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 $m_t^\alpha = -\exp[\beta + \pi m_t^y + (\sigma^2 + \pi^2 v_t^2)/2]$ 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

		MAE (%)			Bias (%)		
Micro Moments Shorthand	Optimal	$\hat{\pi}_1$	$\hat{\pi}_p$	$\hat{\sigma}_p$	$\hat{\pi}_1$	$\hat{\pi}_p$	$\hat{\sigma}_p$
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

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

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

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

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

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

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

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

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 $s_{ijkt} = s_{ik(\text{-}j)t} - s_{ikt}$, the score for ρ on a micro dataset with second choices is $s_{ik(\text{-}j)t}/s_{ijkt}$ 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

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

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

			MAE (%)			Bias (%)		
Micro Moments Shorthand	$\mathcal{J}_t = \mathcal{J}$	Optimal	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\rho}$	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\rho}$
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

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

in which the within-market mean of demographic r is $m^y_{rt} = \sum_{i \in \mathcal{I}_t} w_{it} \cdot y_{rit}$, the within-market covariance between demographics r and r' is $v^y_{rr't} = \sum_{i \in \mathcal{I}_t} w_{it} \cdot (y_{rit} - m^y_{rt}) \cdot (y_{r'it} - m^y_{r't})$, 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^y_{rt}, while "controlling" for each of the $C^2 \times R^2$ covariates $v^y_{rr't} \cdot a_{cc'jt}$.

D. Asymptotics

The micro BLP model has a number of quantities that could be interpreted as "sample sizes." The total number of aggregate observations $N_{\text{A}} = \sum_{t \in \mathcal{T}} |\mathcal{J}_t|$ 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 $N_{\text{A}} \to \infty$ and $N_d \to \infty$ with $N_{\text{A}}/N_d \to \lambda_d^c$. 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) $N_{\text{A}}^{1/2}\hat{g}(\theta_0) \xrightarrow{D} N(0, S)$; (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 $N_{\text{A}}^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, (G'WG)^{-1}G'WSWG(G'WG)^{-1})$. We scale by $N_{\text{A}}^{1/2}$ 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 $N_{\text{A}}^{1/2} g(\theta_0) \xrightarrow{D} N(0, S)$. Let $\hat{g}_{\text{P}}(\theta_0) = \bar{v} - v(\theta_0)$ be micro part sample moments with terms that when passed through micro moment functions $f = (f_1, \ldots, f_{M_{\text{M}}})'$ deliver micro sample "moments" $\hat{g}_{\text{M}}(\theta_0) = f(\bar{v}) - f(v(\theta_0))$. It will suffice to show that $N_{\text{A}}^{1/2}(\hat{g}_{\text{A}}(\theta_0)', \hat{g}_{\text{P}}(\theta_0)')' \xrightarrow{D} N(0, \Omega)$ since the multivariate delta method then delivers (iii) with $S = \text{diag}(I_{M_{\text{A}}}, F)\, \Omega\, \text{diag}(I_{M_{\text{A}}}, F)'$ 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 $N_{\text{A}} = T \cdot \overline{J}$ where the average number of products per market is $\overline{J} = \frac{1}{T} \sum_{t \in \mathcal{T}} J_t$. For each case, we rewrite

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

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

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

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 $(\xi_{jt} \cdot z_{jt}, v_{1i_nj_nt_n} - v_1(\theta_0), \ldots, v_{P_\mathrm{M}i_nj_nt_n} - v_{P_\mathrm{M}}(\theta_0))'$. 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:

$$\Omega^0_{0p} = \mathbb{E}[\xi_{jt} \cdot z_{jt} \cdot \mathbb{E}_\mathrm{A}[v_{pi_nj_nt_n} - v_p(\theta_0)]] = 0. \tag{D3}$$

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

$$\Omega^0_{pq} = \mathbb{E}[(v_{pi_nj_nt_n} - v_p(\theta_0)) \cdot \mathbb{E}_\mathrm{A}[v_{qi_nj_nt_n} - v_q(\theta_0) \mid \mathcal{N}_{d_p}]] = 0. \tag{D4}$$

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

$$\Omega^0_{pq} = \mathbb{E}[\mathbb{C}_\mathrm{A}(v_{pi_nj_nt_n}, v_{qi_nj_nt_n})], \tag{D5}$$

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

$$\begin{aligned}&\mathbb{C}_\mathrm{A}(v_{pi_nj_nt_n}, v_{qi_nj_nt_n}) \\ &\quad = \frac{\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{J}_t\cup\{0\}} w_{it} \cdot s_{ijt}(\theta_0) \cdot w_{dijt} \cdot (v_{pijt} - v_p(\theta_0)) \cdot (v_{qijt} - v_q(\theta_0))}{\sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}_t}\sum_{j\in\mathcal{J}_t\cup\{0\}} w_{it} \cdot s_{ijt}(\theta_0) \cdot w_{dijt}}.\end{aligned} \tag{D6}$$

Case (a) with fixed J_t and $T/T_d \to \lambda^a_d$ 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

$$\begin{aligned}\Omega^a_{00} &= \mathbb{E}[J_t] \cdot \mathbb{V}\left(\frac{1}{J_t}\sum_{j\in\mathcal{J}_t} \xi_{jt} \cdot z_{jt}\right) \\ &= \mathbb{E}[J_t] \cdot \mathbb{E}\left[\mathbb{V}\left(\frac{1}{J_t}\sum_{j\in\mathcal{J}_t} \xi_{jt} \cdot z_{jt} \,\middle|\, J_t\right)\right] + 0 \\ &= \mathbb{E}[J_t] \cdot \mathbb{E}[1/J_t] \cdot \Omega^0_{00}.\end{aligned} \tag{D7}$$

