Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market

by

Steven Berry, James Levinsohn and Ariel Pakes

November 2001

COWLES FOUNDATION DISCUSSION PAPER NO. 1337

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market [1]

Steven Berry
Dept. of Economics
Yale Univ.
and NBER

James Levinsohn
Dept. of Economics
Univ. of Michigan
and NBER

Ariel Pakes
Dept. of Economics
Harvard Univ.
and NBER

November 5, 2001

[1] We thank numerous seminar participants, two referees, and the editor Lars Hansen for helpful suggestions. We also thank the NSF for financial support, through grants 9122672, 9512106 and 9617887. We are particularly grateful to Dr. G. Mustafa Mohatarem at the General Motors Corporation, who made possible our access to the data and for futher help from GM's Dr. Robert Bordley. Gautam Gowrisankaran, Dan Ackerberg, Lanier Benkard, Amil Petrin, and Nadia Soboleva provided invaluable assistance. Now that they are all successful academics, we hope their own research assistants come close to matching their standard. The most recent revision of this paper would not have been possible without Nadia Soboleva's expert advice.

Abstract

In this paper we provide an algorithm for estimating characteristic based demand models from alternative data sources, and apply it to new data on the market for passenger vehicles. We find that, provided care is taken in constructing the demand system and rich enough data are available, the characteristic based model can both rationalize existing results and provide realistic out of sample predictions.

1 Introduction

In Berry, Levinsohn, and Pakes (1995) (BLP) we provide an algorithm for estimating the parameters of a class of differentiated product demand models whose foundations date back at least to Lancaster (1971) and McFadden (1974). In these models, products are bundles of characteristics and consumers have preferences defined on this characteristic space. Each consumer chooses the product that maximizes her utility, and market demand is obtained from the explicit aggregation of consumers' choices. The primitive demand parameters to be estimated are the distribution of consumers' preferences over characteristics. In contrast to the more traditional demand systems defined on a product space, the number of these parameters does not depend on the number of products marketed. Also, characteristic based demand models can be used to analyze the potential demand for products before they are brought to market. Consequently these models are used increasingly in the study of differentiated product markets.

Though BLP models the distribution of household preferences, they use only product-level data (i.e. aggregate quantities, prices, and product characteristics) and information on the distribution of household characteristics available from the Current Population Survey (CPS), data that are available for a wide variety of markets. Since the BLP data did not contain enough information to allow them to obtain reliable parameter estimates from the demand side alone, BLP added information in the form of a pricing assumption. The estimates they then obtained suggested two lessons: realistic estimates of substitution patterns require models which allow consumers to differ in the values they attach to product characteristics, and reasonable own price elasticities require a model which allows for unobserved product characteristics that are correlated with price.

In addition to product-level data in BLP, this paper adds micro data that matches consumer attributes to purchased vehicles. To solve the simultaneity problem generated by the unobserved product characteristics, BLP needed an identifying assumption that justified instruments. Sections 2 and 3, which provide our model and estimation algorithm, show that the simultaneity problem takes on a different form when micro data is available, but does not disappear. With micro data an important subset of parameters are identified without an assumption similar to BLP's, but there are additional parameters which are needed for most (though not all) applications that are not.

A central empirical issue is whether the sources of consumer heterogeneity typically available in micro data sets (income, household demographics, location of residence) are rich enough to account for the heterogeneity in tastes for different characteristics. Our model allows tastes for characteristics to vary as a function of both observed and unobserved consumer attributes. For comparison we also estimate models with only observed attribute differences (analogous to a logit model) and one with only unobserved attribute differences (analogous to BLP's model).

Our micro data, which is described in section 4, is an extremely rich proprietary data set graciously provided to us by the *General Motors Corporation*. In particular it contains the second (as well as the actual) choice of consumers who purchased vehicles in 1993. The second choice data enables us to obtain both reliable estimates of the importance of unobserved attributes, and to compare predicted second choices to those reported in the data. The results, presented in sections 5 and 6, are striking; we *cannot* obtain realistic substitution patterns from a model which does not allow for unobservable consumer attributes. However once we allow for the unobservables, the model does well. Interestingly, the model with only unobserved differences in preferences for vehicle characteristics also does quite well in this respect. Finally, we found that we could not obtain reliable estimates of the model's parameters from combining only the first-choice micro and product level data; we also needed the information on second choices (though either repeat purchase data, or other data with variance in the choice set, might mitigate this need).

We examine the rich set of substitution patterns generated by our model and show that they (and their implications on markups) are in accord with available information. Section 7 uses two "out of sample" prediction exercises to illustrate other applications of our estimates. First, we introduce new "high end" sport utility vehicles similar to those actually introduced in the late 1990's and predict both demand for the new entrant and its effects on the sales of incumbents. Second, we consider the likely implications of an important production decision recently made by *GM*: the shutting down of their Oldsmobile division.

2 The Model

The model in BLP is a model of household choice which is then explicitly aggregated to obtain product level demands. It is therefore able to analyze *both* our micro data on household choices and our aggregate data on product level demands in one consistent framework. We now review that framework emphasizing the role different information sources play in identifying various parameters[1].

Largely for simplicity, we use a linear version of the utility, u_{ij}, that consumer i obtains from the choice of product j (this follows the traditional discrete choice random coefficients literature; *e.g.*, Domenich and McFadden (1975), or Hausman and Wise (1978)). Let $j = 0, \dots, J$ index the products competing in the market, where product $j = 0$ is the "outside" good (so that u_{i0} is the utility of the consumer if she does not purchase any of these J goods and instead allocates all income to other purchases). Let k index the observed (by us) product characteristics, including price, and r index the observed household attributes.

Our model is then

$$u_{ij} = \Sigma_k x_{jk} \tilde{\beta}_{ik} + \xi_j + \epsilon_{ij}, \tag{1}$$

with

$$\tilde{\beta}_{ik} = \bar{\beta}_k + \sum_r z_{ir} \beta^o_{kr} + \beta^u_k \nu_{ik}, \tag{2}$$

where:

- the x_{jk} and ξ_j are, respectively, observed and unobserved *product* characteristics,
- the $\tilde{\beta}_{ik}$ represent the "taste" of consumer i for product characteristic k,

[1]We change BLP's notation to slightly to facilitate our alternative data sources.

- the z_i and ν_i are vectors of observed and unobserved *consumer* attributes, and
- the ϵ_{ij} represent idiosyncratic individual preferences, assumed to be independent of the product attributes and of each other.

Note that the model allows consumers to differ in their tastes for different product characteristics. Those differences (the $\tilde{\beta}$) are allowed (via equation (2)) to depend on *both* consumer attributes observed by the econometrician (through β^o where the "o" superscript is for "observed") and attributes that the econometrician does not observe (through β^u, where "u" is for "unobserved")[2]. In our example the z vectors contain consumer attributes listed in our data (*e.g.* income, family size, and age of household head), while the ν vectors allow for consumer attributes that are not in our data (*e.g.* distance to work or a need to transport a little league team). Similarly, the x_k are auto characteristics that we measure (*e.g.* price, size, and horsepower) and the ξ are unmeasured aspects of car quality.

We want to stress two features of this framework: the interaction terms and the product specific constant terms. First, as noted in the earlier literature (see McFadden, Talvitie, and Associates (1977), Hausman and Wise (1978) and BLP), the interaction between consumer tastes and product characteristics determines substitution patterns in discrete choice models. As the variance in the random tastes for product characteristics increases, similar products (in the space of x's) become better substitutes. Models without individual differences in preferences for characteristics generate demand elasticities that are known to be *a priori* unreasonable (depending only on market shares and not on the characteristics of the vehicles). A goal of this paper is to provide accurate measures of elasticities and so we allow for unobserved (as well as observed) determinants of characteristic preferences.

Second, vehicles (and most other consumer products) are differentiated from one another in many dimensions. We will include characteristics that proxy for the most important sources of differentiation, but even if we had the data we could not hope to estimate the distribution of preferences over a set of characteristics that is large enough to capture all aspects of product differentiation. The role of the unobserved product characteristic, ξ, is to pick up the total impact of the characteristics not included in our specification. As stressed in Berry (1994) and in BLP one might expect ξ to be correlated with price: products with higher unmeasured quality might sell at a higher price. This is the differentiated product analogue of the standard "simultaneity" problem in demand analysis, and our previous work indicates that when we do not account for this correlation we obtain unreasonably small (in absolute value) price elasticities.

The consumer level choice model is found by substituting equation (2) into (1) to obtain

$$u_{ij} = \delta_j + \sum_{kr} x_{jk} z_{ir} \beta^o_{kr} + \sum_{k} x_{jk} \nu_{ik} \beta^u_k + \epsilon_{ij}, \tag{3}$$

[2]Equations (1) and (2) make several simplifying assumptions, including that there is only one unobserved product characteristic, and consumers do not differ in their preferences for it. These simplifications are not necessary to the arguments that follow, though they simplify both the exposition and the subsequent computations; see Heckman and Snyder (1997) for a related model with a higher dimension of unobserved characteristics, and Das, Olley, and Pakes (1995) for an attempt to let consumers differ in their preferences for the unobserved characteristic in this model.

where for $j = 0, 1, \ldots, J$.

$$\delta_j = \sum_k x_{jk}\bar{\beta}_k + \xi_j, \tag{4}$$

This equation clarifies two important points about the identification of our model. First, even without an assumption on the joint distribution of (ξ, x) the micro data allows us to estimate some but *not all* of the parameters of the model. Second, the remaining parameters determine the elasticities of interest and identifying these parameters requires assumptions of the sort used in market-level data.

To see that some parameters are identified without assumptions on (ξ, x), note that equation (3) defines a traditional random coefficients discrete choice model with choice-specific constant terms, δ_j. Given parametric assumptions on (ν, ϵ) and standard regularity conditions, we can therefore obtain consistent estimators of the parameter vector $\theta = (\delta, \beta^o, \beta^u)$ from micro data (like our CAMIP data) without assumptions about the unobservable ξ's [3].

However, knowledge of $\theta = (\delta, \beta^o, \beta^u)$ does not identify own and cross price (and characteristic) elasticities.[4] Unless product characteristics have no systematic effect on demand ($\bar{\beta} \equiv 0$), the choice-specific constant δ is itself a function of product characteristics. Thus to calculate the impact of, say, price on demand, we need to know the impact of price on δ, i.e. we need $\bar{\beta}$.

Equation (4) indicates that the number of observations on δ that can be used to estimate $\bar{\beta}$ equals the number of products: effectively we have to estimate $\bar{\beta}$ from the product level data. Consequently we cannot identify $\bar{\beta}$ without some assumption on the joint distribution of (ξ, x). This is exactly the same identification problem faced by BLP. As noted in BLP and elsewhere (Nevo 2000), different assumptions on the joint distribution of (ξ, x) can be used to identify the remaining parameters. To account for the simultaneity problem, BLP assume the ξ_j are mean independent of the *non-price* characteristics of all of the products. We make use of this and other possible restrictions below.

