

Romeo, Charles J.

Working Paper

Filling Out the Instrument Set in Mixed Logit Demand Systems for Aggregate Data

Economic Analysis Group Discussion Paper, No. EAG 10-3

Provided in Cooperation with:

Expert Analysis Group (EAG), Antitrust Division, United States Department of Justice

Suggested Citation: Romeo, Charles J. (2013) : Filling Out the Instrument Set in Mixed Logit Demand Systems for Aggregate Data, Economic Analysis Group Discussion Paper, No. EAG 10-3, U.S. Department of Justice, Antitrust Division, Economic Analysis Group (EAG), Washington, DC

This Version is available at:

<https://hdl.handle.net/10419/202389>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Filling Out the Instrument Set in Mixed Logit Demand Systems for

Aggregate Data^{*}

Charles J. Romeo
Consumer Financial Protection Bureau
1700 G Street, NW
Washington, DC 20552

charles.romeo@cfpb.gov

(202) 485-9448

This Version: November 2013

^{*}This article is the result of the author's independent research while in the Antitrust Division, US Department of Justice and does not necessarily represent the views of the Consumer Financial Protection Bureau or the United States. I thank Nathan Miller for many helpful discussions and some key insights. I also thank Abe Dunn, Matthew Osborne, Charles Taragin, Jeremy Verlinda, and Tor Winston. Comments by seminar participants at the Bureau of Economic Analysis, the Federal Trade Commission, and Kyoo il Kim for his comments at the International Industrial Organization Conference were greatly appreciated. Insights by all these folks greatly improved the quality of this paper. All remaining errors are mine.

Abstract

The random parameters logit model for aggregate data introduced by Berry, Levinsohn, and Pakes (1995) has been a driving force in empirical industrial organization for more than a decade. In this paper we introduce the means of included demographics as a new set of readily available instruments that have the potential to improve researchers ability to instrument for price level shifts across markets. We compare demographics instruments with market fixed effects proving that both are valid instruments in a correctly specified model, but that demographics remain valid under misspecifications that render market fixed effects invalid. A set of endogenous price simulations demonstrates these points.

Keywords: demographic instruments; instrument validity; random coefficients logit

1. Introduction

We introduce the means of included demographics as a set of instruments for use in random parameters logit demand systems. These instruments have heretofore been overlooked in the literature. We present theory and monte carlo evidence showing that these are valid instruments in a correctly specified model that includes demographic types and importantly, that they continue to be valid under misspecifications in which the econometrician has access to less demographic information than the retailer. We show that market fixed effects are also valid instruments when the model is correctly specified, but they are invalid under these types of misspecifications. The intuition for this result is that demographic information available to the retailer but unavailable to the econometrician is likely to be correlated with the fixed effects, rendering an estimator based on them biased and inconsistent.

The classes of problems in which demographics may prove to be important are ones in which firms vary prices from store-to-store, from county-to-county, or from city-to-city at least partially in response to demographic differences. Marketing and economic applications that deal with grocery products are key candidates for gains in explanatory power with the introduction of these instruments. Alternatively, products whose prices are set nationally, (e.g., online sales) will not gain from the use of these instruments, nor will products whose prices may be set locally, but for which data are only available at the national level (e.g., Berry Levinsohn and Pakes (1995) automobile data).

Mean demographics are likely to be correlated with willingness-to-pay and it is this correlation that potentially makes them important instruments in the mixed logit model. Price level shifts across markets are generally difficult to explain using instruments that are typically employed. In many product markets many of the same products are available in each geographic market leaving exogenous product characteristics unable to explain these shifts. By the same reasoning, instruments developed using the exchangeability arguments in Pakes (1996) and in Berry, Levinsohn and Pakes (1995) use polynomial expansions of the exogenous characteristics to approximate the optimal instruments are also likely to be unable to explain cross-market price level shifts. Petrin (2002, footnote 5) contains the insight that systematic relationships between the random coefficients and market shares provides important identifying power. If the

instruments do not likewise vary systematically, this identifying power is squandered. Nevo's (2001) use of average regional prices as instruments helps capture cross-market price level shifts.¹ But, as he discusses, these instruments are not valid in the logit model if there are regional demand shocks. Unexplained city-specific valuation differences that are due to demographic differences are a likely source of demand shocks. Including mean demographics in the regression model, as he does in one of his logit specifications, controls for these shocks and makes it more likely that average regional prices are valid instruments.

Petrin (2002) and BLP (2004) use national data and limit attention to an instrument set formed by a polynomial expansion of the exogenous variables, but gain additional identifying power using micro moments. BLP (1999) and Villas-Boas (2007) incorporate data on costs: wage rates (BLP (1999)) and manufacturer and retail level input price changes (Villas-Boas).² To the extent these instruments are available at the market level they could substitute for demographics or market fixed effects. Gowrisankaran and Rysman (2009) add to the exogenous product characteristics instruments meant to capture how crowded a product is in characteristic space: mean product characteristics for each firm and time period and the count of product characteristics offered by each firm and all firms. Instruments on crowding could also serve to explain cross-market price variation to the extent that there is enough cross-market variation in crowding.

In principle, additional instruments can be drawn from other moments of the demographic distribution. We limit attention to means because this is most likely the information available to retailers. Census tabulations and company reports typically limit attention to means and their changes over time.

Section 2 develops the mixed logit demand system for aggregate data. The type of model misspecifications we are considering are developed in Section 3. The theory for consistent estimation of the mean utility parameters under these forms of misspecification is in Section 4, while Section 5 contains an endogenous price monte carlo study comparing estimator consistency

¹ Nevo (2000) and Hausman, Leonard and Zona (1994) also use these instruments.