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 $\Omega_{pq} = \mathbb{E}[\mathbb{C}_{\mathrm{A}}(v_{pi_nj_nt_n}, v_{qi_nj_nt_n} \mid n \in \mathcal{N}_{dt})]$, 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 $d_p = d_q = d$,

$$\begin{aligned}\Omega^a_{pq} &= \mathbb{E}[J_t] \cdot \lambda^a_d \cdot \mathbb{C}\left(\frac{1}{N_{dt}}\sum_{n\in\mathcal{N}_{dt}} (v_{pi_nj_nt_n} - v_p(\theta_0)), \frac{1}{N_{dt}}\sum_{n\in\mathcal{N}_{dt}} (v_{qi_nj_nt_n} - v_q(\theta_0))\right) \\ &= \mathbb{E}[J_t] \cdot \lambda^a_d \cdot \mathbb{E}\left[\mathbb{C}_{\mathrm{A}}\left(\frac{1}{N_{dt}}\sum_{n\in\mathcal{N}_{dt}} (v_{pi_nj_nt_n} - v_p(\theta_0)), \frac{1}{N_{dt}}\sum_{n\in\mathcal{N}_{dt}} (v_{qi_nj_nt_n} - v_q(\theta_0)) \,\middle|\, N_{dt}\right)\right] + 0 \\ &= \mathbb{E}[J_t] \cdot \lambda^a_d \cdot \mathbb{E}[1/N_{dt}] \cdot \Omega^0_{pq}.\end{aligned} \tag{D8}$$

Sub-case (a1) with $T/T_d \to \lambda^a_d \in [1, \infty)$ 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 $N_{\mathrm{A}}^{1/2}$, $\max_{d\in\mathcal{D}} T_d^{1/2}$, and $\max_{d\in\mathcal{D}} N_d^{1/2}$.

Sub-case (a2) with $T/T_d \to \lambda^a_d = \infty$ 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, $\Omega^a_{00} = \mathbb{E}[J_t] \cdot \mathbb{V}(\frac{1}{J_t}\sum_{j\in\mathcal{J}_t} \xi_{jt} \cdot z_{jt})$, 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 $\lambda^a_d = \infty$.

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 $\Omega_{pq} = \mathbb{C}_{\text{A}}(v_{pi_nj_nt_n}, v_{qi_nj_nt_n})$ 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 $N_{\text{A}}/N_d \to \lambda_d^c$ 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 $\Omega_{pq}^0 = \mathbb{C}_{\text{A}}(v_{pi_nj_nt_n}, v_{qi_nj_nt_n})$ 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 $N_{\text{A}} \to \infty$, the limit of $\Omega_{pq} = \mathbb{C}_{\text{A}}(v_{pi_nj_nt_n}, v_{qi_nj_nt_n})$ 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 $N_d \cdot \mathbb{E}[\|[v_{pi_nj_nt_n} - v_p(\theta_0)]_{d_p=d}/N_d\|^{2+\delta_d}] \to 0$ as $N_{\text{A}} \to \infty$. 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 $N_{\text{A}}/N_d \to \lambda_d^c \in (0, \infty)$ 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 $N_{\text{A}}/N_d \to \lambda_d^c = \infty$ 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 $N_{\text{A}}/N_d \to \lambda_d^c = 0$ 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,

$$\frac{\partial \hat{g}_{\text{M}}(\theta)}{\partial \theta'} = \begin{bmatrix} \frac{\partial f_1(v(\theta))}{\partial v'}\frac{\partial v(\theta)}{\partial \theta'} \\ \vdots \\ \frac{\partial f_{M_{\text{M}}}(v(\theta))}{\partial v'}\frac{\partial v(\theta)}{\partial \theta'} \end{bmatrix}, \tag{D13}$$

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

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

The micro block is

$$\hat{S}_\text{M} = \hat{F} \hat{S}_\text{P} \hat{F}', \quad \hat{F} = \frac{\partial f(v(\hat{\theta}))}{\partial v'}, \tag{D15}$$

in which element (p, q) of $\hat{S}_\text{P}$ is zero if $d_p \neq d_q$, and otherwise equal to $\frac{N_\text{A}}{N_d} \cdot \hat{\Omega}^0_{pq}$ 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), $\text{diag}(\hat{S}_\text{A}, \hat{S}_\text{P})$ converges in probability to Ω^a, Ω^b, and Ω^c.

[106] For clusters $\ell = 1, \ldots, L$ of products $\mathcal{J}_{\ell t} \subset \mathcal{J}_t$, we compute $\hat{S}_\text{A} = \frac{1}{N_\text{A}} \sum_{\ell=1}^{L} \hat{g}_\ell \hat{g}_\ell'$ 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 $\mathbb{E}[g_{\text{A}}(j, t; \theta_0)] = \mathbb{E}[\xi_{jt}(\theta_0) \cdot b(z_{jt})] = 0$ based on the aggregate data.

Similarly, given conditional likelihoods $\mathbb{P}_{\text{A}}(t_n,\, j_n,\, y_{i_n t_n} \mid n \in \mathcal{N}_d)$, any function v of the aggregate data and $(t_n, j_n, y_{i_n t_n})$ yields valid moments $\mathbb{E}[g_{\text{P}}(n;\theta_0)] = \mathbb{E}[v(n;\theta_0) - v(\theta_0)] = 0$ where $v(\theta) = \mathbb{E}^{\theta}_{\text{A}}[v(n;\theta_0)]$. 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 $g_{\text{P}} = v$. 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

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

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

$$\begin{aligned} V_\tau &= D_\tau^{-1}\mathbb{E}[Y_\tau(j,t,n)Y_\tau(j,t,n)']D_\tau^{-1\prime}, \\ D_\tau = G'WG, \quad Y_\tau(j,t,n) &= G'W\begin{bmatrix} g_{\text{A}}(j,t;\theta_0) \\ F \cdot g_{\text{P}}(n;\theta_0) \end{bmatrix}. \end{aligned} \tag{E2}$$