To return to the implications of our model, market-level aggregate consumer behavior is obtained by summing the choices implied by the individual utility model over the population's distribution of consumer attributes. Let w_i be the vector of both the observed (z_i) and unobserved (ν_i, ϵ_i) individual attributes

$$w_i = (z_i, \nu_i, \epsilon_i),$$

and denote its distribution in the population by $\mathcal{P}_w$. The fraction of households that choose good j (aggregate demand) is given by integrating over the set of attributes that imply a preference for good j:

$$s_j(\delta, \beta^o, \beta^u; x, \mathcal{P}_w) = \int_{A_j(\delta, \beta^o, \beta^u; x)} \mathcal{P}_w(dw) \tag{5}$$

where

$$A_j(\delta, \beta^o, \beta^u; x) = \{w : \max_{r=0,1,\ldots,J} [u_{ir}(w; \delta, \beta^o, \beta^u, x)] = u_{ij}\}.$$

[3] See also Ichimura and Thompson (1998) who discuss non and semi-parametric identification.

[4] Some questions of interest do not require these elasticities. One important example is the calculation of ideal price indices, see Pakes, Berry, and Levinsohn (1993), while section 7 contains another example.

Just as the basic form of equation (1) is familiar from the econometric discrete choice literature (see, for *e.g.* McFadden (1981)), the notion of aggregating discrete choices to market demand has been used extensively in the Industrial Organization literature on product differentiation. An early example is Hotelling (1929), while Anderson, DePalma, and Thisse (1992) provide a more recent discussion with extensive references.

3 Estimation

We begin with an outline of our estimation procedure focusing on the role it gives to alternative data sources. The reader who is not interested in the technical detail should be able to proceed directly from this subsection (3.1) to the section that introduces the data (4). Subsection (3.2) explains how we compute the objective function. An appendix provides the relationship between our estimation procedure and maximum likelihood and outlines how we construct our standard errors.

3.1 Outline of the Estimation Procedure.

Since our micro data allow us to estimate choice specific constant terms, we faced a choice of whether to estimate the vector $\theta = (\beta^o, \beta^u, \delta)$ or to impose enough additional restrictions on the joint distribution of (ξ, x) to enable us to identify $\bar{\beta}$ and only estimate $(\beta^o, \beta^u, \bar{\beta})$. Formally the trade-off here is familiar: gaining efficiency from additional restrictions versus losing consistency if those restrictions are wrong.

We chose to estimate θ without imposing any additional restrictions for two reasons. First the CAMIP data set is large so we are not particularly concerned with precision. Second, as noted in BLP, the distribution of (ξ, x) is partly determined by product development decisions, so *a priori* restrictions on it are hard to evaluate. Our choice implies estimates of (β^o, β^u) that are robust to assumptions on the (ξ, x) distribution. We then use the estimated δ's to estimate $\bar{\beta}$ using various assumptions on (ξ, x) (section 6).

Efficiency considerations argue for using maximum likelihood estimates of θ, but this was too computationally burdensome (see Appendix A). Therefore, we use a method of moments estimator. This compares the moments predicted by our model for different values of θ to our sample's moments and then chooses the value of θ which minimizes the "distance" between the model's predictions and the data.

We matched three "sets" of predicted moments to their data analogs:

1. The covariances of the observed first-choice product characteristics, the x, with the observed consumer attributes, the z (for example, the covariance of family size and first choice vehicle size);

2. The covariances between the first choice product characteristics and the second choice product characteristics (for example, the covariance of the size of the first choice vehicle with the size of the second choice vehicle); and

3. The market shares of the J products.

The first set of moments match observed consumer attributes to the characteristics of the chosen vehicles. We think of these moments as particularly useful for estimating β^o, the coefficients on the interactions between observed product characteristics and household attributes (x and z) [5]. If the first choice car characteristics are denoted by x^1 and z denotes household attributes, we fit the model's predictions for $E(x^1 z')$ and for $E(z)$ to their CAMIP sample analogues. We include in $E(x^1 z')$ a separate moment condition for each interaction term in the utility specification. Since the CAMIP sampling rates are roughly in proportion to market share, the expectation $E(z)$ is roughly the expected value of the attributes of households who chose to buy a car. The $E(z)$ moments are therefore particularly useful in estimating the parameters that define the utility of the outside good.

The second set of moments, between first and second choice characteristics, are particularly useful in identifying the importance of the unobserved consumer characteristics. Note that if all relevant consumer attributes were observed ($\beta^u = 0$), then the coefficients of the observed consumer attributes, β^o, would determine both the first and second choice vehicle characteristics and hence the correlation between them. If the model with $\beta^u \equiv 0$ predicts a first/second choice correlation that is much less than the correlation found in the data, we would conclude that the β^u are necessary to explain observed substitution patterns. Our specification has one element of β^u for each included car characteristic and we include a predicted first/second choice covariance for each such characteristic.

As noted in Berry (1994), given $\beta \equiv (\beta^o, \beta^u)$ there is a unique δ which matches the observed market shares equal to the model's predicted share. So the third set of moments are particularly useful in estimating the δ parameters.

3.2 The Fitted Moments

This section explains how we compute the moments that go into our method of moments estimation algorithm and considers the limit distribution of the parameter estimates. This requires some additional notation, an introduction to our data sets, and assumptions on the joint distribution of the household attributes.

Letting N indicate the number of households in the U.S. population (about 100 million), the product level data consists of J couples, (s_j^N, x_j), where s_j^N is the share of the population that purchased vehicle j, and x_j is a vector of the vehicle's observed characteristics (one of which is price, p_j). $s_0^N = 1 - \sum_j s_j^N$ is the fraction of the population that does not purchase one of our J vehicles. Our model implies that the market shares observed in the data, say, s^N distributes multinomially about $s(\delta_0, \beta_0; x, \mathcal{P}_w)$, where (β_0, δ_0) represent the true value of that vector, and has a covariance matrix whose elements are all less than N^{-1} .

The consumer level, or CAMIP, data is a choice based sample drawn from new vehicle registrations. GM determines the number of households to sample from the registrations

[5] If $\beta^o = 0$, and we used only first choice data, then the aggregate shares used in BLP would be sufficient statistics for the first choice data, and the match of individuals to the car they chose would contain no additional information.

for each vehicle, say n_j, and then the characteristics of the households sampled and their second choice vehicle are found. We let $n = \sum_j n_j$ and index the number of households in the CAMIP data by $i = 1, \ldots, n$. $y_i^1 = j$ is our notation for the event that the first choice of household i is vehicle j, while $y_i^2 = k$ indicates that the second choice is vehicle k.

To derive the predictions of the model we have to specify a joint distribution for the observed and unobserved consumer attributes; the z_i, and the (ν_i, ϵ_i) couples. Since the CPS is a random sample of US households, we can use it to sample from $\mathcal{P}_z$ directly. The (ν, ϵ) couples are assumed to distribute independently of z and of each other. Recall that the means of these variables go into the constant terms (the δ). We assume that the deviation from the means (our ν) are independent, normal random variables. Thus β_k^u can be interpreted as the standard deviation of the unobserved distribution of tastes for vehicle characteristic k. The sole exception to this is the unobserved characteristic that interacts with price which is assumed to be log-normal (this allows us to impose the constraint that no one prefers higher prices, see equation (14) below for more detail). These assumptions give us the marginal distribution of ν, denoted $\mathcal{P}_\nu$.

Finally, for computational simplicity we assume that the idiosyncratic errors, the ϵ_{ij}, have an i.i.d. extreme value "double exponential" distribution. This assumption yields the logit functional form for the model's choice probabilities *conditional* on a (z, ν) couple

$$Pr(y_i^1 = j | z_i, \nu_i, \theta, x) = \frac{exp[\delta_j + \sum_{kr} x_{jk} z_{ir} \beta_{kr}^o + \sum_k x_{jk} \nu_{ik} \beta_k^u]}{1 + \sum_q exp[\delta_q + \sum_{kr} x_{qk} z_{ir} \beta_{kr}^o + \sum_k x_{qk} \nu_{ik} \beta_k^u]}. \quad (6)$$

Note that the choice probabilities in (6) are an easy to calculate function of z, ν and θ.

We now move to the computation of our moments. The moments for the aggregate shares are treated slightly differently in order to solve another computational problem. Since we have over 200 car models, δ has 200 elements and a search over θ is a search over about 250 dimensions. Since we cannot search over that many dimensions effectively, we use the aggregate moments to "concentrate out" the δ parameter, and search only over β.

Recall that the variance of $s^N - s(\delta_0, \beta_0; x, \mathcal{P}_w)$ is of order N^{-1} and $N^{-1} \approx 0$. Consequently if we could calculate $s(\cdot)$ exactly an efficient method of moments algorithm would chose θ so that $s^N \approx s(\cdot)$. Berry (1994) shows that for any $[\beta, \mathcal{P}_w, s^N]$ there is a unique δ vector that makes $s^N = s(\delta, \beta; \cdot)$ and BLP provide a contraction mapping which *quickly* computes its value, say $\delta(\beta, s^N; \cdot)$.

We would like to; (i) use this contraction mapping to find the unique value of δ that makes $s^N \equiv s(\beta, \delta; \cdot)$ for each guess at β, (ii) substitute that $\delta(\beta, s^N; \cdot)$ for δ into the model's predictions for the micro moments making them a function of $(\beta, \delta(\beta, s^N; \cdot))$, and (ii) then search to find the value of β that minimizes the distance between those predictions and the data (thus eliminating any search over δ).

To do this we need to compute the market shares predicted by our model for different values of θ; i.e. to integrate the probability in equation(6) over the distribution of (z, ν). Unfortunately that integral does not have an analytic form. Consequently we follow Pakes (1986) and use simulation to approximate its value. Specifically, let (z_r, ν_r) for $r = 1, \ldots, ns$, index ns random draws on a couple whose first component, z_r, is taken from the CPS and whose second component, ν_r, is taken from the assumed distribution of ν. We then define $\delta^{ns,N}(\beta)$ implicitly as the value of this vector that sets [6]

$$G^3_{ns,N}(\theta) = s^N_j - \frac{1}{ns}\sum_{r=1}^{ns} Pr(y_1 = j | z_r, \nu_r, \beta, \delta^{ns,M}(\beta)) \tag{7}$$

to zero (this can be found quickly with BLP's contraction mapping).

Note that we draw the (z_r, ν_r) couples once at the beginning of the algorithm and hold them constant thereafter. This insures that the limit theorems in Pakes and Pollard (1989) apply to our estimators. This use of simulation does, however, put simulation error in our estimates of δ given β and this affects the asymptotic variance of the estimates of β (see the appendix).

Next we calculate the model's predictions for the covariances between the first choice car characteristics and household attributes. Since the CAMIP data is choice based the moments we have to fit to the data are the model's predictions for the attributes of a household who chose a particular vehicle. To form the sample moment we interact the average attributes of households who chose vehicle j with the characteristics of that vehicle, and then average over the different vehicles (using the CAMIP sampling weights). That is, our first choice moments are

$$G^1_{n,ns,N}(\beta) \ \approx \ \Sigma_j \frac{n_j}{n} x^1_{kj} \left\{ (n_j)^{-1} \Sigma_{i_j=1}^{n_j} z_{i_j} \ - \ E[z | y^1_i = j, \beta] \right\}, \tag{8}$$

where, at the risk of some misunderstanding, it is now understood that when we condition on β we are conditioning on $(\beta, \delta^{ns,N}(\beta; \cdot))$.