²Villas-Boas's data is for a single urban area in the Midwest. She forms input price changes from a variety of census, plant and retail chain data.

and performance of correctly and misspecified models using different instrument sets. Section 6 contains conclusions.

2. Demand Model

We represent the conditional indirect utility of consumer type i in market m at time t with preferences $(a_{im}, v_{im}, \varepsilon_{ijmt})$ from the purchase of one unit of the j th product as

$$\begin{aligned} u_{ijmt} &= [x_{jmt}, p_{jmt}] \theta_{im} + \xi_{jmt} + \varepsilon_{ijmt}, \\ \theta_{im} &= \bar{\theta} + \Gamma a_{im} + \Sigma v_{im}, \\ i &= 1, \dots, I, \quad j = 0, \dots, J_{mt}, \quad m = 1, \dots, M, \quad t = 1, \dots, T. \end{aligned} \tag{1}$$

x_{jmt} and p_{jmt} are observed product characteristics and prices respectively. The ξ_{jmt} represent product characteristics that are observed by market participants, but unobserved by the econometrician. Assuming prices are set strategically, each p_{jmt} will be correlated with all $J+1$ ξ_{jmt} in each market-time period m, t . The ε_{ijmt} are i.i.d. type 1 extreme value errors.

The second equation in (1) is a hierarchical regression of $k \times 1$ vector θ_{im} on $D \times 1$ vector of demographics a_{im} , with standard normal error vector v_{im} . Γ and Σ are unknown parameters: Γ is a $k \times D$ vector and Σ is a $k \times k$ diagonal matrix.

Normalize u_{0jmt} to zero and assume that each consumer maximizes utility by purchasing one unit of product j at time t if and only if $u_{ijmt} \geq u_{irmt}$, $r = 0, \dots, J$. Assuming that there are no ties, the market share of good j is given as

$$\begin{aligned} s_j(x_{mt}, p_{mt}, \xi_{mt}; \bar{\theta}, \lambda) &= \int \int \int dP_0(a, v, \varepsilon | m) \\ &= \int \int \int dP_0(a | m) dP_0(v | m) dP_0(\varepsilon), \end{aligned} \tag{2}$$

where P_0 is the population distribution for individual level unobservables, and where the second line results from the assumption of mutual independence of a , v , and ε .

Given that we have endowed ε with an i.i.d. type 1 extreme value distribution, the outer integral in (2) has a logit distribution as its analytical solution. Solving this integral yields an expression for market shares of the form

$$s_j(\delta_{mt}, P_{0|m}; \bar{\theta}, \lambda) = \iint \frac{\exp(\delta_{jmt} + \mu_{ijmt})}{1 + \sum_r \exp(\delta_{rmt} + \mu_{irmt})} dP_0(a|m) dP_0(v|m) \quad (3)$$

where $\delta_{jmt} = r_{jmt} \bar{\theta} + \xi_{jmt}$, $r_{jmt} = [x_{jmt}, P_{jmt}]$ and $\mu_{ijmt} = r_{jmt}(\Gamma a_{im} + \Sigma v_{im})$. The integrals in (3) generally do not have analytical solutions.

Mixed logit demand systems of the type discussed here do not generally include market fixed effects in mean utility. Nevo (2001) includes product fixed effects in lieu of product characteristics, but market fixed effects would fall out of his and many other mixed logit models in the literature. Placing them in mean utility would require one to assume that they do not enter the outside good. This is untenable as the outside good is often the unobserved portion of the market containing the inside goods. In Nevo (2001) the outside good is unobserved sales of cereal, in Villas-Boas (2007) the outside good is “yogurt sold by small retail stores, or grocery stores not considered in the analysis, as well as yogurt of small manufacturers sold in the three retail stores studied.” In Romeo (2012) the inside goods are grocery store beer sales and the outside good is beer sold through other venues.

3. Misspecification due to unobserved demographics

Without loss of generality let x_{jmt} be a scalar so that the correctly specified model has two random coefficients.³ Write μ as

$$\mu_{ijmt} = x_{jmt}(\gamma_x a_{xim} + v_x v_{xim}) + p_{jmt}(\gamma_p a_{pim} + v_p v_{pim}).$$

The misspecifications we formulate result from the retailer using more demographic information in setting prices than the econometrician has available. We form misspecifications by decomposing μ as $\mu = \mu^o + \mu^u$, where superscripts “o” and “u” identify the parts of μ that are

³Though we do not extend out notation to include an intercept. We assume mean utility contains an intercept with a fixed coefficient.

observed and unobserved by the econometrician respectively. Allow μ^o and μ^u to take three possible forms:

$$i) \mu^o = \mu, \mu^u = 0;$$

$$ii) \mu_{ijmt}^o = p_{jmt}(\gamma_p^o a_{pim}^o + v_p v_{pim}), \mu_{ijmt}^u = p_{jmt} \gamma_p^u a_{pim}^u \text{ assuming } \gamma_x = 0, v_x = 0;$$

$$iii) \mu_{ijmt}^o = p_{jmt}(\gamma_p a_{pim} + v_p v_{pim}), \mu_{ijmt}^u = x_{jmt}(\gamma_x a_{ixm} + v_x v_{xim}).$$

In (i) μ is correctly specified. In this case, it is straightforward to prove that mean demographics and market fixed effects are both valid instrumental variables. In (ii) the correctly specified model only has a random coefficient on price, but the model is misspecified because the econometrician is missing demographics a^u . In (iii) the correctly specified model has random coefficients on both x and p , but the econometrician only put a random coefficient on price. We leave open for now whether the demographic factors, a_x , in (iii) are the same as those associated with price.