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

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

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

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

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

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

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

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

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,

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

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 $v_{mi_n j_n t_n}$ 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 $v_{mi_n j_n t_n}$ 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 $v_{pi_n j_n t_n}$, averages of which could reasonably be collected by a survey administrator. For example, we could restrict ourselves to second-order polynomial interactions of demographics $y_{ri_n t_n}$ and merged-in aggregate data, such as product characteristics $x_{cj_n t_n}$ and market shares $\mathcal{S}_{j_n t_n}$. We can exclude trivial candidates such as a constant, $v_{pi_n j_n t_n} = 1$, 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 $m = 1, \ldots, M_{\text{M}}$ 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, $v_{pi_n j_n t_n} = y_{ri_n t_n}$, gives an expected micro value equal to the mean of that demographic, $v_m(\theta) = \sum_{i \in \mathcal{I}_{t_n}} w_{it_n} \cdot y_{rit}$, 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 $v(\theta) = \mathbb{E}^{\theta}_{\mathrm{A}}[v(n)]$ and $v(n) = (v_{1 i_n j_n t_n}, \ldots, v_{M_{\mathrm{M}} i_n j_n t_n})'$. 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

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

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 $M_{\text{M}} = 1$ micro moment with $\mathbb{E}[v(n)\mathbb{S}_{\text{A}}(n \mid d)] = \mathbb{C}(v(n), \mathbb{S}_{\text{A}}(n \mid d))$. Since we normalized $\mathbb{V}(v(n)) = 1$ and $\mathbb{V}(\mathbb{S}_{\text{A}}(n \mid d))$ does not depend on $v(\cdot)$, we can rewrite (G3) as

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

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 $Y = (\mathbb{S}_{\text{A}}(1 \mid d), \ldots, \mathbb{S}_{\text{A}}(N_d \mid d))$ be a $N_d \times 1$ vector and let $X = (v(1), \ldots, v(N_d))'$ be a $N_d \times M_{\text{M}}$ 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