We use an approximation sign in equation (8) to indicate that we can not calculate $E[z|y^1 = j, \beta]$ exactly. To obtain our approximation we use Bayes rule to rewrite[7]

$$E[z|y^1 = j, \beta] = \int_z z\mathcal{P}(dz|y^1 = j, \beta) = \frac{\int_z zPr(y_1 = j|z, \beta)\mathcal{P}(dz)}{Pr(y_1 = j, \beta)}$$

and substitute from the model's predictions for the choice probabilities (equation 6) to obtain

$$E[z|y^1 = j, \beta] = \frac{\int_z \int_\nu zPr(y_1 = j|z, \nu, \beta)\mathcal{P}(dz, d\nu)}{Pr(y_1 = j, \beta)}. \tag{9}$$

For each value of β, our model's prediction for the denominator of (9) will, by virtue of the choice of $\delta^{N,ns}(\beta)$, exactly equal s^N_j. However we have to simulate the integral in the numerator. Using the same draws on (z_r, ν_r) we used in equation (7) we obtain our approximation as

$$E[z|y^1 = j, \beta] \approx \frac{(ns)^{-1}\Sigma_r z_r Pr\left(y_1 = j|z_r, \nu_r, \beta, \delta^{ns,N}(\beta)\right)}{s^N_J}. \tag{10}$$

[6]In practice we don't just take random draws from the distributions of z and ν but rather use importance sampling techniques, analogous to those used in BLP, to reduce the variance of our estimated integrals.

[7]This follows the literature on choice based sampling; see Manski and Lerman (1977) Cosslett (1981), and Imbens and Lancaster (1994)

The first choice moments we use are formed by substituting (10) into (8).

An analogous procedure is used to form the moments for the covariances between the characteristics of the first and second choice vehicles. Consider only the households whose first choice was vehicle j. For those households, the difference between the the average value of characteristic k of the second choice vehicle they list in their responses, and the average value of characteristic k for the second choice vehicles predicted by our model is

$$\left(\frac{1}{n_j}\sum_{i=1}^{n}\sum_{q\neq j} x_{kq}\{y_i^2 = q\}\{y_i^1 = j\}\right) - \left(E[\sum_{q\neq j} x_{kq}\{y_i^2 = q\} \mid y^1 = j, \beta]\right), \tag{11}$$

where $\{y_i^2 = q\}$ is the indicator function for the event that vehicle q is the second-choice. We interact this difference with x_{kj}^1 and use the CAMIP sample weights to average over first choices to obtain the moment

$$G^2_{n,ns,N}(\beta) \approx \tag{12}$$

$$\sum_j \frac{n_j}{n} x_{kj}^1 \sum_{q\neq j} x_{kq}\left[(\frac{1}{n_j}\sum_{i=1}^{n}\{y_i^2 = q\}\{y_i^1 = j\}) - \int_z\int_\nu Pr(y^2 = q \mid y^1 = j, z, \nu, \beta)\mathcal{P}_z(dz)\mathcal{P}_\nu(d\nu)\right]$$

To calculate the expectation in (12) we note that the second choice probabilities conditional on $(y^1 = j, z, \nu, \beta)$, i.e., $Pr(y^2 = k \mid y^1 = j, z, \nu, \beta)$, are given by the standard "logit" form in (6) modified to take both vehicle j and the outside alternative out of the choice set (this changes the denominator in the choice probability, eliminating both the "one" and the "j^{th}" element in the summation sign). After substituting this into the integrand in (12) we approximate that integral by simulation (as in 8).

We stack $G^1(\cdot)$ and $G^2(\cdot)$ and use the two step generalized method of moments (GMM) estimator (see (Hansen 1982)) of β from the stacked moments. Provided $ns \to \infty$ and $N \to \infty$ as $n \to \infty$ standard arguments show that this estimator is consistent. Since N is large relative to n and ns in our example, we use the limit distribution for β that assumes that as $n \to \infty$, $N/n \to \infty$, but ns/n converges to a positive constant (this insures that we adjust our variances for simulation error). That limit distribution is normal and the appendix explains how to obtain consistent estimates of its covariance matrix.

4 Data

We begin with a description of the CAMIP data. It contains the results of a propriety survey conducted on behalf of the *General Motors Corporation* (GM) and is generally not available to researchers outside of the company. This survey is a sample from the set of vehicle registrations in the 1993 model year. For each vehicle, a given number of purchasers is sampled. The intent is to create a random sample conditional on purchased vehicle. The sampled vehicles consist of almost all vehicles sold in the U.S. in 1993, not just GM products. The subsample we use contains 37,500 observations (see appendix C for more details).

The CAMIP questionnaire asks about a limited number of household attributes, including income, age of the household head, family size and place of residence (urban, rural, etc.). We match each of the household attribute questions to a question in the CPS[8]. Table 1 compares the distribution of household characteristics in the CAMIP sample to those in the CPS. Not surprisingly CAMIP samples disproportionately from higher income groups. Households who buy new vehicles, especially high priced ones, tend to have disproportionately high incomes. A more surprising difference between the two samples is that the CAMIP sample is significantly less urban and more rural than the overall U.S. population. Apparently, the rural population purchases a disproportionate number of vehicles, which helps explain the high share of trucks in total vehicle sales.

Table 1: Comparison of Consumer Samples.

Income (in thousands)

Income Range	% in CPS	% in CAMIP	CPS Group Mean	CAMIP Mean
0−36.5	64.17	25.00	16.90	25.96
36.5−55	16.97	23.16	44.89	45.43
55−85	12.34	26.71	66.93	67.46
85−	6.52	25.13	114.25	148.19
all	100.00	100.00	34.17	72.27

Other Demographics

Variable	CPS Mean	CAMIP Mean
Family Size	2.36	2.65
Age of Household Head	46.80	46.18
Number of Kids	0.66	0.58
Urban	0.46	0.35
Rural	0.25	0.35
Suburban	0.29	0.30

The Choice Set.

To define a choice set, we need to classify vehicles into a list of distinct models and associate characteristics and quantities sold with those models. Roughly, our list of vehicles was determined by the sampling cells used to form the data GM provided to us (see Appendix C for details). This was detailed enough to allow us to construct a choice set of 203 vehicles (147 cars, 25 sport utility vehicles, 17 vans, and 14 pickup trucks)[9].

[8] The match is generally good, although the CPS questions are usually less ambiguously worded than the CAMIP questions. CAMIP does not ask about the education of the household head. There is a question about the education of the driver of the car, but that is hard to match to a question in the CPS.

[9] In most of the runs we used 218 vehicles. However in the later runs (reported below) we aggregated 15 very expensive vehicles (an average price of $74,000 and a composite market share of .3% of vehicles sold)

CAMIP contains information on the characteristics of the cars actually sold and on their transaction prices (most studies must make do with the characteristics of a "base" model and list prices). As our x_j we used the characteristics of the modal vehicle for each CAMIP vehicle sample cell (i.e. the combination of options that was most commonly purchased), and for our p_j we used the average price of the modal vehicle. Table 2 provides vehicle characteristics by type of vehicle and the definitions of the vehicle characteristics used throughout the paper. There were about 10.6 million vehicles sold in 1993 and they were sold at an average price of 18.5 thousand dollars. This gives total sales of about 196 billion dollars. The light truck market alone had sales of 81.2 billion dollars.

Table 3 provides the characteristics of a selected set of vehicles. Many of the interesting implications of our estimates are best evaluated at a vehicle level of aggregation. To give some idea of these implications without overwhelming the reader with details we display them only for the illustrative sample of sixteen vehicles in Table 3. These vehicles were selected because they all have sales that are large relative to the sales of vehicles of their type and because, between them, they cover the major types of vehicles sold.[10]

Characteristics of the Micro Data.

Table 4 provides the mean characteristics of vehicles chosen by the different demographic groups in the CAMIP sample. A number of interactions between observed household attributes and car characteristics stand out including; kids with minivan, income with price, rural with pickup and with allwheel drive, and age and nearly everything[11]. We used this table and others like it to suggest interactions to include in our specification for utility.

One of the very useful features of the CAMIP data is the presence of second choice information. Table 5 provides information on second choices for our "representative" sample of vehicles. The first column gives the first choice vehicle, while the second column gives the CAMIP sample size n. The next columns, in order, give: the modal second choice, the number of sampled consumers making that choice, the second choice with the second highest number of consumers, the fraction of n that chose one of the two second choices listed, and the number of different second choices made. For example the sample contains 199 purchasers of the Ford Escort. Their modal second choice was the Ford Tempo, while the second choice with the next highest number of consumers was the Ford Taurus. Together these two second choices accounted for 39, or 18%, of the consumers who chose the Escort. There were 51 other second choices registered among Escort purchasers.

There are a large number of different second choices for the same first choice car but the second choices are more concentrated for light trucks and for higher priced cars. Note also that the second choice is often produced by the same company as the first choice car; a fact

into one "super-luxury" model. Because of the very small shares of these luxury cars, this cut computational time considerably without changing the nature of the results.

[10]The list includes: ten cars (three of them luxury cars), a relatively low and a high priced minivan, a relatively low and a high priced jeep, a compact and a full sized pickup, and a full sized van.

[11]Older households tend to purchase larger (and therefore heavier) cars with both more safety features and more accessories. They also tend to stay away from sports utility vehicles and pickups.

Table 2: Vehicle Characteristics by Size/Type of Vehicle*

Vehicle Type	Total Q+	Mean Price+	Mean Pass	Mean HP	Mean Safe	Mean Acc	Mean MPG	Mean Allw	Mean PUPayl	Mean SUPayl	# of Vehicles
Car, pass = 2	57.5	28.5	2	7.1	2	4	20	0	0	0	6
Car, pass = 4	951.3	15.7	4	4.8	1	3	26	.004	0	0	35
Car, pass = 5	3829.7	17.5	5	4.7	1	3	23	.005	0	0	84
Car, pass ≥ 6	1374.1	21.5	6	4.8	1	4	19	0	0	0	22
Miniv	858.3	19.4	7	4.2	1	3	18	0	0	0	13
SU	1163.9	23.3	5	4.4	1	3	15	0.9	0	1.3	25
PU	2049.2	15.0	3	4.2	1	2	18	.003	2.0	0	14
Van	269.8	25.0	7	4.1	1	3	14	0	0	0	04
Total	10553.7	18.4	4.9	4.6	1	2.9	20	0.11	0.39	0.14	203

Variable Definitions for Vehicle Characteristics.

Q	US Sales and leases to consumers (from Polk)
P	Average price for modal car
HP	Horsepower/weight for engine of modal car ("acceleration")
Pass	Number of Passengers ("size")
MPG	City Miles per Gallon from EPA for modal engine/bodystyle
Acc	Number of power accessories of modal car (*e.g.* power windows, power doors)
Safe	Safety features: sum of ABS plus Airbags
Payl	Payload in thousands of pounds, for light trucks (from Wards and Automotive News)
Miniv	Dummy equal one if Minivan
SU	Dummy equal one if Sport Utility
PU	Dummy equal one if Pickup
Van	Dummy equal one if Full Size Van
Sport	Dummy equal one if Sport Car (as defined by consumer publications)
Allw	Dummy equal one if 4-wheel or all-wheel drive
PUPayl	PU × Payl
SUPayl	SU × Payl

*All means are sales weighted.

+ In thousands.

