficient on price. Since market shares are usually very small, $(1 - \mathcal{S}_{jt}) \approx 1$, and the own-price elasticity ϵ_{jjt} is mostly unaffected by the choice of market size. However, the elasticity of substitution to the outside option ϵ_{j0t} is biased upwards by a large outside share $\mathcal{S}_{0t}$ due to a too-large market size $\mathcal{M}_t$, and substitution to other inside goods $k \notin \{0, j\}$ is biased downward.

Incorporating random coefficients can help discipline substitution with data rather than often untestable assumptions about the size of the market. Ideally, we would like to estimate a random coefficient on the outside good, or equivalently, on a constant for all inside goods. The degree of preference heterogeneity for the outside good versus inside goods is closely related to how consumers substitute between the two. One challenge is that with only aggregate data and market fixed effects, the distribution of this random coefficient is not identified.[130] In our setting, micro moments identify demographic-specific preferences for all inside goods. Since we only consider two demographics, however, incorporating second choice data is important to more credibly estimate the degree of unobserved preference

---

[130]Recall the FRAC regression from Section 3. The artificial regressor on a constant characteristic $x_{jt} = 1$ is $a_{jt} = \mathcal{S}_{0t} - 1/2$, the outside share minus one-half, variation of which is absorbed by market t fixed effects. Moments of the demographic distribution are also collinear with market fixed effects.

heterogeneity for all inside goods. We discuss how we collect second choices below.

Micro Data

The NielsenIQ Consumer Panel dataset tracks the purchase decisions of households over time. The dataset is split into sub-datasets, one for each panel year. We keep all panel years from 2007–2016, and restrict to trips to stores in our product data made by households who live in Seattle. We only use data on inside purchases of products in our product data.

To account for non-random participation of households in the NielsenIQ panel, we construct sampling weights for each household-year. We use these weights whenever computing statistics from the NielsenIQ data, including above when computing average revenue per trip. For each demographic bin and year, the sampling weight is equal to the share of Seattle households in this bin from the ACS, divided by the same share from NielsenIQ. Reweighting households helps adjust for non-random selection into the NielsenIQ dataset.[131]

Our micro moments match means and covariances. When computing micro moment sample values $f_m(\bar{v})$, we compute weighted averages and weighted covariances, with weights equal to the household's sampling weight, multiplied by total ounces purchased, and divided by the quarter's market size.

By weighting in this way, we are assuming that NielsenIQ Consumer Panel purchase data are generated as follows. First, a household is selected to be in the dataset based on its households weights. Second, each purchased ounce of inside beverages is recorded with probability proportional to the market size of that quarter, which in turn is proportional to the number of trips made that quarter by Seattle households.

Each purchase n made by a household is associated with a quarter t_n, a product j_n, and a group of agent types i_n with the same observed demographics $y_{i_n t_n}$ as the household. Like in the ACS data, we define a high income household as one with deflated pre-tax household income above \$67,106. We directly use the NielsenIQ variable that measures the presence of at least one child in the household.

We observe multiple choices per household-quarter, but our model assumes that each consumer makes a discrete choice per quarter. To bridge this disconnect, we assume that each "consumer" in the model is a household-ounce choice. When conducting statistical inference, we set the number of micro observations in the dataset d equal to $N_d = 10{,}455$, the number of trips with a purchase of an inside good $j \neq 0$.

If there were only observed demographics y_{it}, this type of assumption would be fairly

---

[131]NielsenIQ also provides projection weights, but our understanding is that these are for computing nationally representative statistics, not sub-nationally representative statistics.

innocuous because we always condition on these observables when computing micro moment scores. But with unobserved preferences ν_{it}, this assumption of many independent purchases for each household translates to re-drawing ν_{it} for each ounce the household purchases.

To use the panel aspect of the data, we would have to extend our likelihood to be the product of purchase probabilities over many time periods (e.g., Chintagunta and Dubé, 2005), which is beyond the scope of this paper that considers only the static BLP model. Incorporating multiple time periods would be similar to incorporating second or third choices, although with enough time periods estimation may run into numerical errors, since there would be a product of many small probabilities.

Second Choice Survey

To estimate the degree of preference heterogeneity for the outside good and for diet beverages, we use second choice data and match the probability consumers divert from some of the most popular non-diet brands to the outside good or a diet beverage, respectively, if their first choice brand were to be eliminated from their choice set. We could have matched a number of other similar second choice statistics, but we view these as particularly interpretable and tightly-connected to the counterfactual of interest, given our focus on substitution from non-diet beverages to the outside good and diet beverages.