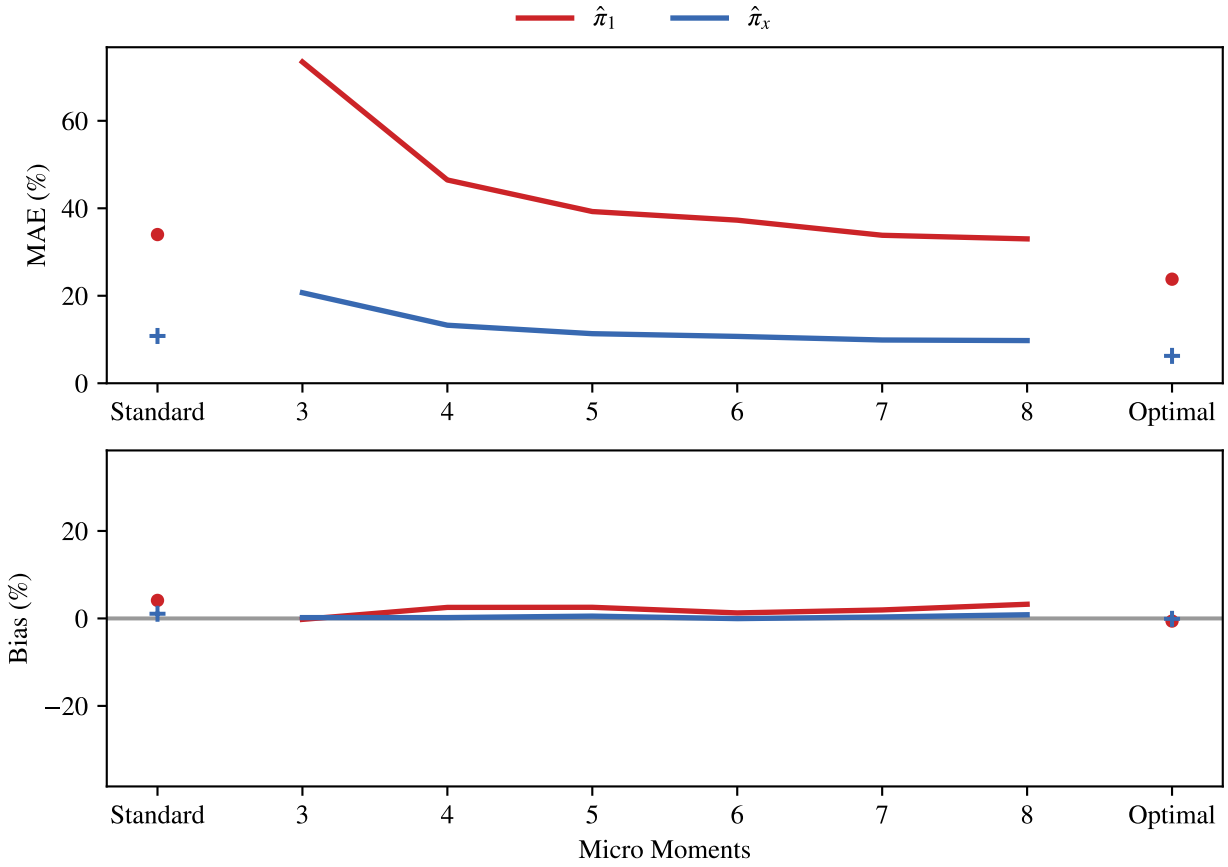
```
# Simulate micro data
simulated_data = problem_results.simulate_micro_data(micro_dataset, seed=0)

# 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.

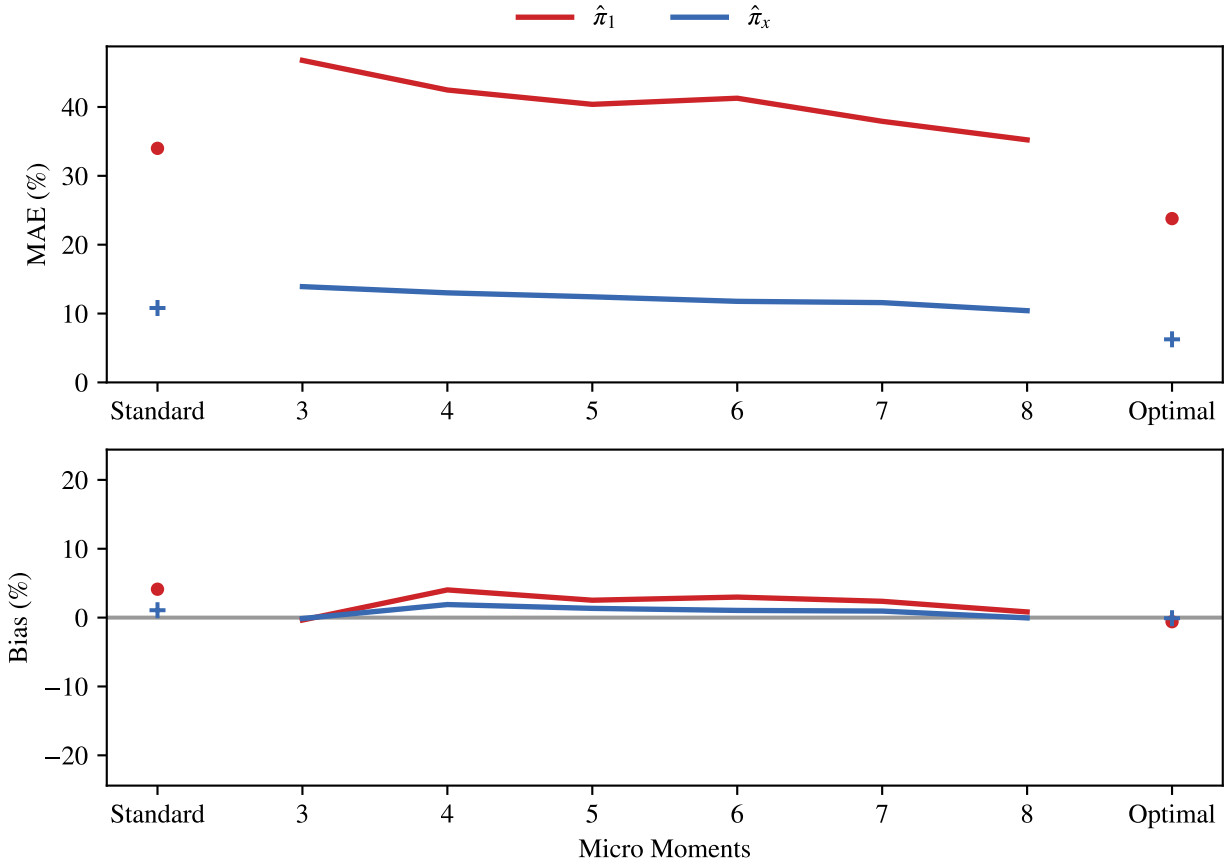
[108] We interact $y_{i_n t_n}$, $1\{y_{i_n t_n} < \bar{y}_{t_n}\}$, $1\{y_{i_n t_n} \geq \bar{y}_{t_n}\}$, $1\{j_n \neq 0\}$, $x_{2j_n t_n}$, $1\{x_{2j_n t_n} < \bar{x}_{2t_n}\}$, $1\{x_{2j_n t_n} \geq \bar{x}_{2t_n}\}$, and $\mathcal{S}_{j_n t_n}$ where $\bar{y}_t$ and $\bar{x}_{2t}$ are medians of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

Figure G2: R^2 Approach



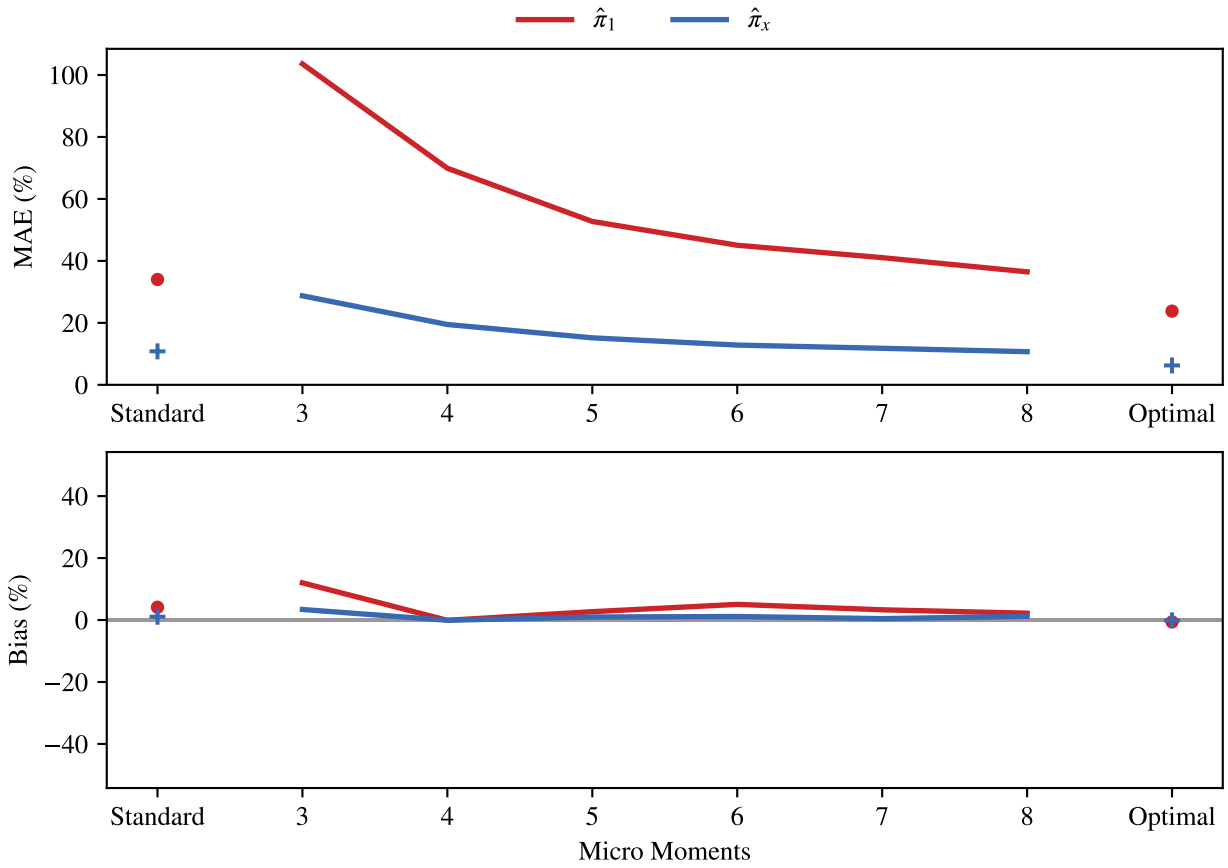
This figure reports median absolute error (MAE) and median bias of parameter estimates over 1,000 simulated datasets for approximations to the optimal micro moments. Values above "Standard" correspond to the same "$\mathbb{E}[y_{it} \mid j \neq 0]$" and "$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$" moments 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



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 $\hat{\theta}$:

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

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
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{E}[x_{2jt} \mid y_{it} < \overline{y}_t, j \neq 0]$"	95.0	94.8	-1.6	-2.3
"$\mathbb{E}[y_{it} \mid j \neq 0]$, $\mathbb{E}[x_{2jt} \mid y_{it} < \overline{y}_t, j \neq 0]$, $\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"	84.9	81.7	-28.2	-33.1

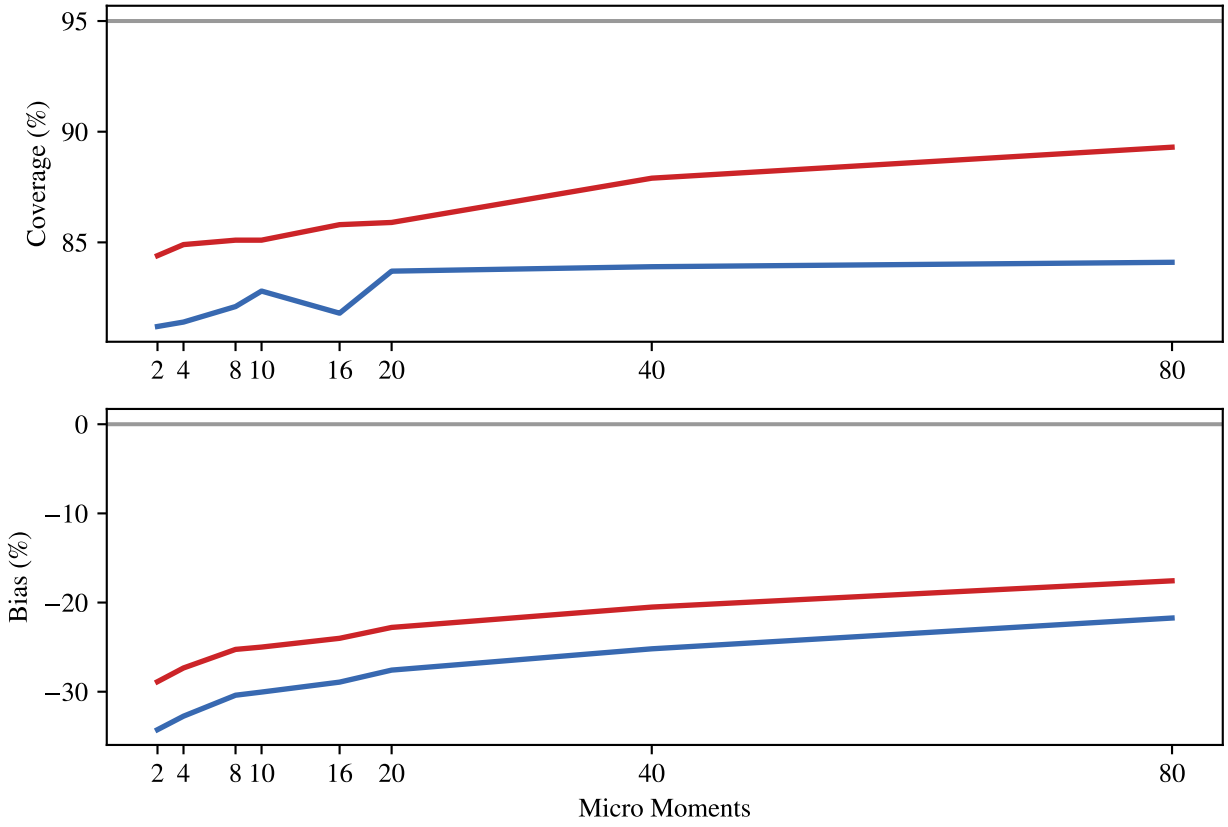