Table 3: Characteristics of Selected Vehicles

Model	Q*	Price*	Pass	HP	Safe	Acc	MPG	Allw	Miniv	SU	PU	Van	PUPayl	Spay
Geo Metro	83.7	7.8	4	3.0	0	0	46	0	0	0	0	0	0.00	0.00
Cavalier	184.8	11.5	5	4.4	1	2	23	0	0	0	0	0	0	0
Escort	207.7	11.5	5	3.6	0	1	25	0	0	0	0	0	0	0
Corolla	140.0	14.5	5	5.0	1	1	26	0	0	0	0	0	0	0
Sentra	134.0	11.8	4	4.7	0	2	29	0	0	0	0	0	0	0
Accord	321.2	17.3	5	4.5	1	4	22	0	0	0	0	0	0	0
Taurus	221.7	17.7	6	4.5	1	4	21	0	0	0	0	0	0	0
Legend	42.5	32.4	5	5.7	2	4	19	0	0	0	0	0	0	0
Seville	33.7	43.8	5	7.9	2	5	16	0	0	0	0	0	0	0
Lex LS400	21.9	51.3	5	6.5	2	5	18	0	0	0	0	0	0	0
Caravan	216.9	17.6	7	4.3	1	2	19	0	1	0	0	0	0	0
Quest	38.2	20.5	7	3.9	0	4	17	0	1	0	0	0	0	0
G Cherokee	160.3	25.9	5	5.4	2	4	15	1	0	1	0	0	0	1.15
Trooper	18.7	22.8	5	4.5	1	4	15	1	0	1	0	0	0	1.21
GMC FS PU	141.2	16.8	3	4.2	1	3	17	0	0	0	1	0	2.2	0
Toyota PU	175.1	13.8	3	4.4	0	0	23	0	0	0	1	0	1.64	0
Econovan	116.3	24.5	7	3.4	1	3	14	0	0	0	0	1	0	0

* In thousands.

Table 4: Vehicle Characteristics of Different Demographic Groups*

Group	Price	HP	Pass	Acc	Safe	Sport	MPG	Allw	Miniv	SU	Van	PU Payl	SU Payl
Age ≤ 30	16.6	4.7	4.5	2.6	.8	.20	22.0	.13	.03	.15	.001	.24	.18
Age $\in (30, 50]$	20.1	4.8	4.9	3.1	1.1	.15	20.4	.13	.08	.13	.009	.18	.18
Age > 50	22.4	4.9	5.1	3.4	1.3	.07	19.8	.06	.04	.04	.011	.19	.07
0 Kids	20.9	4.9	4.8	3.2	1.1	.14	20.4	.10	.03	.09	.006	.20	.12
1 Kids	19.2	4.7	4.8	3.0	1.0	.13	21.0	.12	.06	.11	.006	.20	.15
2+ Kids	20.1	4.6	5.3	3.1	1.0	.08	19.9	.12	.18	.13	.020	.16	.18
1 Fam	19.8	4.9	4.7	3.1	1.1	.20	21.2	.09	.01	.08	.003	.20	.12
2 Fam	21.5	4.9	4.9	3.3	1.2	.11	20.1	.10	.04	.09	.007	.20	.12
3+ Fam	19.7	4.7	5.0	3.1	1.0	.12	20.5	.11	.10	.12	.012	.19	.16
Urban	20.6	4.8	4.9	3.2	1.1	.13	20.7	.10	.05	.10	.009	.14	.14
Subrb	21.7	5.0	4.9	3.4	1.2	.15	20.3	.10	.06	.10	.006	.10	.14
Rural	19.2	4.7	4.9	3.0	1.0	.11	20.2	.12	.06	.11	.010	.31	.14
$y \leq 37$	16.6	4.6	4.8	2.6	.88	.12	21.9	.08	.04	.07	.008	.25	.08
$y \in (37, 55]$	18.5	4.7	4.9	3.0	1.0	.12	20.7	.10	.07	.10	.011	.24	.13
$y \in (55, 85]$	20.3	4.8	4.9	3.2	1.1	.14	20.0	.13	.07	.13	.009	.19	.17
$y > 85$	26.3	5.2	4.9	3.7	1.4	.14	19.1	.11	.05	.12	.006	.08	.17

*$a = age$ and $y = income$.

Table 5: Examples of Second Choices

Model	n_j	Modal 2nd Choice	# Choosing	Next 2nd Choice	(Modal + Next)/n	# Different Choices
Metro	188	Escort	22	Geo Storm	0.22	49
Cavalier	238	Escort	16	Lebaron	0.12	59
Escort	166	Tempo	16	Taurus	0.18	53
Corolla	250	Civic	42	Camry	0.33	55
Sentra	203	Corolla	34	Civic	0.31	60
Accord	223	Camry	58	Taurus	0.35	61
Taurus	147	Camry	18	Sable	0.22	45
Legend	119	Lex ES300	19	Lex SC300	0.24	40
Seville	243	Deville	38	Lin MK8	0.26	49
Lex LS400	148	Deville	33	Inf Q45	0.39	27
Caravan	166	Voyager	31	Aerostar	0.32	36
Quest	232	Caravan	50	Villager	0.43	31
G Cherokee	137	Explorer	75	Blazer	0.59	34
Trooper	137	Explorer	43	Rodeo	0.41	27
GMC FS PU	469	Chv FS PU	222	Ford FS PU	0.55	29
Toyota PU	113	Ford Ranger	29	Nissan PU	0.43	25
Econovan	90	Chv FS Van	20	Suburban	0.44	23

which argues strongly for pricing policies that maximize the joint profits of the firm across all the products it produces.

As expected, the second choice vehicles have characteristics that are similar to those of the first choices. The correlations of the different vehicle characteristics across the first and second choices of the households were all positive and highly significant (the correlations for price and Minivan were largest, about .7; those for MPG, Size and other type dummies were about .6; and the rest were between .3 and .5). Unfortunately, the surveyed consumers are not asked whether they would have purchased a vehicle at all if their first choice had not been available, so we cannot provide any descriptive evidence on how many consumers might substitute out of the new vehicle market altogether if their first choice was unavailable[12].

5 The Estimates of β^o and β^u

We begin with details of our specification. Recall that utility (equation 1) has interaction terms of the form $\sum_k \tilde{\beta}_{ik} x_{jk}$, where k indexes characteristics, i indexes household and j

[12]Some households listed a second choice that was broader than our first choice cells (*e.g.* a Ford pickup). The empirical analysis explicitly aggregates the respective cell probabilities for the second choices of these consumers.

indexes products. For all characteristics except price we assume that

$$\tilde{\beta}_{ik} = \bar{\beta}_k + \sum_r z_{ir}\beta^o_{kr} + \beta^u_k \nu_{ik}. \tag{13}$$

As in (2), the $\bar{\beta}$'s are subsumed in the product specific constants, δ, while the ν's are assumed to have independent (both across consumers and characteristics) standard normal distributions. Thus the β^u are the standard deviations of the contribution of unmeasured consumer attributes to the variance in the marginal utility for characteristics k. We let the descriptive tables and a number of preliminary runs guide our choice of which z_i to interact with the different x_j. Observed interactions were dropped from our early runs if we found them to be consistently unimportant.[13]

We assume the price coefficient to be a function of effective wealth, say W, and then model W in terms of household attributes. I.e. our price coefficient is $-e^{-W}$, so that its log is a decreasing function of

$$W_i \equiv \sum_r z_{ir}\beta^o_{w,r} + \beta^u_w \nu_{iw}. \tag{14}$$

Initially the $z_{i,r}$ included a constant, family size, a spline in income that was allowed to change derivatives at each of the quartiles of the CAMIP income distribution, and a lognormally distributed $\nu_{i,w}$ (for determinants of wealth not contained in our data). The data indicated only needed a change in the derivative of the income/price interaction in the spline at the 75th income percentile.

We have little *a priori* information on the outside option of not buying a car, so in early runs we let it be a linear function of all observed household attributes, a random normal disturbance, and the "logit" error. These runs indicated that the only attributes that mattered were income, family size, and, sometimes, the number of adults.

Table 6 (broken down into 6a and 6b) provides the estimates from our full model (the first result column), and compares them to those from more traditional models. Table 6a presents estimates of the β^o coefficients of interactions with observed household attributes, while Table 6b presents estimates of the β^u coefficients of interactions with unobserved attributes. There are three comparison models. The first two are obtained from our full specification but with $\beta^u = 0$, giving us a standard logit model with closed-form probabilities. This model has *both* choice specific intercepts and interactions between *observed* household attributes and vehicle characteristics (so we still have to use simulation to obtain predictions for aggregate shares; see Appendix A). The column labeled "Logit 1^{st}" provides the estimates obtained when by using only first choice data, while the column labeled "Logit 1^{st} & 2^{nd}" provides the estimates using both first and second choice data. The third comparison model sets $\beta^o = 0$ and so does not appear in Table 6a (just in 6b). This model is like BLP's model in that it has no observed consumer attributes.

There was one other comparison model we tried to estimate; our full model using only the first choice data (like the "Logit 1^{st}" results). However, even after *substantial* experimentation we had convergence problems with these runs and it eventually became clear that

[13]Our use of preliminary runs gives us some confidence that our results are reasonably robust to the inclusion of further interactions. However, it makes our standard errors suspect in the usual way.

Table 6a: Estimates of Interaction Terms, β^o

Vehicle Characteristic	Household Attribute	Full Model	Logit 1^{st}	Logit 1^{st} & 2^{nd}
Price	Constant	-2.18 (0.142)	0.092 (0.0001)	0.139 (0.0003)
Price	Income $\times$ (Income $<$ 75 percentile)	0.714 (0.044)	0.299 (0.002)	0.344 (0.001)
Price	Income $\times$ (Income $>$ 75 percentile)	1.17 (0.083)	0.466 (0.091)	0.603 (0.007)
Price	Family Size	-0.565 (0.010)	-0.144 (0.001)	-0.143 (0.006)
Miniv	Kids (kids have age $\leq$ 16)	1.973 (0.242)	0.765 (0.098)	0.771 (0.323)
Pass	Adults (adults have $age > 16$)	0.203 (0.095)	0.018 (0.0004)	-0.067 (0.009)
Pass	Family Size	.536 (0.052)	-0.055 (0.003)	-0.006 (0.0002)
Pass	Age (of household head)	0.019 (0.003)	0.002 (0.00001)	0.005 (0.00001)
HP	Age	-0.002 (0.001)	-0.010 (0.0004)	-0.012 (0.0001)
Acc	Age	0.0004 (0.001)	0.001 (0.00001)	-0.002 (0.0001)
Acc	Age^2	0.0001 (0.00001)	0.000 (0.00001)	0.000 (0.00001)
PUPayl	Age	0.0174 (0.002)	-0.003 (0.0001)	0.000 (0.00001)
PUPayl	Rural Dummy	1.075 (0.179)	.512 (0.005)	0.376 (0.008)
Safe	Age	0.013 (0.0006)	0.015 (0.001)	0.016 (0.0004)
SU	Age	-0.219 (0.010)	-0.043 (0.003)	-0.043 (0.004)
SU	Rural Dummy	0.332 (0.156)	0.403 (0.007)	-0.016 (0.002)
Allw	Rural Dummy	0.278 (0.247)	0.142 (0.005)	0.734 (0.246)
Outside Good	Total Income	5.151 (0.228)	-0.228 (0.096)	-0.305 (0.063)
Outside Good	Family Size	-0.007 (0.002)	0.532 (0.057)	-0.346 (0.004)
Outside Good	Adults	-0.428 (0.766)	0.851 (0.112)	1.953 (0.148)