We rewrite (ii) and (iii) with μ^u averaged over I as this is the form it will appear in the mean utility regression. Specifically,

$$ii) \bar{\mu}_{jmt}^o = p_{jmt}(\gamma_p^o \bar{a}_{pm}^o + v_p \bar{v}_{pm}), \bar{\mu}_{jmt}^u = p_{jmt} \gamma_p^u \bar{a}_{pm}^u \text{ assuming } \gamma_x = 0, v_x = 0;;$$

$$iii) \bar{\mu}_{jmt}^o = p_{jmt}(\gamma_p \bar{a}_{pm} + v_p \bar{v}_{pm}), \bar{\mu}_{jmt}^u = x_{jmt}(\gamma_x \bar{a}_{xm} + v_x \bar{v}_{xm}) \approx x_{jmt} \gamma_x \bar{a}_{xm}.$$

where the approximate equality in (iii) follows because $E[\bar{v}_{xm}] = 0$ for all m .

4. Consistent estimation of mean utility with unobserved demographics

We develop key elements of the estimator using BLP's Nested Fixed Point estimation algorithm. If elements of μ are unspecified the contraction mapping will be affected. Setting $(\Gamma, \Sigma) = (\Gamma^*, \Sigma^*)$ sets $\mu^o = \mu^o(\Gamma^*, \Sigma^*)$ leaving δ to explain the difference in observed and model shares not explained by $\mu^o(\Gamma^*, \Sigma^*)$. Using the decomposition of μ specified above, we write the monte carlo estimator of the integrals in (3) as

$$s_j(\delta_{mt}, \bar{\mu}_{mt}^u, \mu_{mt}^o; P_I) = I^{-1} \sum_i \frac{\exp\{(\delta_{jmt} + \bar{\mu}_{jmt}^u) + \mu_{ijmt}^o\}}{1 + \sum_k \exp\{(\delta_{kmt} + \bar{\mu}_{kmt}^u) + \mu_{ikmt}^o\}}. \quad (4)$$

The contraction mapping for (4) solves for $\tilde{\delta} = \delta + \bar{\mu}^u$. Form the instrumental variables estimator of mean utility parameters $\bar{\theta}$ as

$$\bar{\theta}_{IV} = (r'zW_Nz'r)^{-1}r'zW_Nz'\tilde{\delta},$$

where W_N is a positive definite weighting matrix with $N = JMT$, and z is a matrix of instruments that includes either the means of observed demographics, \bar{a}_m^o , or market fixed effects d_m . Define $z = [\tilde{z}, f]$, where \tilde{z} includes exogenous variables in x , and other available instruments and $f \in \{\bar{a}_m^o, d_m\}$.

Assume: (a) $\text{plim } r'z/N = Q_{rz}$ a full rank matrix, (b) $\text{plim } W_N = W$ a positive definite matrix, (c) $\text{plim } z'\tilde{\xi}/N = 0$ and (d) $\text{plim } \tilde{z}'\bar{\mu}^u/N = 0$.

For our purposes it is sufficient that the probability limit in N is reached due to $M \rightarrow \infty$.

Assumptions (a) and (b) are standard. (c) assumes that all instruments are asymptotically uncorrelated with the demand error and (d) assumes that all instruments except \bar{a}_m^o and d_m are likewise asymptotically uncorrelated with mean unobserved heterogeneity $\bar{\mu}^u$. Assumption (d) is reasonable in circumstances where the set of products available in a market does not vary systematically with its demographics. This is precisely the grocery product case: many products are available in all markets so the standard BLP instruments are unable to explain cross-market price shifts. Under these assumptions, it is straightforward to show that

$$\text{plim } \bar{\theta}_{IV} = \bar{\theta} + (Q_{rz}WQ_{rz}')^{-1}Q_{rz}W[\mathbf{0}', \text{plim } \frac{f'\bar{\mu}^u}{N}]',$$

where $\mathbf{0}$ is column vector with length equal to the columns of $\tilde{\mathbf{z}}$. Clearly, consistency of $\bar{\boldsymbol{\theta}}_{IV}$ for $\bar{\boldsymbol{\theta}}$ depends on whether instruments $\bar{\mathbf{a}}_m^o$ and d_m are valid in the face of misspecification $\bar{\boldsymbol{\mu}}^u$. In case (i), $\bar{\boldsymbol{\mu}}^u = 0$ and there is no misspecification and it follows that the estimator is consistent for $f = \bar{\mathbf{a}}_m^o$ or d_m . In cases (ii) and (iii) when $f = \bar{\mathbf{a}}_m^o$ consistency is achieved under assumptions presented below, while consistency cannot be achieved if $f = d_m$. Market fixed effects are invalid instruments under both of these types of misspecification.

Additional assumptions are needed to provide conditions under which $\bar{\mathbf{a}}_m^o$ are valid instruments in cases (ii) and (iii). Define $\bar{\mathbf{p}}_m = \sum_{j,t} p_{jmt} / JT$.