This table reports standard error coverage and median bias for Table 5.

Table H3: Standard Errors, Optimal Micro Moments and Compatibility

			Coverage (%)		Bias (%)	
Micro Moments (plus $\mathbb{E}[y_{it} \mid j \neq 0]$)	Incompatible	Optimal	$\hat{\pi}_1$	$\hat{\pi}_x$	$\hat{\pi}_1$	$\hat{\pi}_x$
"$\mathbb{C}(x_{2jt}, y_{it} \mid j \neq 0)$"			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



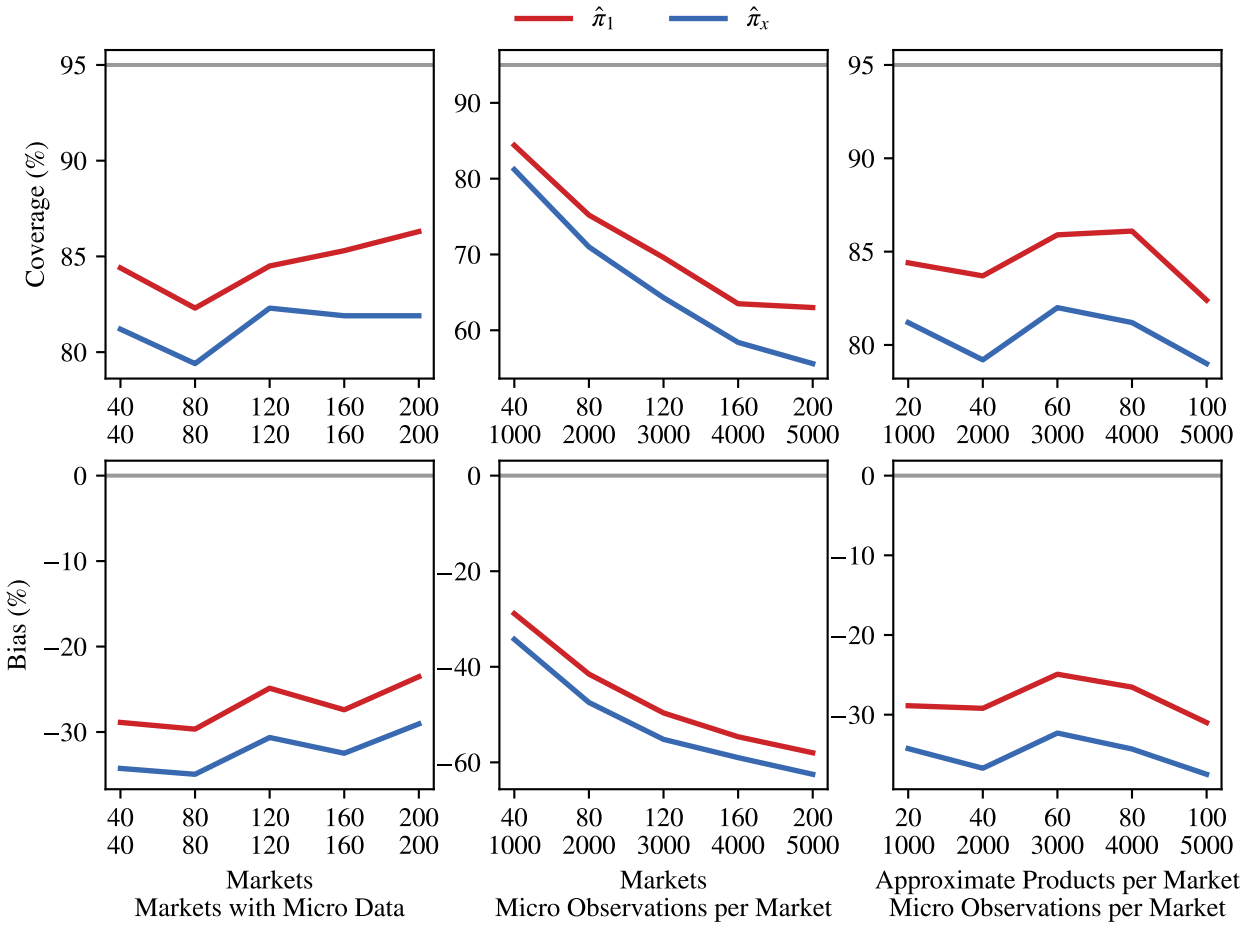
This figure reports standard error coverage and median bias for Figure 2.

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 (%) $\hat{\pi}_1$	Coverage (%) $\hat{\pi}_x$	Coverage (%) $\hat{\sigma}_x$	Bias (%) $\hat{\pi}_1$	Bias (%) $\hat{\pi}_x$	Bias (%) $\hat{\sigma}_x$
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 (%) $\hat{\pi}_1$	Coverage (%) $\hat{\pi}_x$	Coverage (%) $\hat{\sigma}_x$	Bias (%) $\hat{\pi}_1$	Bias (%) $\hat{\pi}_x$	Bias (%) $\hat{\sigma}_x$
No Second Choice Moments		96.1	89.1	98.5	-4.7	-19.7	160.7
"$\mathbb{C}(x_{2jt}, x_{2k(\text{-}j)t} \mid j, k \neq 0)$"		91.0	90.0	94.0	-23.2	-22.1	-23.6
"$\mathbb{E}[x_{2jt} + x_{2k(\text{-}j)t} \mid j, k \neq 0]$"		90.2	86.5	93.3	-25.4	-32.7	-55.3
"$\mathbb{P}(x_{2k(\text{-}j)t} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t},\ j, k \neq 0)$"		91.2	89.9	92.2	-25.9	-34.5	-49.6
"$\mathbb{P}(x_{2k(\text{-}j)t} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t},\ j, k \neq 0)$"	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 (%) $\hat{\pi}_1$	Coverage (%) $\hat{\pi}_p$	Coverage (%) $\hat{\sigma}_p$	Bias (%) $\hat{\pi}_1$	Bias (%) $\hat{\pi}_p$	Bias (%) $\hat{\sigma}_p$