Table 6b: Estimates of Interaction Terms, β^u

Parm Name	Full Model	$\beta^o \equiv 0$
Price	0.449	0.055
	(0.026)	(0.004)
HP	0.030	.183
	(0.016)	(0.020)
Pass	2.74	1.444
	(0.147)	(0.055)
Sport	0.002	2.763
	(0.0004)	(0.068)
Acc	0.554	0.515
	(0.078)	(0.055)
Safe	0.260	0.376
	(0.130)	(0.093)
MPG Y	0.488	0.430
	(0.018)	(0.017)
Allw	0.740	0.431
	(0.179)	(0.049)
Miniv	4.787	6.641
	(0.353)	(0.113)
SU	3.076	3.231
	(0.292)	(0.114)
Van	1.713	6.888
	(0.289)	(0.266)
PUPayl	2.160	4.301
	(0.092)	(0.210)
SUPayl	.356	0.015
	(0.072)	(0.013)
Chrysl	1.689	1.383
	(0.058)	(0.051)
Ford	0.915	1.410
	(0.072)	(0.051)
GM	1.885	1.844
	(0.057)	(0.105)
Honda	0.329	0.086
	(0.128)	(0.043)
Nissan	0.506	1.588
	(0.142)	(0.071)
Toyota	0.169	0.576
	(0.134)	(0.094)
Sm Asia*	1.467	2.155
	(0.068)	(0.022)
Europe*	0.454	1.883
	(0.084)	(0.034)
OutG	27.858	10.256
	(1.004)	(.506)

*We constrained the coefficients on the dummies for the different European firms to be the same, and we did the same for the smaller Asian producers.

very different parameter values could generate values of the objective function that were essentially the same as that of the minimum of that function. Apparently it is the availability of second choice data which enables us to focus in on a set of precise parameter estimates. Note that since we have only a single cross-section there is no variance in the choice set across observations [14]. In applications to other datasets, variation in the choice set over time might provide the information necessary to estimate the random coefficients.

The first panel of Table 6a shows that all three observed interactions with price are sharply estimated and have the expected sign (all else equal, larger families have lower "wealth"). Indeed almost all interactions in Table 6a had both an expected sign and were precisely estimated in *all three* specifications.[15] In addition to the price interactions this includes the interactions between Minivans and Kids (+), Age and Passengers (+), Age and Safety (+), HP and Age (-), SU and Age (-), and Rural and Pickup-payload (+).

The full model had only one parameter estimate that might be considered an anomaly (the positive age/Pickup-Payload interaction), while the first choice logit estimates had as its sole clear anomaly a negative interaction between number of passengers and family size (and the implication of this is ameliorated by the highly positive interactions between the minivan dummy and kids and between adults and passenger size). The second choice logits do a little worse, predicting negative interactions between family size and passengers and between rural and the sport utility dummy. The logits also have a pattern of outside good coefficients which is counter-intuitive. While estimates from our full model imply that households with more income and smaller families tend to have larger values for the outside option, the logits predict the opposite.[16] However, the outside good's coefficients are reduced form and hence more difficult to interpret.

On the whole the logits performed quite well in terms of producing sensible signs for coefficients, so the increased computational burden of the full model is not obviously justified by the pattern of estimated interactions between x and z. However, while the demographic interaction terms both seem to make sense and are sharply estimated, Table 6b indicates that they apparently *do not* explain the full pattern of substitution in the data. The estimated β^u coefficients are large and very precisely estimated. No matter how many observed interactions we allowed for, we needed numerous additional unobserved interactions to explain the data. Of course if we had richer consumer data we would hope to capture more with household observables, but the CAMIP data does have most of the household attributes generally available in large consumer choice data sets.

Looking at Table 6b more closely, nineteen out of twenty two coefficients are highly significant (eleven with t-values over ten) and two are marginally significant. Interestingly, there

[14] A referee noted that random coefficients models have been found unstable in many related cross-sectional contexts. For a review of random coefficients models see Rossi and McCulloch (2000), and the literature cited there.

[15] We did not present the breakdown of the variance in the estimated coefficients into portions caused by simulation and sampling error but typically somewhat less than half of this variance is due to simulation.

[16] Note that though our full model predicts a higher value of the outside good for higher income people, it also predicts a higher probability of purchasing a vehicle for higher income people, since the negative price interactions with income more than offsets the positive interactions with the outside good.

seems to be a wider dispersion of preferences for vehicles of U.S. than for those of Japanese companies. The model with no observed attributes has even more precisely estimated β^u coefficients (the $\beta^o \equiv 0$ column) as it has less other coefficients to estimate. Indeed the $\beta^o \equiv 0$ model has *all* β^u coefficients significant and several with t-values over fifty.

A clear pattern emerged when we compared the fit of the various models. The full model fit the (uncentered) moments derived from the interactions between observed consumer attributes and first choice car characteristics (equation 8) about as well as did the first and the second choice logits, while the model with no observed interactions could not fit these moments at all. On the other hand the model with no observed interactions fit the (uncentered) covariance of the first and second choice car characteristics (equation 12) about as well as did the full model, but the percentage errors in the first and second choice logits for these moments was typically five to ten times as large.

The logits, then, provide an adequate fit for the correlations between observed household and vehicle characteristics, but do very poorly in matching the characteristics of the first and second choice car. This might lead us to believe that the logits will predict the demographics of consumers well, but do a poor job of predicting substitution patterns. The no observed attribute model provides an adequate fit for the correlations of the characteristics of the first and second choice car, but has no prediction at all for the correlations between the observed household and the observed vehicle characteristics. Our full model (which nests all specifications) does about as well as the best of the alternatives in both these dimensions.

6 $\bar{\beta}$ and Substitution Patterns.

The only demand parameters left to estimate are the $\bar{\beta}$, the effects of the characteristics on the choice specific intercepts (the $\{\delta_j\}$). Recall that

$$\delta_j = p_j \bar{\beta}_p + \Sigma_{k \neq p}^{K} x_{jk} \bar{\beta}_k + \xi_j. \tag{15}$$

The problems encountered in estimating equation (15) are similar to the problems discussed in BLP in the context of estimating demand systems from product level data. In particular, consistent estimation of (15) requires instruments at least for the endogenous prices. Note that in contrast to our single 1993 cross-section, BLP had twenty annual cross-sections. Still their estimates that used only the demand system were too imprecise to be useful. This suggests that we also will have a precision problem, but this time only for a subset of the parameters, $\bar{\beta}$.

A number of additional sources of information could be used to increase the precision the estimated $\bar{\beta}$. First, we could mimic BLP. They assumed: [i] a functional form for marginal costs and [ii] that the equilibrium is Nash in prices. This generates a pricing equation that can be used in conjunction with the δ equation to increase the precision of our estimates of $\bar{\beta}$. In particular, if marginal costs are given by

$$mc_j = \sum_k x_{kj} \gamma_k + \omega_j, \tag{16}$$

where ω_j is an unobserved productivity term which is mean independent of x, and the γ are a set of parameters to be estimated, then the equilibrium assumption implies that price is equal to marginal cost plus a markup

$$p_j = \Sigma x_{kj}\gamma_k + b(x, p, \delta, \bar{\beta}_1, \beta^o, \beta^u)_j + \omega_j, \tag{17}$$

where the form of $b(x, p, \delta, \bar{\beta}_1, \beta^o, \beta^u)$ is determined by the demand-side parameters and the Nash pricing assumption.

With single product firms, the markup would be the (familiar) inverse of the semi-elasticity of demand with respect to price. Since we have multiproduct firms we must use the more complex formula for that case (see, for *e.g.* BLP).

The equilibrium markup in (17) is determined, in part, by ξ, ω, and p, and hence needs to be instrumented when that equation is estimated. In addition to x_j, the instruments we use are predictions of the markup:

$$\hat{b}_j \equiv b_j(x, \hat{p}, \hat{\delta}, \hat{\bar{\beta}}_1, \hat{\beta}^o, \hat{\beta}^u)_j \tag{18}$$

where $(\hat{\delta}, \hat{p})$ are obtained by projecting our estimate of δ and the observed p onto the $x's$, while $\hat{\bar{\beta}}_p$ is obtained from an initial IV estimate of the δ equation. So $\hat{b}_j$ is only a function of the x's and consistent parameter estimates [17].

Notice that this method of identifying $\bar{\beta}$ relies on our pricing assumption (though our estimates of (β^o, β^u) do not), and relies quite heavily on functional form restrictions (we do not observe multiple prices for a given vehicle). This suggests looking for other ways of identifying $\bar{\beta}$. Moreover since the equilibrium markups and price elasticities depend only on the coefficients estimated in the first stage analysis and on $\partial\delta_j/\partial p_j$, and equation (15) implies that $\partial\delta_j/\partial p_j = \bar{\beta}_p$, we can analyze all price change effects from the estimates of (δ, β^o, β^u) and any single restriction which identifies $\bar{\beta}_p$ [18]. Based on their experience, the staff at the *General Motors Corporation* suggested that the aggregate (market) price elasticity in the market for new vehicles was near one. An alternative estimate of $\bar{\beta}_p$ is then the value that sets the 1993 market elasticity equal to one.

When we use the δ equation (15) alone, the IV estimates of $\bar{\beta}$ are too imprecise to be of much use (our estimate of $\bar{\beta}_p$ had a standard error ten times the point estimate: 25 vs. 2.5). The IV estimate of $\bar{\beta}_p$ from the two equation model (which uses the δ equation *and* the pricing assumption) is -3.58 and has a standard error of .22. The estimate of $\bar{\beta}_p$ that "calibrates" to $GM's$ market elasticity of -1, is -11. We consider these two estimates as well as the estimate implicit in studies that ignore the correlation between the product-specific constant terms and price: $\bar{\beta}_p = 0$.

[17] Actually we iterate on this procedure several times, *i.e.* we use an initial simple IV estimate from the δ equation alone to produce our first estimate of $\hat{b}$. Then, we construct $\hat{b}$ and use it in a method of moments routine based on the orthogonality conditions from both equations. This produces a new estimate for $\bar{\beta}_p$, which is used to produce another estimate of $\hat{b}$ which was used in another method of moments routine. We continued in this way until convergence.

[18] Similarly, if we were interested in elasticities with respect to any other characteristic, say MPG or HP, we would require only the $\bar{\beta}$ associated with the characteristic of interest.

Table 7: Implications of Alternative Estimates of $\bar{\beta}_p$

Value of $\bar{\beta}_p$	0	-3.58	-11
Mean Semi-Elasticity	$-.75$	-3.94	-10.56
Total Market Elasticity	$-.2$	$-.4$	-1
Coefficients From Projecting Semi-Elasticities.			
Price	-0.016	-0.031	-0.063
	(0.003)	(0.006)	(0.014)
HP	0.023	-0.025	-0.122
	(0.025)	(0.044)	(0.102)
Pass	0.023	0.057	0.127
	(0.029)	(0.052)	(0.121)
Sport	-0.235	-0.230	-0.219
	(0.069)	(0.117)	(0.273)
Acc	-0.086	-0.066	-0.023
	(0.023)	(0.040)	(0.093)
Safe	-0.177	-0.137	-0.052
	(0.038)	(0.067)	(0.126)
MPG	0.010	-0.034	-0.126
	(0.007)	(0.013)	(0.029)
Allw	0.084	0.275	0.671
	(0.103)	(0.182)	(0.425)
Miniv	-0.174	-0.730	-1.882
	(0.099)	(0.174)	(0.406)
SU	-0.480	-0.923	-1.841
	(0.179)	(0.316)	(0.735)
Van	-0.339	-1.112	-2.714
	(0.154)	(0.272)	(0.633)
PUPayl	-0.173	-0.625	-1.562
	(0.050)	(0.088)	(0.204)
SUPayl	-0.107	-0.058	-0.400
	(0.101)	(0.144)	(0.416)

Firm dummies suppressed.