Assume: (e) $E[\bar{\mathbf{a}}^o \bar{\mathbf{a}}^u | \bar{\mathbf{p}}_m] = E[\bar{\mathbf{a}}^o \bar{\mathbf{a}}^u]$ and (f) $E[\bar{\mathbf{a}}^o | \bar{\mathbf{a}}_{pm}^u] = E[\bar{\mathbf{a}}^o] = \mathbf{0}$.

(e) implies that the demographics of an area are not affected by the local price level.⁴ This assumption implies that consumers do not move in or out of an area because of the local price level. This is a reasonable assumption for grocery products as the price of bread or beer is unlikely to influence a consumer's choice of where to live, and it may be reasonable for cars, both of which have been studied using these models, but is not likely to be appropriate for housing. The first equality in (f) implies that $\bar{\mathbf{a}}^o$ and $\bar{\mathbf{a}}^u$ are mean independent. Note that the distributions of \mathbf{a}^o and \mathbf{a}^u can be highly correlated, only the distribution of their means is required to be uncorrelated. The second equality in (f) is straightforward to enforce by standardizing the distribution of $\bar{\mathbf{a}}^o$.⁵

⁴We do, of course, expect that causality does flow in the other direction, from demographics to the price level: $E[\bar{\mathbf{p}} | \bar{\mathbf{a}}_m^o, \bar{\mathbf{a}}_m^u] \neq E[\bar{\mathbf{p}}]$. Otherwise mean demographics would be uninformative as instruments.

⁵Experience indicates that it is good practice to standardize all data in these models as it can dramatically improve optimization performance. Some older literature (e.g. Draper and Smith 1981, Chapter 5.5) note the importance of data standardization for reducing round off error. Rounding error continues to be a problem, in particular, when using estimators like the Nested Fixed Point algorithm used here. As Dubé, Fox and Su (2012) discuss, errors propagate from the inner loop of the algorithm and from the use of numerical derivatives to the outer loop. They demonstrate a link between the inner and outer loop tolerances and suggest an inner loop tolerance of 10^{-14} . Unscaled data may make it impossible to achieve an inner loop tolerance this tight.

Theorem 1: Under assumptions (a - f), in case (ii) with $\bar{\mu}_{jmt}^u = p_{jmt} \gamma_p^u \bar{a}_{pm}^u$ and $f = \bar{a}_{pm}^o$,

$$plim \frac{\bar{a}_{pm}^o \bar{\mu}_m^u}{N} = \gamma_p^u plim \frac{\bar{a}_{pm}^o p_{jmt} \bar{a}_{pm}^u}{N} = \gamma_p^u plim \frac{\bar{a}_{pm}^o \bar{p}_m \bar{a}_{pm}^u}{M} = 0, \text{ and } \bar{\theta}_{IV} \text{ is consistent for } \bar{\theta}.$$

There are no equivalent conditions under which we can achieve consistency when the d_m are included. $E[\bar{a}^u | d_m] \neq E[\bar{a}^u]$ as the fixed effects are correlated with the mean unobserved demographics.

Obtaining consistency in case (iii), in which x is modeled as having a fixed coefficient when in truth it is given heterogeneous valuations by consumers, will likewise require a condition on the relationship between a_x and a_p similar to (f). We specify this as:

Assume: (g) $E[\bar{a}_p | \bar{a}_{xm}] = E[\bar{a}_p] = 0$.

A necessary condition for (g) to hold is that \bar{a}_x and \bar{a}_p have no demographics in common, the second equality is again enforced by standardization. To build intuition, income heterogeneity may be important when consumers decide on the price range they are willing to consider for a product, but the specific product within that range may appeal to other demographics such as age, ethnicity or family characteristics.⁶

Corollary: In case (iii) with $\bar{\mu}_{jmt}^u = x_{jmt} \gamma_x^u \bar{a}_{xm}$, consistency of $\bar{\theta}_{IV}$ for $\bar{\theta}$ follows directly from assumptions (a - d) and (g).

5. A simulation study with endogenous prices

To evaluate the effect of incorporating mean demographics in the instrument set and to compare demographic and fixed effects instruments we simulate data from a model with

⁶We provide insight into bias that results when the model is misspecified in accordance with case (iii) but \bar{a}_x and \bar{a}_p share a common demographic so that assumption (g) is not satisfied. The output for this model is in the online appendix file Table 3_ExcludedOutput2.pdf. <http://hq.ssrn.com/submissions/MyPapers.cfm?partid=263724>

endogenous prices. In all simulations endogeneity results from Bertrand markups entering the price function; one simulation introduces additional endogeneity from correlation in the demand and supply shocks. The model has four dimensions: products J , markets M , time periods T and individual types I . This is more dimensions than other simulations in the literature. Berry, Linton and Pakes (2004), Romeo (2007) and Armstrong (2012) limit dimensions to J and I , while Dubé, Fox and Su (2012) and Skrainka (2011) add in markets M . We allow for time periods T to allow market fixed effects to be included in the instrument set. In addition we allow for market specific demographic draws not found in the other simulations.

Results in Armstrong (2012) inform our model formulation. We allow for multi-product firms that grow in size as J increases and allow for systematic differences in product sets between owners. This captures some of the realism of actual markets and avoids the problem of instruments based on product characteristics becoming asymptotically invalid.