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 (%) $\hat{\pi}_1$	Coverage (%) $\hat{\pi}_x$	Coverage (%) $\hat{\rho}$	Bias (%) $\hat{\pi}_1$	Bias (%) $\hat{\pi}_x$	Bias (%) $\hat{\rho}$
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 "$\mathbb{P}(x_{2kt} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t},\ j, k \neq 0)$"	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

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

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

	MAE (%)		Bias (%)	
Micro Moments Shorthand	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

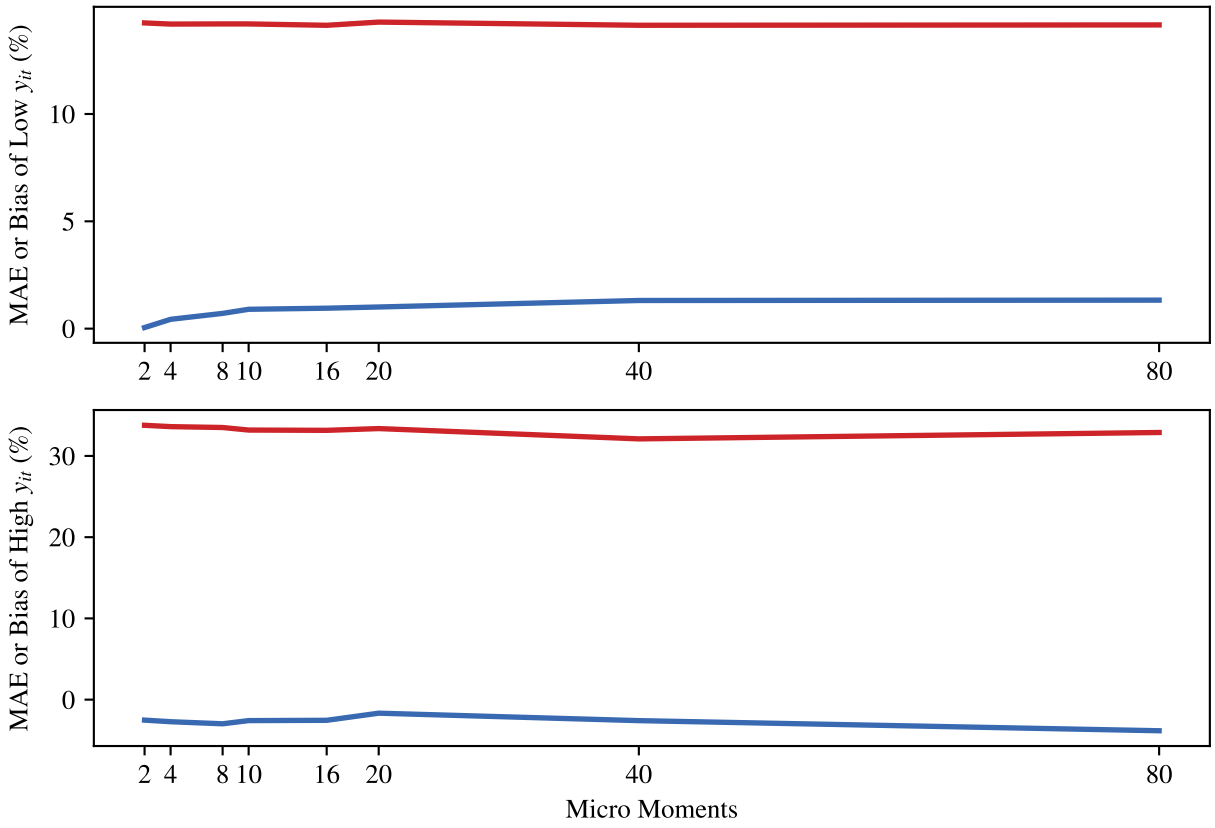
This table reports median absolute error (MAE) and median bias of each *CS* for Table 5.

Table I3: Counterfactual, Optimal Micro Moments and Compatibility

			MAE (%)		Bias (%)	
Micro Moments (plus $\mathbb{E}[y_{it} \mid j \neq 0]$)	Incompatible	Optimal	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



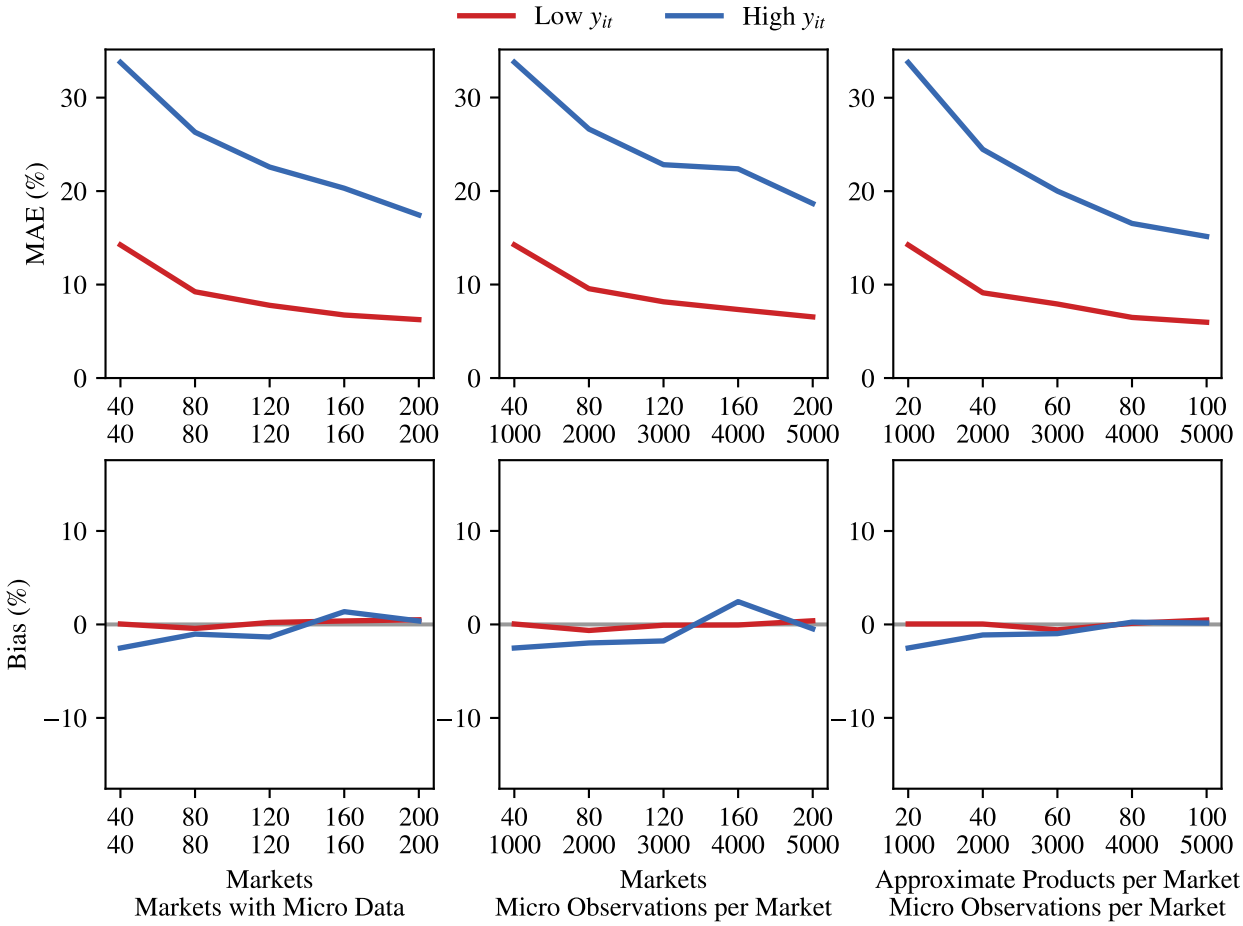
This figure reports median absolute error (MAE) and median bias of each *CS* for Figure 2.

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
"$\mathbb{C}(x_{2jt}, x_{2k(\text{-}j)t} \mid j, k \neq 0)$"		6.5	6.0	0.1	0.2
"$\mathbb{E}[x_{2jt} + x_{2k(\text{-}j)t} \mid j, k \neq 0]$"		2.8	2.6	-0.0	0.0
"$\mathbb{P}(x_{2k(\text{-}j)t} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t},\ j, k \neq 0)$"		5.4	5.1	-0.4	-0.4
"$\mathbb{P}(x_{2k(\text{-}j)t} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t},\ j, k \neq 0)$"	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 "$\mathbb{P}(x_{2kt} < \bar{x}_{2t} \mid x_{2jt} \gtrless \bar{x}_{2t},\ j, k \neq 0)$"	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.