Table 7 examines the implications of these three estimates of $\bar{\beta}_p$. The first rows provide the implied average (across vehicles) price semi-elasticities and total market price elasticities. The rest of the table presents the coefficients obtained from the projection of the implied price semi-elasticities onto car characteristics.

Clearly the *level* of the price elasticities increase with the value of the estimate of $\bar{\beta}_p$. On the other hand the *pattern* of the elasticities seems fairly robust across our estimates of $\bar{\beta}_p$ and accords well with industry reports (especially to reports *circa* 1993). Semi-elasticities decrease in price and given price, vans (both mini and full sized), pickups, sport utilities and, to a lesser extent, sport cars, have noticeably smaller elasticities than other vehicles. This goes a long way in explaining reports of high markups to these vehicles.

We now come to the patterns of substitution across cars. The two types of substitution patterns we consider are; (i) substitution induced by price changes, and (ii) substitution induced by deleting vehicles from the choice set. The two sets of substitution patterns differ because when price increases only a selected sample of consumers that purchased the given vehicle substitute out of that vehicle (the more price-sensitive consumers), whereas when a vehicle is deleted from the choice set all of them must make an alternative choice. These substitution patterns were virtually independent of the estimates of $\bar{\beta}_p$ so we present only one set of results (with $\bar{\beta}_p = -3.58$).

Table 8a presents our model's predictions for the substitution patterns that would result from a small increase in price of the vehicle in the first column. The table provides the name of the vehicle chosen by the largest fraction of the substituting consumers, the price of that vehicle, and the fraction of those who substitute out of the first choice vehicle who move to that "best" substitute. It then provides the same information for the vehicle chosen by the second highest fraction of the substituting consumers. The last column of the table provides the fraction of the substituting consumers who substitute to the outside alternative. Thus the best (price) substitute for the Toyota Corolla is the Honda Civic and the second best is the Ford Escort. Together these two cars account for about 25% of those who substitute out of the Corolla when its price rises. About 5% of those who substitute out do not purchase a car at all.

The substitution patterns in table 8a make a lot of sense. Both substitutes tend to be the same type of vehicle as the vehicle whose price rose (minivans substitute to minivans, . . .). Among vehicles of the same type, the substitutes tend to be vehicles with similar prices and of similar size as the car whose price increased.

Table 8b compares best price substitutes from our model to those from our comparison models. It is clear that the intuitive features of the predictions of our model *are not* shared by the results from the logit models, but are, for the most part, shared by the results from the no observed attributes model. The first choice logit predicts the Dodge Caravan, a minivan, to be the "best substitute" for nine of the ten first choice cars, and predicts the Ford Econovan to be the best substitute for the tenth car (a 400 series, or "high end", Lexus). It also predicts the Dodge Caravan to be the best substitute for both pickups, both sport utility vehicles, and the full size van. The first and second choice logit has the Ford full sized pickup as the best substitute for *all ten* cars.

Apparently the observed characteristics of households do not capture enough of the variation in individual tastes to produce reasonable substitution patterns [19]. On the other hand the no observed attribute ($\beta^o \equiv 0$) model produces the same best substitutes as our full model in twelve out of the seventeen cases (though its substitute for the Escort, and to a lessor extent for the Metro, seem questionable). If our primary interest is in substitution patterns, allowing for interactions between unobserved consumer and product characteristics seems far more important than allowing for the interactions between the observed consumer and product characteristics in our data. Again, recall that our consumer level data contains most of the variables that are generally available in large micro data sets.

Because of our second choice data, we are able to compare the models' predictions for substitution patterns to the data. Table 9 provides the most popular second choice as predicted by the four models. These are the "best substitutes" when the good in the first-column is taken off the market. We also ranked the actual data on second choices and placed the data rank of the model's best substitute next to the name of the predicted substitute. Thus, if the Honda Accord were taken off the market, both our model and the $\beta^o = 0$ model predict that the biggest beneficiary would be the Toyota Camry, and the data indicate that the Camry is in fact the most popular second choice among Accord purchasers. Our full model predicts exactly the same best substitute as the data nine out of seventeen times, predicts one of the top three best substitutes fifteen out of seventeen times, and never picks a best substitute that the data ranks higher than tenth (out of over 200 possible models). The model with $\beta^o \equiv 0$ predicts the same best substitute as the data twelve out of seventeen times, but has two best substitutes which the data ranks above ten[20]. Meanwhile, the logit models (i.e. $\beta^u \equiv 0$) perform as poorly here as they did in Table 8b with the Ford Full Size Pickup being predicted as the best substitute for every car in all the logit specifications. Note also that the best price substitutes and the best second choices are different for about half the cars and one of the light trucks.

7 Prediction Exercises.

Having shown that the implications of our estimate are consistent with available information we move on to two prediction exercises. First, We evaluate the potential demand for new models; in particular we introduce "high-end" sport utility vehicles (SUV). Second, we use the system to evaluate a major production decision; shutting down the Oldsmobile division of General Motors. We ask what Oldsmobile purchasers would do were the cars they bought not available. These examples were chosen for their relevance. Several new sport utility vehicles were introduced in the late 1990's (an apparent response to the high markups being

[19] This might have been expected from the logits inability to fit the moments for the characteristics of the first and second choice cars. Note that it is in spite of our allowing for choice specific constant terms.

[20] The one set of substitutes that might be considered an anomaly are the predicted substitutes for the Legend. Our model predicts the much cheaper Civic, which is in fact the choice of a small though significant number of Legend buyers. The $\beta^o = 0$ model predicts the Lincoln Towncar, which is priced close to the Legend but in fact Legend consumers almost never indicate it as a second choice.

Table 8a: Price Substitutes for Selected Vehicles, Estimates from the Full Model

Vehicle	Price	Semi- -Elas	Best Sub	Price	% of Movers[a]	2^{nd} Best	Price	% of Movers[a]	% to Outside[b]
Metro	7.84	−1.77	Tercel	9.70	14.96	Festiva	7.41	10.57	17.96
Cavalier	11.46	−4.08	Escort	11.49	8.62	Tempo	10.78	6.80	6.81
Escort	11.49	−4.02	Tempo	10.78	8.21	Cavalier	11.49	7.29	6.56
Corolla	14.51	−3.92	Civic	14.00	8.08	Escort	11.49	7.91	5.00
Sentra	11.78	−3.79	Civic	14.00	13.36	Escort	11.49	4.70	6.55
Accord	17.25	−3.92	Camry	18.20	8.60	Civic	13.00	4.47	5.06
Taurus	17.65	−3.73	Accord	17.25	6.25	MerSab	18.66	6.09	3.97
Legend	32.42	−3.73	Accord	17.25	3.96	Camry	18.20	3.87	4.38
Seville	43.83	−3.16	Deville	34.40	10.12	El Dorado	35.74	8.04	5.57
Lex LS400	51.29	−3.43	MB 300	47.71	7.97	LinTnc	35.68	6.29	5.87
Caravan	17.56	−3.32	Voyager	17.59	35.11	Aerostar	18.13	10.19	5.20
Quest	20.55	−3.98	Aerostar	18.13	12.50	Caravan	17.56	10.38	5.48
G Cherokee	25.84	−3.06	Explorer	24.27	17.60	Cherokee	20.10	9.51	6.38
Trooper	22.78	−3.96	Explorer	24.27	17.53	G.Cherokee	25.85	8.50	5.42
GMC FS PU	16.76	−3.78	Chv FS PU	16.78	43.74	Ford FS PU	16.68	13.56	6.03
Toyota PU	13.77	−3.34	Ranger	11.74	20.53	Nissan PU	11.10	11.93	9.35
Econovan	24.54	−2.86	Chevy Van	25.96	12.90	Dodge Van	23.71	9.73	5.38

[a]Of those who substitute away from the given good in response to the price change, the fraction who substitute to this good.

[b]Of those who substitute away from the given good in response to the price change, the fraction who substitute to the outside good.

Table 8b: Price Substitutes for Selected Vehicles, A Comparison Among Models.

Vehicle	Full Model	Logit 1^{st}	Logit 1^{st} & 2^{nd}	Sigma Only
Metro	Tercel	Caravan	Ford FS PU	Civic
Cavalier	Escort	Caravan	Ford FS PU	Escort
Escort	Tempo	Caravan	Ford FS PU	Ranger
Corolla	Escort	Caravan	Ford FS PU	Civic
Sentra	Civic	Caravan	Ford FS PU	Civic
Accord	Camry	Caravan	Ford FS PU	Camry
Taurus	Accord	Caravan	Ford FS PU	Accord
Legend	Town Car	Caravan	Ford FS PU	LinTnc
Seville	Deville	Caravan	Ford FS PU	Deville
Lex LS400	MB 300	Econovan	Ford FS PU	Seville
Caravan	Voyager	Voyager	Voyager	Voyager
Quest	Aerostar	Caravan	Caravan	Aerostar
G Cherokee	Explorer	Caravan	Chv FS PU	Explorer
Trooper	Explorer	Caravan	Chv FS PU	Rodeo
GMC FS PU	Chv FS PU	Caravan	Chv FS PU	Chv FS PU
Toyota PU	Ranger	Caravan	Chv FS PU	Ranger
Econovan	Dodge Van	Caravan	Ford FS PU	Dodge Van

Table 9: Most Popular Second Choices, A Comparison Among Models and to the Data

Vehicle	Full Model	Rank	Logit 1^{st}	Rank	Logit 1^{st}&2^{nd}	Rank	$\beta^o \equiv 0$	Rank
Metro	Chevsto	2	Ford FS PU	≥ 25	Ford FS PU	≥ 25	Tercel	12
Cavalier	Sun Bird	3	Ford FS PU	≥ 25	Ford FS PU	≥ 25	Ford Escort	1
Escort	Tempo	1	Ford FS PU	≥ 25	Ford FS PU	≥ 25	Tempo	1
Corolla	Escort	6	Ford FS PU	≥ 25	Ford FS PU	≥ 25	Civic	1
Sentra	Civic	2	Ford FS PU	≥ 25	Ford FS PU	≥ 25	Civic	2
Accord	Camry	1	Ford FS PU	≥ 25	Ford FS PU	≥ 25	Camry	1
Taurus	Mer. Sable	2	Ford FS PU	≥ 25	Ford FS PU	≥ 25	Accord	4
Legend	Civic	10	Ford FS PU	≥ 25	Ford FS PU	≥ 25	LinTnc	$\geq$25
Seville	Deville	1	Ford FS PU	≥ 25	Ford FS PU	≥ 25	Deville	1
Lex LS400	MB 300	3	Ford FS PU	≥ 25	Ford FS PU	≥ 25	Devill2	1
Caravan	Voyager	1	Ford FS PU	≥ 25	Voyager	1	Voyager	1
Quest	Aerostar	7	Ford FS PU	≥ 25	Caravan	1	Caravan	1
G Cherokee	Explorer	1	Chv FS PU	$\geq$25	Chv FS PU	≥ 25	Explorer	1
Trooper	Explorer	1	Chv FS PU	22	Chv FS PU	22	Rodeo	2
GMC FS PU	Chv FS PU	1	Chv FS PU	1	Ford FS PU	2	Chv FS PU	1
Toyota PU	Ranger	1	Chv FS PU	4	Chv FS PU	4	Ranger	1
Econovan	Chevy Van	1	Ford FS PU	6	Ford FS PU	6	Chevy Van	1

earned on those vehicles in the period of our data; see Table 7), and GM announced its intention to close down its Oldsmobile division in 2000.