We demonstrate how one can collect second choice data by building the short online survey in Figure J1 with Qualtrics and recruiting participants from Prolific Academic.[132] After providing informed consent, participants were asked questions about their first choice brand of non-diet soft drink, how much of this brand they purchased during the last 30 days, and what they would do if it were no longer available. We also collect basic demographic information.

Our survey design is similar to that used by choice-based conjoint analysis, and we try to follow best practices outlined in Allenby, Hardt, and Rossi (2019). See Stantcheva (2023), for example, for a more general introduction to collecting data with surveys. To elicit first choices, we first ask participants which of eight popular non-diet brands of soda they have purchased the most of during the last 30 days.[133] We discard the responses of about 30% of participants who say they have purchased none of these brands.[134] Following standard practice, we provide brand images to help focus participant attention (Wedel and Pieters,

---

[132]The Harvard University IRB determined the survey to be exempt from full IRB review.

[133]We select these brands from the top-purchased brands in Seattle near the end of our sample.

[134]Adding more brands would decrease this share of discarded responses at the cost of having a more complicated survey.

2000) and randomly order brands.

Since our demand system is for ounces, we next use a volumetric task similar those in Howell, Lee, and Allenby (2016). Instead of directly asking for volume in ounces, which participants are unlikely to be able to estimate very well, we ask participants to record how many of each common first choice beverage size their household purchased during the last 30 days.[135] Again, we provide first choice brand-specific images of bottles and cans to help focus the attention of participants. Responses seem to be compatible with the aggregate data. For example, 12-packs of 12-ounce cans (i.e., multiples of 12 entered next to 12-ounce cans in the survey) seem to account for the most Coke volume both in our survey and in the NielsenIQ data.

For second choices, we ask respondents what they would have done had their first choice brand been unavailable. In addition to the diet version of their first choice brand and both non-diet and diet versions of the other seven popular brands that we presented for first choice options, we also have options for another non-diet soft drink, another diet soft drink, or the outside option: no drink or a non-soft drink.[136] Again, we expect that explicit choices along with brand logos help focus participants' attention. Ideally, having "other" options means that which brands we put on the page should not matter for measuring substitution from non-diet to diet or the outside good. However, concern about participants treating these "other" options differently depending on which brands were shown (and in particular, the presence of the diet version of the first choice brand) led us to randomly display only half of the eight brands to each respondent. We discuss results from this randomization below.

Finally, we ask participants demographic questions to mirror ACS and NielsenIQ data: whether their pre-tax household income was above the previously-mentioned median of $67,106 in 2023 dollars, and whether their household has at least one child. As we will discuss shortly, we collect participants from all of Washington State, so we also ask whether participants live in Seattle to check whether the responses of Seattle residents are particularly different.

We recruited 139 participants from Prolific Academic to achieve a target sample size of 100 participants who selected a brand in the first choice question. We chose Prolific over other online survey platforms such as Amazon's Mechanical Turk because other researchers

---

[135]We obtain common beverage sizes from the aggregate NielsenIQ scanner data. We use the same sizes for all first choice brands (16.9-ounce bottles, 1-liter bottles, 3-liter bottles, 12-ounce cans, and 7.5-ounce cans) except Gatorade and Powerade, which have different common sizes (32-ounce, 20-ounce, and 12-ounce bottles).

[136]We describe what we mean by each of these categories in the question text. See Page 5 in Figure J1. These definitions mirror the definitions that we use when defining our sample of aggregate data.

have found that Prolific response quality tends to be the best among online platforms with large participant pools (Peer et al., 2017; Eyal et al., 2021). To take part in the study, we use Prolific filters for residents of Washington State,[137] and for those using non-mobile devices because our choice exercises take up a large amount of screen space. We conservatively paid each participant $1 for taking the survey, which for a median time spent on the survey of only one and a half minutes ended up being far above the Washington minimum wage of $15.74 per hour. If we were instead to pay this minimum wage, running the survey would have cost around $60.

Like our micro moments constructed from NielsenIQ micro data, we adjust for both non-random sampling by demographic group and by ounces purchased. Weights are the share of Seattle households in the demographic bin divided by the same share in our survey responses, multiplied by ounces in the volume task. We compute two weighted averages: the share of respondents who would substitute to a diet soft drink, and the share who would substitute to the outside good.

In Table J1 we report these weighted averages both for all survey responses, which is what we match in Table 9, and also for a number of subsamples of responses. Across subsamples, both shares are between 0.1 and 0.3, suggesting that, as our demand estimates reflect, there is substantial unobserved preference heterogeneity for both diet soft drinks and the outside good.