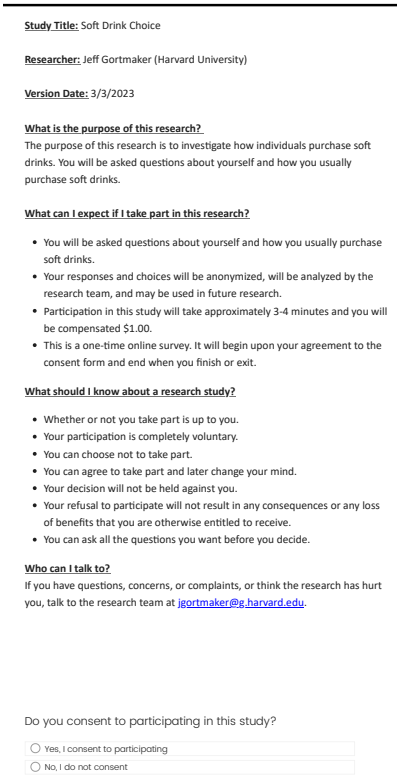
Like our micro moments constructed from NielsenIQ micro data, we adjust for both non-random sampling by demographic group and by ounces purchased. Weights are the share of Seattle households in the demographic bin divided by the same share in our survey responses, multiplied by ounces in the volume task. 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



Study Title: Soft Drink Choice

Researcher: Jeff Gortmaker (Harvard University)

Version Date: 3/3/2023

What is the purpose of this research?
The purpose of this research is to investigate how individuals purchase soft drinks. You will be asked questions about yourself and how you usually purchase soft drinks.

What can I expect if I take part in this research?

- You will be asked questions about yourself and how you usually purchase soft drinks.
- Your responses and choices will be anonymized, will be analyzed by the research team, and may be used in future research.
- Participation in this study will take approximately 3-4 minutes and you will be compensated $1.00.
- This is a one-time online survey. It will begin upon your agreement to the consent form and end when you finish or exit.

What should I know about a research study?

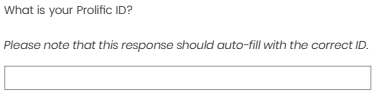
- Whether or not you take part is up to you.
- Your participation is completely voluntary.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.