Two caveats are worth noting before going to the results. First, all the data used in our investigations is 1993 data. The market has changed since 1993 and those changes might well effect our estimates. Second, in the exercises done here we do not allow other actors in the market to respond to the change we are investigating. I.e. when we shut down the Oldsmobile division we do not allow for either a re-alignment of the prices of other products in response to the shutdown, or for the introduction of the new models that might follow such a shut down. Similarly when we introduce a new model we investigate demand responses under the twin assumptions that prices of other vehicles do not respond to the introduction of that model and that no further new vehicles are introduced.

It is not much more difficult to modify our procedure to find a set of prices that would be a Nash equilibrium to the situation we study. This would, however, require (i) estimates of costs as well as of demand functions and, (ii) an assumption on how prices are set. In the past when we have tried similar exercises we found that the impact of the price response to be "second" order in cases similar to the cases we investigate here, but to be central to the analysis of other issues [21]. On the other hand we have done very little which examines the

[21] These studies used product level data and BLP's methodology. Induced price effects were second order in our analysis of the response of demand to the increase in gas prices in the early 1970's which appears in the *A. E. R.*, 1993. However we found the price effects to be central in our analysis of voluntary export restraints which appears in the *A. E. R.*, 1999, and in unpublished analysis of particular mergers.

longer term responses of the other characteristics (other than price) of the vehicles marketed to changes in the environment.

New Models.

The two new models we introduce into the 1993 market are a new Mercedes and a new Toyota SUV. Both new models were introduced with all characteristics *but* price and the unobserved characteristic (i.e. ξ) set equal to the characteristics of the Ford Explorer. The explorer was the biggest selling sport utility vehicle in 1993.

Recall that ξ captures the effect of all the detailed characteristics that are omitted from our specification; we think of it as "unobserved quality". The ξ of the new Toyota SUV was set equal to the mean ξ of all Toyota cars marketed in that year and the price of that vehicle was obtained from a regression of price onto a large set of vehicle characteristics and company dummies. This latter regression had a very good fit, and using it allowed us to avoid using the explicit pricing and cost assumptions that would be needed to obtain price from a more complete model. The ξ and p of the new Mercedes SUV were set in the same way using the "low end" of the Mercedes vehicles marketed in 1993[22]. Both vehicles introduced are at the very upper end of the quality and price distributions of the SUV's offered in 1993; the Toyota SUV's price ($30,240) is $4,500 more than that most expensive SUV sold in 1993, and the Mercedes' price is $3,500 above that.

Table 10 summarizes results from introducing the Mercedes SUV. It did well capturing about a third of the market share of the Explorer. The total number of vehicles sold hardly changed at all with the introduction; the demand for the Mercedes SUV comes largely at the expense of other sports utility vehicles, and to a far lesser extent, from luxury cars. The Toyota SUV's introduction was somewhat less successful at our predicted price; its market share was only .05. To increase the Toyota SUV's market share to that of the Mercedes we found that Toyota would have had to cut a thousand dollars off the price of its entrant. Our top predicted losers from the introduction of the Toyota SUV were the same as those for the introduction of the Mercedes SUV, but when the Toyota was introduced the fall in the market share of luxury cars was much smaller. The Toyota Camry was the only non-luxury car which was in the top 15 of falls in sales, and it was in that list when either new SUV was introduced.

Discontinuing the Oldsmobile Division.

Table 11 provides the results from discontinuing the Oldsmobile division of *GM*. This is of interest because GM has in fact recently announced the phase-out of that division. In 1993 Oldsmobile had a market share of about 2.44% of the total number of vehicles purchased, while *GM*'s total share of vehicles purchased was 32.2% . When we drop the Oldsmobile

[22]The mean Mercedes quality and price were much higher than the quality and price of any SUV marketed at the time. So if we used the means of the Mercedes we would have been doing prediction way out of the range of the data which we used in our estimation (and probably also out of the range of the SUV eventually marketed by Mercedes).

Table 10: Introducing a Mercedes SUV.*

Model	Price	Old Share	New Share	New - Old Share
New Car	33.659	0.0000	0.0762	0.0762
Biggest Declines in Sales.				
Ford Explorer	24.2740	0.2518	0.2373	−0.0144
Jeep G Cherokee	25.8490	0.1475	0.1376	−0.010
Chevy S10 Blazer	22.6510	0.1106	0.1071	−0.0036
Toyota 4Runner	25.5480	0.0380	0.0347	−0.0033
Nissan Pathfinder	24.943	0.0397	0.0375	−0.0022
Luxury cars	**	.1610	.1565	−.0045
All Vehicles	*n.r.*	9.711	9.711	.000

* See the text for the characteristics of the new car.
** Cars priced above $30,000.

models from the choice set, the three vehicles which benefit the most are all family sized *GM* cars (Chevy Lumina, Buick Lesabre, and Pontiac Grandam). Still some of the Olds purchasers shift to high selling family sized cars produced by other companies; notably the Honda Acccord, Ford Taurus and the Toyota Camry. Overall 43% of Oldsmobile car purchaser substitute to a non-*GM* alternative, and *GM*'s market share falls to 31.1%. Of course the profit change to *GM* depends on the costs saved by discontinuing Oldsmobile and on the markups of the *GM* cars that the Olds purchasers substitute to (numbers which *GM* presumably has detailed information on).

8 Conclusion

We find that a carefully constructed characteristic based model can provide an approximation to demand patterns that rationalizes what we observe (both in our data and in industry publications), and provide realistic out of sample predictions. To do so the model needed to allow for unobserved consumer attributes which affect preferences for characteristics. Also to obtain reliable parameter estimates we needed to use use the second (as well as first) choice data in our single cross section. On the other hand the number of parameters we needed to estimate was small relative to the number of parameters that would be needed for a flexible demand system in product space, and our parameters were quite precisely estimated. Moreover our system can be used for predicting the demand for new products, as well as for exploring the implications of policy and environmental changes on the demand for the products that already exist.

Table 11: Discontinuing the Oldsmobile Division

	Old Share	New Share	New-Old Share
All Oldsmobiles	.237	0	-.237
All GM	3.126	3.016	-.110
All Cars	9.711	9.695	-.016
Non-Olds Share Changes.			
Chevy Lumina	0.1354	0.1548	0.0194
Buick LeSabre	0.1216	0.1336	0.0120
Pontiac Grand Am	0.1322	0.1441	0.0119
Honda Accord	0.2955	0.3039	0.0084
Ford Taurus	0.2040	0.2115	0.0075
Saturn SL	0.1465	0.1539	.0074
Toyota Camry	0.2343	0.2415	0.0072
Buick Century	0.0614	0.0683	0.0069
Pontiac Grand Prix	0.0517	0.0584	0.0067
Chevy Cavalier	0.1700	0.1767	0.0067
Pontiac Bonneville	0.0658	0.0721	0.0064

The original Oldsmobile models in the data (and their shares) are: Ciera (0.068), Cutlass Supreme (0.059), Olds 88 (0.050), Achieva (0.033), Olds 98 (0.019) and Bravada (0.008).

9 Appendix

A. Comparison to Maximum Likelihood.

The likelihood of the combined (product level and CAMIP) data sets conditional on; $\theta \equiv (\delta, \beta^o, \beta^k)$, the car characteristics (x), the distribution of consumer attributes ($\mathcal{P}_w$), and the model in equations (3), (4) and (5), is given by

$$\begin{aligned} Pr(Camip, s^N \mid x, \theta, \mathcal{P}_w) &= Pr(Camip \mid s^N, x, \theta, \mathcal{P}_w) Pr(s^N \mid x, \theta, \mathcal{P}_w) \\ &= [Pr(Camip \mid x, \theta, \mathcal{P}_w) + O_p(n/N)] Pr(s^N \mid x, \theta, \mathcal{P}_w) \\ &\approx Pr(Camip \mid x, \theta, \mathcal{P}_w) Pr(s^N \mid x, \theta, \mathcal{P}_w). \end{aligned} \tag{19}$$

That is, though the CAMIP sample and the product level data are not independent (the households in the CAMIP sample contribute to aggregate market shares), the joint likelihood of the two samples differ from the product of their marginal likelihoods by a factor which is of order n/N, and since this is $\approx .0003$ in our sample, it can be safely ignored.

The CAMIP component of (19) is the product (over vehicles) of the likelihoods that each of the n_j randomly sampled purchasers of vehicle j would have the attributes and the second choice observed in the data. Since our model does not condition on the vehicle purchased and then predict z_i and second choices, but rather, it conditions on consumer attributes and then predicts first and second choices, we need to use Bayes' rule to derive this term. Letting $\prod$ be the product operator we have

$$Pr(Camip \mid x, \theta, \mathcal{P}_w) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} Pr(y_i^2, z_i \mid y_i^1 = j, x, \theta) \tag{20}$$

$$= \prod_{j=1}^{J} \prod_{i=1}^{n_j} \frac{Pr(y_i^2 \mid z_i, y_i^1 = j, x, \theta) Pr(y_i^1 \mid z_i, x, \theta) Pr(z_i)}{Pr(y_i^1 \mid x, \theta)}. \tag{21}$$

As in the text, $Pr(y_i^2 \mid z_i, y_i^1, x, \theta)$, $Pr(y_i^1 \mid z_i, x, \theta)$ and $Pr(y_i^1 \mid x, \theta)$ can be derived from the model, and $Pr(z_i)$ is taken from the CPS.

We also need the likelihood of the observed aggregate shares (the other term in (19)). Those shares distribute as a multinomial with sample size equal to the number of households in the US (our N) and probabilities given by the shares predicted by our model. Recalling that $\{s_j^N\}_{j=1}^J$ are the market shares and $s_0^N = 1 - \sum_j s_j^N$, the log of that likelihood is

$$log[Pr(s^N \mid x, \theta)] = N \sum_{j=0}^{J} s_j^N log[s_j(\theta \mid x)] \tag{22}$$

plus a constant term, where $s_j(\theta \mid x)$ is taken from (5).

As noted in the text, maximizing a likelihood based on (21) and (22) would involve a search over about 250 parameters. Note however that (21) implies that the contribution of the CAMIP data to the log likelihood increases with n, while (22) implies that the contribution of the the market shares increases with N. Thus when $N^{-1} \approx 0$ the m.l.e. will choose a θ that comes very close to maximizing $\sum_{j=0}^{J} s_j^N log[s_j(\theta \mid x)]$. This expression is maximized when $s_j^N = s_j(\theta \mid x)$ for $j = 0, \ldots, J$. As noted in the text, for each β there is a unique δ, say $\delta^N(\beta)$, which satisfies these equations and we can solve for it quickly using the contraction mapping provided in BLP. We call the estimator that maximizes the likelihood of the CAMIP sample *conditional* on the restriction that $s^N = s(\delta, \beta)$, "near" m.l.e. It searches only over β (which, in our case only has about fifty parameters), and should have nearly the same asymptotic distribution as the unrestricted m.l.e.