To check for dataset compatibility and standard survey biases, we also re-compute these numbers for different subsamples. The one-fourth of responses who live in Seattle tend to have slightly higher diversion to the outside good, but similar diversion to diet soft drinks. As discussed above, we can also check whether the diet version of the first choice brand being visible in the second choice question affects results. It does seem to somewhat increase diversion to diet soft drinks and decrease diversion to the outside good, but not by much. Finally, we verify that our results are unlikely to be driven by low-quality, fast-clicking respondents by re-computing our statistics for those who take little time on the second choice question, and for those who have taken relatively few surveys on Prolific. Shares are very similar to those for the full sample.

---

[137]We could have also set up a pre-screening survey for Seattle residents, although at this point our pool of potential respondents would likely have been much less than 200, much less than the recently-active (as of March 2023) Prolific participants in Washington.

Figure J1: Second Choice Survey

What is your Prolific ID?

Please note that this response should auto-fill with the correct ID.

During the last 30 days, which of these *non-diet* brands of soft drinks have you purchased the *most* of?

Coke — Powerade — Mountain Dew — Pepsi — Canada Dry — Seven Up — Gatorade — Dr Pepper — Have purchased none of these non-diet brands during the last 30 days

Page 1: Standard consent form.	Page 2: Prolific ID collection. Prolific IDs can be used to merge in demographics and past survey experience provided by Prolific.	Page 3: First choice brand question. Brands are randomly ordered when shown to participants.

Continued on the next page.

If non-diet Coke was *not available*, what type of drink would you have purchased instead?

"Diet" means advertised as diet, zero sugar, light, etc.

"Non-soft drink" means juice, milk, plain water, an alcoholic beverage, etc.

"Soft drink" means soda, sports drink, fruit-flavored drink, flavored water, energy drink, etc.

Non-diet Pepsi	Diet Pepsi
○	○
Non-diet Gatorade	Gatorade Zero
○	○
Non-diet Powerade	Powerade Zero
○	○
Non-diet Canada Dry	Canada Dry Zero
○	○
Non-diet Dr Pepper	Diet Dr Pepper
○	○
Non-diet Mountain Dew	Diet Mountain Dew
○	○
Non-diet Seven Up	Seven Up Zero
○	○
○ No drink or non-soft drink	Diet Coke ○
○ Other non-diet soft drink	○ Other diet soft drink

During the last 30 days, how many non-diet Coke drinks did your household purchase?

If you purchased a package, add the number of drinks in the package. If a size is missing, use the closest.

[ ] 16.9-ounce bottles   [ ] 12-ounce cans

[ ] 1-liter bottles   [ ] 7.5-ounce cans

[ ] 2-liter bottles

Last year, was the total pre-tax income of your household above $85,000?

○ Yes, above $85,000
○ No, below $85,000

Does your household have at least one child below the age of 18?

○ Yes, household has at least one child
○ No, household does not have any children

Is your household in Seattle?

○ Yes, in Seattle
○ No, not in Seattle

Page 4: Volume question for respondents who selected Coke on Page 3. Images depend on the response to Page 3. All volumes are the same except Gatorade and Powerade, which display standard 32, 20, and 12 ounce sizes.

Page 5: Second choice question for respondents who selected Coke on Page 3. Only half of the eight brands are randomly visible when shown to participants. The three text options at the bottom are always visible.

Page 6: Demographic questions. Choices are randomly ordered when shown to participants.

Table J1: Second Choice Survey Subsamples

	Weighted Average Diversion	
	Diet	Outside
All Survey Responses ($N = 100$)	0.17	0.16
	(0.04)	(0.04)
Respondent Household Location:		
↪ Seattle ($N = 25$)	0.15	0.30
	(0.07)	(0.09)
↪ Elsewhere in Washington ($N = 75$)	0.17	0.10
	(0.04)	(0.03)
Diet Version of First Choice Brand:		
↪ Visible ($N = 57$)	0.20	0.10
	(0.05)	(0.04)
↪ Not Visible ($N = 43$)	0.12	0.24
	(0.05)	(0.07)
Time Spent on Second Choice Question:		
↪ At Least 16 Seconds ($N = 50$)	0.15	0.14
	(0.05)	(0.05)
↪ Less than 16 Seconds ($N = 50$)	0.18	0.18
	(0.05)	(0.05)
Experience with Taking Other Surveys:		
↪ At Least 1058 Prolific Approvals ($N = 50$)	0.19	0.18
	(0.06)	(0.05)
↪ Less Than 1058 Prolific Approvals ($N = 50$)	0.15	0.14
	(0.05)	(0.05)

This table reports the diversion ratios we match computed on the full sample of survey responses and for different subsamples. Standard errors are in parentheses.