We formulate the utility function as

$$\begin{aligned}
u_{ijmt} &= \theta_0 + [x_{jmt} \ln(p_{jmt})] \theta_{im} + \xi_{jmt} + \varepsilon_{ijmt} \\
\theta_{xim} &= \bar{\theta}_x + \gamma_{1x} a_{1im} + v_x v_{xim} \\
\theta_{pim} &= \bar{\theta}_p + \gamma_{1p} a_{1im} + \gamma_{2p} a_{2im} + v_p v_{pim} \\
i &= 1, \dots, I, \quad j = 0, \dots, J, \quad m = 1, \dots, M, \quad t = 1, \dots, T \\
\{\theta_0, \bar{\theta}_x, \bar{\theta}_p, \gamma_{1x}, \gamma_{1p}, \gamma_{2p}, v_x, v_p\} &= \{-8, 1, -1, -0.20, 0.25, -0.4, 0.25, 0.25\},
\end{aligned}$$

where x is a scalar and we formulate marginal costs and the price function as

$$\begin{aligned}
\ln(mc_{jmt}) &= \beta_0 + \beta_1 c_{jmt} + \eta_{jmt} \\
p_{jmt} &= -[\partial s_{jmt} / \partial p_{jmt} * O]^{-1} s_{jmt} + mc_{jmt}, \\
\{\beta_0, \beta_1\} &= \{-4, 1\},
\end{aligned}$$

where $[\partial s_{jm}/\partial p_{jm} * O]^{-1} s_{jm}$ are Bertrand markups. O is the ownership matrix and “*” indicates element-by-element multiplication.

We set $I = 100$ in all runs to limit the number of simulations and let J , M and T each take on three possible values though we exclude some M , T permutations:

$(J,M,T) = \{J = (3, 10, 25), M = (25, 50, 75), T = (5, 15, 26): \text{for each } J, \text{ with } M = 50, T = (15,26), \text{ with } M = 75, T = 26\}$.

This reduces the number of simulations from 27 to 18 for each model specification thereby substantially reducing computational time while maintaining the full range of sample sizes. Our smallest simulation has 375 simulated observations, the largest has 48,750. Each model is simulated 100 times.

We alter the ownership structure with J as follows:

$$J = \begin{cases} 3 & 3 \text{ single product firms} \\ 10 & 5 \text{ two product firms} \\ 25 & 5 \text{ five product firms.} \end{cases}$$

In constructing the single product characteristic, x , we allow it to have owner specific parameterizations, and then add a small non-systematic random component to all x to allow cross-market variation. With these design components, x acts like a composite characteristic for a product mix that shifts randomly across markets and time periods.

Marginal cost, c , also evolves slowly across markets reflecting possibly differential transportation and menu costs. Demographic a_1 is drawn from a Bernoulli distribution to give it the character of an age or ethnicity, etc. dummy variable; demographic a_2 is drawn from a log normal distribution so as to have the character of an income variable. The specifications used for a , x , and c are

$$a_{1im} \sim \text{Bernoulli}(\lambda_m), \text{ where } \lambda_m \sim U[0.2, 0.6]$$

$$a_{2im} \sim \exp(N(\mu_m, \sigma_m^2)), \text{ where } \mu_m \sim U[0.5, 1.5] \text{ and } \sigma_m^2 \sim U[0.5, 0.6],$$

$$x_{jm} = N(\alpha_j, \sigma_j) + N_{jm}(0, 0.2),$$

$$\alpha_j = \begin{cases} \{-0.15, 0, 0.15\} & \text{if } J = 3 \\ \{-0.25, -0.15, 0, 0.15, 0.25\} & \text{if } J = 10, 25 \end{cases}$$

$$\sigma_j = \begin{cases} \{1, 1.1, 1.2\} & \text{if } J = 3 \\ \{1, 1.1, 1.2, 1.3, 1.4\} & \text{if } J = 10, 25 \end{cases}$$

$$c_{jm} = N_j(0, 0.5) + N_{jm}(0, 0.1).$$

In most runs, ξ_{jm} , v_{jm} , and η_{jm} are mutually independent and are each specified as having iid standard normal distributions. We allow $\text{corr}(\xi_{jm}, \eta_{jm}) = 0.9$ in one run but find that this has little effect on the results and so limit attention to the case where correlation is strictly induced by ξ 's presence in the Bertrand markups. ε_{ijm} is given an iid Type 1 extreme value distribution. ε is integrated out and then the model is solved for the $J \times 1$ vectors of prices and shares p_m and s_m respectively for each m and t .

All instrument matrices include the BLP instruments and their squares.⁷ Costs, c , are included in most specifications, but estimator performance without costs is evaluated in two runs. In addition, since demographics induce systematic cross-market differences in the price distribution through the markup term we include instruments to control for this variation; instrument sets include either the mean of the distribution of demographics in each market or market fixed effects.

Results are presented in Tables 1 - 3 and correspond to cases (i) - (iii) respectively. Tables 1 and 2 are limited to a model where the correct specification has a fixed coefficient on x . Table 3 contains consistent estimates for one correctly and one misspecified model in which

⁷We use the exchangeability of the products in the demand system to formulate $\bar{x}_{m,-j} = (J-1)^{-1}(\sum_r x_{rm} - x_{jm})$, and include $\bar{x}_{m,-j}$ and it's square in z .

assumption (g) is satisfied. The online appendix includes two additional Table 3 results. Both are for misspecified models where assumption (g) is not satisfied.

Table 1 contains results for $\bar{\theta}_p$, the mean utility price coefficient, for correctly specified models. In this and in subsequent tables we limit attention to the mean and root mean squared error (RMSE) of $\bar{\theta}_p$ over 100 simulations and for reasons of brevity. Output files containing a full range of statistics for all demand parameters can be found in the online appendix.⁸

This table contains five columns. In Column 1 we do not include costs, mean demographics or fixed effects in the instrument set. The estimator performs poorly, the estimates are biased and the RMSEs in many cases are an order of magnitude larger than those for all other estimates in the table. This set of results indicates clearly the importance of instrumenting for cross-market price variation.