The near m.l.e.'s solution for $\delta^N(\beta)$ insures that the denominator of (21) equals s_j^N (no matter β), so it maximizes

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} \Big(log[Pr(y_i^2 \mid z_i, y_i^1 = j, x, \beta, \delta^N(\beta))] + log[Pr(y_i^1 \mid z_i, x, \beta, \delta^N(\beta))] \Big). \tag{23}$$

For given $\beta, \delta^N(\beta)$, this is the likelihood of an *random* sample of vehicle purchases; i.e. the method of choosing $\delta^N(\beta)$ corrects for the fact that the sample is choice based.

In practice, the near MLE runs into two problems. First, as with GMM, it is not feasible to solve the integral defining δ^N exactly and so we must use simulation to derive $\delta^{ns,N}$ (see the discussion surrounding (7).) This introduces a non-linear simulation error into the likelihood in (23). However, for large N, the error in our estimate $\delta^{ns,N}$ will converge to zero as ns grows large, and since $ns \approx n$ in our sample, it is reasonable to consider limits as ns grows in proportion to n, in which case the simulation error in $\delta^{ns,N}$ does not effect the consistency of the near m.l.e. (its contribution to the variance in $\hat{\beta}$ is analyzed as is its contribution to the GMM estimator described in the text).

The problem which deterred us from using the near m.l.e. for our *general* model is that the probabilities in (23) conditional on $\delta^N(\beta)$ cannot be computed exactly, *e.g.*, the integral

$$Pr(y_i^1 \mid z_i, x, \beta, \delta^N(\beta)) = \int_\nu Pr(y_i^1 \mid z_i, \nu, x, \beta, \delta^N(\beta))\mathcal{P}_\nu(d\nu)$$

has no analytic form and thus has to be simulated. This was also true for the GMM procedure discussed in the text, but unlike in that case the simulation error here enters the objective function in (23) non-linearly (it is inside the log function). As a result the simulated near m.l.e. is not consistent unless there is a large number of simulation draws for each individual probability. Moreover many of our probabilities are very small (the average is only $\approx .005$) and the log function is very sensitive to measurement error near zero, so in practice we found that we needed a very large number of simulation draws per individual to obtain reasonable estimates of the needed probabilities. Since we have over thirty thousand individuals, even a moderate number of simulation draws per individual was computationally prohibitive.

However when ν has no effect, that is for the special case where $\beta^u \equiv 0$, we *can* evaluate (23) directly. Note that in this case, $\delta^{ns,N}$ still has simulation error; we still have to simulate over the CPS distribution of z to approximate the model's prediction for aggregate market shares. Thus we still have to correct the standard errors of the estimate for simulation error, but this can be done using techniques analogous to those discussed in the text. This explains why we report results for the near MLE estimates for those special cases with $\beta^u \equiv 0$, that is when unobserved individual attributes are not important, and GMM estimates for both our general case, and for the case where $\beta^o \equiv 0$, i.e. where observed consumer attributes are not important.

B. Variances of Parameter Estimates.

The variance-covariance of the parameters is determined by; (i) the variance-covariance of the first order conditions that define the estimator evaluated at the true value of the parameters, and (ii) the expectation of the derivative, with respect to β, of the first order conditions that define the estimator evaluated at β_0 (see (Hansen 1982) for the formula given these two matrices).

The variance in our moments when evaluated at θ_0 is generated by two sources of randomness

- sampling error in the CAMIP means (*e.g.* from the variance in $(n_j)^{-1}\Sigma_{i_j=1}^{n_j} z_{i_j}$),
- simulation error in our calculations of the model's predictions.

Since the simulation and sampling errors are independent of each other and it is the difference between the sample mean and our model's predictions that enter our objective function (see equations 8 and 12), the variance of the moment conditions can be expressed as the sum of the variances due to sampling and simulation errors. The variance due to sampling error can be consistently estimated by calculating the variance of the moment conditions at the estimate of the parameter values holding the simulation draws constant. The variance due to simulation error can be consistently estimated by simulating the sample moment at the estimate of β for many independent sets of ns simulation draws and calculating the variance across the calculated moment vectors[23].

[23]For each set of draws we have to solve the contraction mapping for the $\delta^{N,ns}(\hat{\beta})$ that corresponds to that set of draws and use that estimate of $\delta^{N,ns}(\hat{\beta})$ in the calculation of the moments that go into (8) and (12). This is to account for the fact that the simulation effects both the prediction of the micro moments given an estimate of $\delta(\beta_0)$ and the estimate $\delta_0(\beta_0)$, i.e. $\delta^{N,ns}(\beta_0)$, itself.

The derivative matrix can be consistently estimated by taking the derivative of the sample first order condition evaluated at the estimate of β, remembering that, since we use a two step estimator, that derivative is the sum of two terms: one accounting for the direct effect of β on the moments given the estimate of $\delta(\beta, \cdot)$, and one accounting for the effect of β on $\delta(\beta)$ (see, for example, Pakes and Olley (1995)).

C. The CAMIP Sample and The Choice Set.

The original 1993 CAMIP sample is very large (about 57,000 observations). We deleted observations with missing values for any of the consumer attributes we used, and were left with about 37,500 observations[24]. The ratio of sampled purchasers to vehicle sales, a number set by GM, tends to decrease slightly in sales. That is, GM over-samples the buyers of less popular vehicles (these tend to be higher priced vehicles), so the overall distribution of characteristics in the CAMIP sample is not quite representative of the attributes of vehicle buying households. Almost all (actually about 34,500) respondents also report their second choice vehicle, and the first choices of the 2877 individuals who had no missing data except for second choice data *are used* in the estimation[25].

GM also gave us sales data for a list of just over two hundred models compiled by the Polk company. Since the CAMIP sampling cells were finer than those in the Polk data, we aggregated them to the Polk level of aggregation. The data on sales plus leases to households were divided by the number of households in the U.S. to give us our market shares (since the CAMIP do not include any sales or leases to businesses, neither could we). We need to attach characteristics to these models.

Previous empirical studies of this sort (including our own) have largely relied on published data from *Automotive News* or similar publications for both the model classification and the characteristics of the cars classified. *Automotive News*, for example, gives the *base* model characteristics of cars together with the list price of those cars. In contrast, we would like to have a measure of the *typical* characteristics of vehicle models, together with the average transaction (as opposed to list) price, so we use the CAMIP data to construct both the characteristics and the transaction prices.

For each vehicle purchased, the CAMIP data give a very detailed list of vehicle characteristics and the transaction price of the car (including sales taxes but excluding trade-in allowances)[26]. As our x_j we used the characteristics of the modal vehicle for each CAMIP vehicle sample cell (i.e. the combination of options that was most commonly purchased), and for our p_j we used the average price of the modal vehicle. Car characteristics that were not in the CAMIP survey (such as exterior

[24] We treat the missing data as if they were randomly missing. Though there were a significant number of missing values for all of our variables, data on income, and to a lesser extent on age, were missing disproportionately. We did compare means of observed variables conditional on income being present to the means when income was absent, and there were some differences (most notably the average age of a household which did not report income was 46.2, while the average age of those who did was 52). Though there is room for a deeper analysis of the impact of this selection criteria, such an analysis is beyond the scope of this paper.

[25] About 800 of these individuals had second choices that were deleted because they were identical to their first choices.

[26] Some of the characteristics (make, model, body style, and engine type) are known from the vehicle identification number of the car, but most are self-reported by the consumer. We informally compared the transaction prices, which were self-reported, to industry publications on suggested transaction prices and the CAMIP prices look quite reasonable.

size or fuel efficiency) were obtained from industry and/or government publications[27].

Any list of vehicles and their characteristics obscures the optional equipment (and, in some cases, the range of body styles and engines) that are available to the consumer; i.e. some aggregation of the choice set is necessary. Without denying the compromises inherent in our procedure, we would like to emphasize the improvement that our data provide over earlier studies (our own and others) that use list prices of base model cars (or, worse, the average characteristics of cars together with the list price of the base model), and a much smaller number of models. Also, unlike many previous studies, we include light trucks – minivans, sport utility vehicles and pickup trucks – in our analysis. Light trucks in 1993 accounted for about 40% of sales, so it is hard to get a complete picture of demand patterns without them.

References

Anderson, S., A. DePalma, and F. Thisse (1992): *Discrete Choice Theory of Product Differentiation.* MIT Press, Cambridge MA.

Berry, S. (1994): "Estimating Discrete Choice Models of Product Differentiation," *RAND Journal of Economics*, 23(2), 242–262.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 60(4), 889–917.

Cosslett, S. (1981): "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*, 49(5), 1289–1316.

Das, S., G. S. Olley, and A. Pakes (1995): "The Market for TVs," Discussion paper, Yale Univeristy.

Domenich, T., and D. McFadden (1975): *Urban Travel Demand: A Behavioral Analysis.* North Holland, Amsterdam.

Hansen, L. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hausman, J., and D. Wise (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46, 403–426.

Heckman, J. J., and J. M. Snyder (1997): "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators," *RAND.*

[27]For fuel efficiency (miles per gallon of gasoline), we matched the engine of the modal vehicle to EPA test data. In some cases the Polk sales data is more aggregated than the CAMIP data and in this case we aggregate the CAMIP to the Polk model definitions by taking the best-selling car within the Polk vehicle definition.

Hotelling, H. (1929): "Stability in Competition," *Economic Journal*, 39, 41–57.

Ichimura, H., and T. S. Thompson (1998): "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution," *Journal of Econometrics*, 86(2), 269–95.

Imbens, G. W., and T. Lancaster (1994): "Combining Micro and Macro Data in Microeconometric Models," *Review of Economic Studies*, 61(4), 655–80.

Lancaster, K. (1971): *Consumer Demand: A New Approach*. Columbia University Press, New York.

Manski, C. F., and S. R. Lerman (1977): "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica*, 45(8), 1977–88.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers of Econometrics*, ed. by P. Zarembka. Academic Press, New York.

——— (1981): "Econometric Models of Probabilistic Choice," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski, and D. McFadden. MIT Press, Cambridge, MA.

McFadden, D., A. Talvitie, and Associates (1977): *Demand Model Estimation and Validation*. Institute of Transportation Studies, Berkeley CA.

Nevo, A. (2000): "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry," *RAND Journal ofEconomics*, 31(3), 395–421.

Pakes, A. (1986): "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 54, 755–784.

Pakes, A., S. Berry, and J. Levinsohn (1993): "Some Applications and Limitations of Recent Advances in Empirical Industrial Organization: Price Indexes and the Analysis of Environmental Change," *American Economic Review, Paper and Proceedings*, 83, 240–246.

Pakes, A., and S. Olley (1995): "A Limit Theorem for a Smooth Class of Semiparametric Estimators," *Journal of Econometrics*, 65(1), 295–332.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 54, 1027–1057.

Rossi, P., and R. McCulloch (2000): "Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," *Journal of Econometrics*, 99, 173–193.