- You can ask all the questions you want before you decide.

Who can I talk to?
If you have questions, concerns, or complaints, or think the research has hurt you, talk to the research team at jgortmaker@g.harvard.edu.

Do you consent to participating in this study?

- ○ Yes, I consent to participating
- ○ No, I do not consent

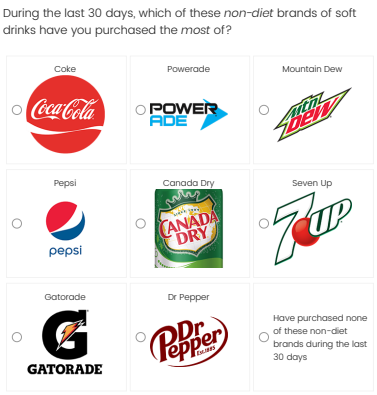
Page 1: Standard consent form.



What is your Prolific ID?

Please note that this response should auto-fill with the correct ID.

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.

Continued from the previous page.

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 

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.

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

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.

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 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:		
$\hookrightarrow$ Seattle ($N = 25$)	0.15	0.30
	(0.07)	(0.09)
$\hookrightarrow$ Elsewhere in Washington ($N = 75$)	0.17	0.10
	(0.04)	(0.03)
Diet Version of First Choice Brand:		
$\hookrightarrow$ Visible ($N = 57$)	0.20	0.10
	(0.05)	(0.04)
$\hookrightarrow$ Not Visible ($N = 43$)	0.12	0.24
	(0.05)	(0.07)
Time Spent on Second Choice Question:		
$\hookrightarrow$ At Least 16 Seconds ($N = 50$)	0.15	0.14
	(0.05)	(0.05)
$\hookrightarrow$ Less than 16 Seconds ($N = 50$)	0.18	0.18
	(0.05)	(0.05)
Experience with Taking Other Surveys:		
$\hookrightarrow$ At Least 1058 Prolific Approvals ($N = 50$)	0.19	0.18
	(0.06)	(0.05)
$\hookrightarrow$ 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.