Column 2 includes fixed effects in the instrument set while Column 3 substitutes in mean demographics. Both instrument sets include cost data. Scanning the columns shows the estimator to perform well using either instrument set. The means are close to the true value of -1.00 in all but the smallest sample sizes. Comparison of the RMSEs shows fixed effects to be the preferred instruments as they produce RMSEs that are roughly half of those produced using mean demographics as instruments.

In Column 4 we retain mean demographics as instruments but remove costs. The means remain close to the true value but the RMSE roughly doubles relative to Column 3. Turning to Column 5, the results include mean demographics and costs in the instrument set and allow for correlation between the demand and supply errors. We set $\text{corr}(\xi, \eta) = 0.90$ in this run. The performance deterioration relative to Column 3 is minor. There is some increase in RMSE but no discernable deterioration in mean performance. Given this, we set $\text{corr}(\xi, \eta) = 0.0$ in all remaining runs.

8

The online appendix includes $P(\bar{\theta}_{IV} < \bar{\theta}_{\text{Truth}})$, the minimum, median, mean, maximum, bias, standard deviation and RMSE for all demand parameters. It can be found at: <http://hq.ssrn.com/submission/MyPapers.cfm?partid=263724>

Table 2 contains results of four misspecified models corresponding to case (ii). In the first two columns, the econometrician does not observe the log normal “income” demographic, while in Columns 3 and 4 the econometrician does not observe the Bernoulli “dummy” demographic. All four columns include cost instruments, Columns 1 and 3 include fixed effects, while 2 and 4 include mean demographic instruments.

A comparison of the market fixed effects and mean demographics results indicates that the fixed effects estimates are biased and inconsistent, while those using mean demographics remain consistent. The online appendix includes the empirical probability $P(\bar{\theta}_{IV} < \bar{\theta}_{Truth})$ for the 100 simulations. With fixed effects instruments in place $P(\bar{\theta}_{IV} < \bar{\theta}_{Truth}) = 1$ for almost all fixed effect results and shows no improvement with sample size.

Table 3 contains results for one correctly specified and one misspecified model corresponding to case (iii). In the correctly specified model, in Columns 1 and 2, price is interacted with the log normal “income” draws while x is interacted with the Bernoulli “dummy” draws. Misspecification is introduced in Columns 3 and 4 by dropping the random coefficient on x . Mean demographics and costs are included in the instrument sets for both models.

The correctly specified model in the first two columns shows some bias in the mean estimates of the price coefficient in particular that diminishes as sample sizes increase. The RMSEs for the price estimates are at least 20 percent larger than those for the model in Table 1, Column 2. This reduction in performance appears to be due to the addition of a second variable with a random coefficient. This is confirmed in Columns 3 and 4 of Table 3. Misspecifying the model by imposing a fixed coefficient on x improves the performance of both the price and x coefficients.

This last finding, that restricting some product characteristics to have fixed coefficients can actually improve estimator performance is valuable as the researcher is not generally aware of what characteristics and demographics are driving retailer pricing decisions. It is, however, of limited value, in that it only applies if the characteristic with the missing random coefficient does not have any demographic interactions in common with the characteristics that include random coefficients. The online appendix contains results for a model in which assumption (g) is

violated; price is interacted with both the log normal and Bernoulli demographics while x is interacted with the Bernoulli draws. Introducing a misspecification in this model by restricting x to have a fixed coefficient produces a substantial downward bias in the x coefficient, though the price coefficient still appears to be consistent.

6. Conclusions

Our sense is that finding that mean demographic are both valid and informative instruments is like found money: they have been overlooked in the literature to-date, are readily available, and are likely to be valuable in many contexts in marketing and economics. A constructive proof identifies the conditions under which these instruments are valid and shows them to be valid under misspecifications that render market fixed effects invalid. Our endogenous price simulations estimates demonstrate their validity and show that including them in the instrument set substantially improves estimator performance.

References

- Armstrong, Timothy, (2012) "Large Market Asymptotics for Differentiated Product Demand Estimators with Economic Models of Supply, ITechnical Report, Stanford University.
- Berry, Steven, James Levinsohn, and Ariel Pakes, (1995), "Automobile Prices in Market Equilibrium," Econometrica, 63, 4, 841-90.
- _____, (1999), "Voluntary Export Restraints on Automobiles: Evaluating a Trade Policy," American Economic Review, 89, 3, 400-430.
- _____, (2004), "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: the New Car Market," Journal of Political Economy, 112, 1, 68-105.
- Berry, Steven, Oliver Linton and Ariel Pakes (2004), "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems," Review of Economic Studies, 71, 613-54.
- Draper, Norman, and Harry Smith, (1981), Applied Regression Analysis, Second Edition, John Wiley & Sons, New York.
- Dubé, Jean-Pierre, Jeremy Fox, and Che-Lin Su, (2012), "Improving the Numerical Performance of BLP Static and Dynamic Discrete Choice Random Coefficients Demand Estimation," Econometrica, 80, 5, 2231-67.
- Gowrisankaran, Gautam and Marc Rysman, (2009), "Dynamics of Consumer Demand for New Durable Goods," mimeo University of Arizona and Boston University.
- Hausman, Jerry, Gergory Leonard and J. Douglas Zona, (1994), "Competitive Analysis with Differentiated Products," Annales D'Economie et de Statistique, 34, 159-80.
- Nevo, Aviv, (2000), "Mergers with Differtiated Products: the Case of the Ready-to-Eat Cereal Industry," RAND Journal of Economics, 31, 3, 395-421.
- _____, (2001), "Measuring Market Power in the Ready-to-Eat Cereal Industry," Econometrica, 69, 2, 307-42.
- Newey, Whitney, (1990), "Efficient Instrumental Variables Estimation of Nonlinear Models," Econometrica, 58,4, 809-38.

- Pakes, Ariel, (1996), "Dynamic Structural Models, Problems and Prospects: Mixed Continuous Discrete Controls and Market Interactions," in Advances in Econometrics: Sixth World Congress, Vol.. II, edited by Christopher Sims, 171-260.
- Petrin, Amil, (2002), "Quantifying the Benefits of New Products: the Case of the Minivan," Journal of Political Economy, 110, 4, 705-29.
- Romeo, Charles, (2012), "Incorporating Prior Information into a GMM Objective for Mixed Logit Demand Systems," Working Paper, Consumer Financial Protection Bureau.
- _____ (2007), "A Gibbs Sampler for Mixed Logit Analysis of Differentiated Product Markets using Aggregate Data," Computational Economics, 29, 33-68.
- Skrainka, Benjamin, (2011), "A Large-Scale Study of the Small Sample Performance of Random Coefficient Models of Demand, Working Paper, University of Chicago, 2011.
- Villas-Boas, Sofia, (2007), "Vertical Relationships between Manufacturers and Retailers: Inference with Limited Data," Review of Economic Studies, 74, 2, 625-52.

Table 1. Results for correctly specified case (i) models with a fixed coefficient on x based on 100 simulations. Mean and RMSE of the price coefficient using fixed effects or mean demographics in the instrument set with and without costs and correlation in the demand-supply error distributions: truth = -1.00

| fixed effects: mean demographics: costs: corr(ξ, η): | | | no no no 0 | yes no yes 0 | no yes yes 0 | no yes no 0 | no yes yes 0.90 |
|--|----|----|---------------------|-----------------------|-----------------------|----------------------|--------------------------|
| J | M | T | mean (RMSE) | | | | |
| 3 | 25 | 5 | -0.429 (4.214) | -1.149 (0.237) | -0.816 (1.042) | -1.030 (1.236) | -0.697 (1.239) |
| | | 15 | -0.386 (1.865) | -1.061 (0.126) | -1.010 (0.180) | -1.076 (0.570) | -0.957 (0.212) |
| | | 26 | -0.763 (3.180) | -1.047 (0.104) | -1.018 (0.146) | -1.048 (0.406) | -0.994 (0.196) |
| | 50 | 15 | -1.012 (4.979) | -1.040 (0.087) | -0.935 (0.444) | -1.106 (0.752) | -0.967 (0.370) |
| | | 26 | -0.481 (1.753) | -1.032 (0.064) | -0.974 (0.363) | -0.934 (0.816) | -0.871 (0.554) |
| | 75 | 26 | -0.964 (3.767) | -1.054 (0.077) | -1.033 (0.119) | -1.036 (0.433) | -1.006 (0.194) |
| 10 | 25 | 5 | -0.529 (1.095) | -1.149 (0.170) | -1.063 (0.241) | -1.036 (0.324) | -1.033 (0.561) |
| | | 15 | -0.648 (0.817) | -1.064 (0.086) | -1.035 (0.117) | -1.022 (0.206) | -1.034 (0.136) |
| | | 26 | -0.779 (0.451) | -1.043 (0.058) | -1.023 (0.086) | -1.001 (0.145) | -1.011 (0.126) |
| | 50 | 15 | -0.874 (0.370) | -1.011 (0.039) | -1.013 (0.105) | -1.006 (0.192) | -0.999 (0.152) |
| | | 26 | -0.851 (0.353) | -1.006 (0.030) | -1.007 (0.078) | -1.025 (0.110) | -1.011 (0.106) |
| | 75 | 26 | -0.803 (0.453) | -1.032 (0.042) | -1.013 (0.083) | -1.002 (0.106) | -1.027 (0.107) |
| 25 | 25 | 5 | -0.625 (0.740) | -1.092 (0.104) | -0.963 (0.272) | -0.914 (0.399) | -0.982 (0.404) |
| | | 15 | -0.851 (0.425) | -1.047 (0.057) | -1.005 (0.101) | -1.001 (0.269) | -1.010 (0.127) |
| | | 26 | -0.822 (0.380) | -1.032 (0.042) | -0.999 (0.064) | -0.996 (0.149) | -1.005 (0.086) |
| | 50 | 15 | -0.825 (0.475) | -0.998 (0.024) | -1.036 (0.076) | -1.062 (0.185) | -1.040 (0.093) |
| | | 26 | -0.869 (0.304) | -0.998 (0.019) | -1.018 (0.068) | -1.016 (0.114) | -1.023 (0.092) |
| | 75 | 26 | -0.837 (0.343) | -1.010 (0.017) | -0.978 (0.072) | -0.974 (0.088) | -0.985 (0.088) |

Table 2. Results for case (ii) misspecifications based on 100 simulations. Mean and RMSE of the price coefficient using fixed effects or mean demographics in the instrument set with costs in misspecified models: truth = -1.00.

| fixed effects: mean demographics: costs: missing random coef: | | | yes no yes income | no yes yes income | yes no yes dummy | no yes yes dummy |
|--|----|----|----------------------------|----------------------------|---------------------------|---------------------------|
| J | M | T | mean (RMSE) | | | |
| 3 | 25 | 5 | -1.190 (0.262) | -0.861 (0.828) | -1.372 (0.412) | -1.011 (0.546) |
| | | 15 | -1.128 (0.159) | -1.053 (0.239) | -1.392 (0.405) | -0.987 (0.393) |
| | | 26 | -1.141 (0.169) | -1.072 (0.215) | -1.406 (0.416) | -1.037 (0.277) |
| | 50 | 15 | -1.179 (0.196) | -0.886 (0.751) | -1.378 (0.387) | -0.984 (0.297) |
| | | 26 | -1.180 (0.190) | -0.999 (0.404) | -1.385 (0.391) | -1.029 (0.232) |
| | 75 | 26 | -1.272 (0.276) | -1.061 (0.152) | -1.427 (0.430) | -1.040 (0.189) |
| 10 | 25 | 5 | -1.224 (0.235) | -1.047 (0.305) | -1.356 (0.361) | -1.051 (0.387) |
| | | 15 | -1.176 (0.181) | -1.029 (0.217) | -1.345 (0.348) | -0.971 (0.289) |
| | | 26 | -1.173 (0.177) | -1.044 (0.160) | -1.360 (0.362) | -0.999 (0.220) |
| | 50 | 15 | -1.193 (0.196) | -1.023 (0.165) | -1.338 (0.340) | -1.009 (0.213) |
| | | 26 | -1.194 (0.196) | -1.023 (0.157) | -1.343 (0.344) | -0.955 (0.178) |
| | 75 | 26 | -1.249 (0.250) | -1.037 (0.110) | -1.373 (0.375) | -1.003 (0.159) |
| 25 | 25 | 5 | -1.157 (0.162) | -0.928 (0.357) | -1.252 (0.256) | -1.069 (0.232) |
| | | 15 | -1.154 (0.156) | -0.991 (0.197) | -1.262 (0.263) | -1.027 (0.218) |
| | | 26 | -1.157 (0.159) | -0.983 (0.153) | -1.270 (0.271) | -0.976 (0.210) |
| | 50 | 15 | -1.167 (0.169) | -0.997 (0.117) | -1.260 (0.261) | -0.990 (0.185) |
| | | 26 | -1.172 (0.173) | -0.993 (0.361) | -1.260 (0.261) | -1.002 (0.139) |
| | 75 | 26 | -1.193 (0.194) | -0.988 (0.088) | -1.268 (0.269) | -0.973 (0.166) |

Table 3. Models having random coefficients on price and x. Results for correctly specified and case (iii) misspecified models based on 100 simulations. Mean and RMSE of price and x coefficients using mean demographics and costs in the instrument set: truth = -1.00 for price; 1 for x.

| missing random coef: | | | none | | dummy on x | |
|----------------------|----|----|----------------|---------------|----------------|---------------|
| coefficient: | | | price | x | price | x |
| J | M | T | mean (RMSE) | | | |
| 3 | 25 | 5 | -0.910 (0.322) | 1.043 (0.270) | -0.921 (0.282) | 0.988 (0.060) |
| | | 15 | -0.945 (0.235) | 1.053 (0.223) | -0.956 (0.176) | 0.992 (0.039) |
| | | 26 | -0.959 (0.208) | 1.037 (0.197) | -0.965 (0.165) | 0.997 (0.029) |
| | 50 | 15 | -0.920 (0.260) | 1.038 (0.208) | -0.950 (0.193) | 1.000 (0.028) |
| | | 26 | -0.962 (0.229) | 1.035 (0.172) | -0.972 (0.147) | 1.003 (0.020) |
| | 75 | 26 | -0.934 (0.232) | 1.061 (0.256) | -0.973 (0.145) | 0.997 (0.019) |
| 10 | 25 | 5 | -0.859 (0.291) | 1.073 (0.234) | -0.872 (0.273) | 0.996 (0.029) |
| | | 15 | -0.905 (0.171) | 1.032 (0.134) | -0.919 (0.136) | 1.001 (0.020) |
| | | 26 | -0.965 (0.088) | 1.023 (0.102) | -0.967 (0.073) | 0.999 (0.012) |
| | 50 | 15 | -0.972 (0.142) | 0.999 (0.094) | -0.990 (0.114) | 0.999 (0.014) |
| | | 26 | -0.951 (0.136) | 1.015 (0.101) | -0.959 (0.107) | 1.001 (0.011) |
| | 75 | 26 | -0.982 (0.132) | 0.999 (0.106) | -0.990 (0.084) | 1.000 (0.008) |
| 25 | 25 | 5 | -0.865 (0.186) | 1.067 (0.107) | -0.852 (0.199) | 0.996 (0.018) |
| | | 15 | -0.928 (0.108) | 1.046 (0.107) | -0.925 (0.109) | 0.999 (0.011) |
| | | 26 | -0.954 (0.084) | 1.033 (0.087) | -0.953 (0.070) | 1.000 (0.009) |
| | 50 | 15 | -0.981 (0.092) | 1.007 (0.066) | -0.999 (0.075) | 0.999 (0.008) |
| | | 26 | -0.978 (0.086) | 1.021 (0.092) | -0.995 (0.064) | 0.999 (0.006) |
| | 75 | 26 | -0.940 (0.130) | 1.047 (0.117) | -0.982 (0.065) | 0.999 (0.005) |