The BLP Method of Demand Curve Estimation in Industrial Organization

14 December 2007

Eric Rasmusen

Abstract

This is an exposition of the BLP method of structural demand estimation using the random-coefficients logit model. It is forthcoming Japanese in Gendai Keizaigaku 1, mikuro-bunseki, edited by Isao Miura and Tohru Naito, Tokyo: Keiso shobo. The English version will stay posted on the web and most likely be revised later.

Dan R. and Catherine M. Dalton Professor, Department of Business Economics and Public Policy, Kelley School of Business, Indiana University, BU 456, 1309 E. 10th Street, Bloomington, Indiana, 47405- 1701. Office: (812) 855-9219. Fax: 812-855-3354. Erasmuse@indiana.edu. http://www.rasmusen.org. Copies of this paper can be found at http://www.rasmusen.org/papers/blp-rasmusen.pdf.

I thank Alberto Gaggero, Timothy Havel, and Fei Lu for their comments.

I. Introduction

I came to know Professor Moriki Hosoe when he led the project of translating the first edition of my book, *Games and Information*, into Japanese in 1990 and visited him at Kyushu University. He realized, as I did, that a great manybnuseful new tools had been developed for economists, and that explaining the tools to applied modellers would be extraordinarily useful because of the difficulty of reading the original expositions in journal articles. Here, my objective is similar: to try to explain a new technique, but in this case the statistical one of the "BLP Method" of econometric estimation, named after Berry, Levinsohn & Pakes (1995). I hope it will be a useful contribution to the festschrift in honor of Professor Hosoe. I, alas, do not speak Japanese, so I thank xxx for their translation of this paper. I know from editing the 1997 *Public Policy and Economic Analysis* with Professor Hosoe how time-consuming it can be to edit a collection of articles, especially with authors of mixed linguistic backgrounds, and so I thank Isao Miura and Tohru Naito for their editing of this volume.

The BLP Method is a way to estimate demand curves, a way that lends itself to testing theories of industrial organization. It combines a variety of new econometric techniques of the 1980's and 90's. Philosophically, it is in the style of structural modelling, in which empirical work starts with a rigorous theoretical model in which players maximize utility and profit functions, and everything, including the disturbance terms, has an economic interpretation, the style for which McFadden won the Nobel Prize. That is in contrast to the older, reduced-form, approach, in which the economist essentially looks for conditional correlations consistent with his theory. Both approaches remain useful. What we get with the structural approach is the assurance that we do have a self-consistent theory and the ability to test much finer hypotheses about economic behavior. What we lose is simplicity and robustness to specification error.

Other people have written explanations of the BLP Method. Nevo (2000) is justly famous for doing this, an interesting case of a commentary on a paper being a leading article in the literature itself. (Recall, though, Bertrand's 1883 comment on Cournot or Hicks's 1937 "Mr Keynes and the Classics.") The present paper will be something of a commentary on a commentary, because I will use Nevo's article as my foundation. I will use his

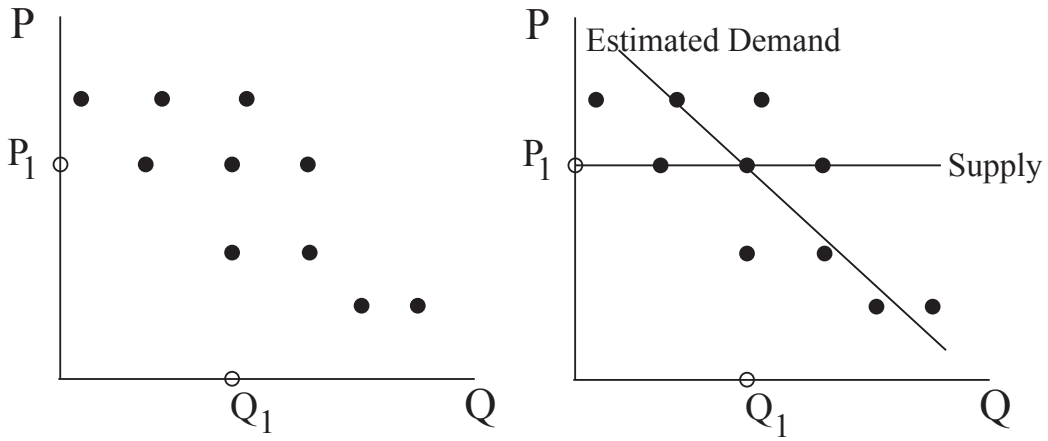
notation and equation numbering, and point out typos in his article as I go along. I hope that this paper, starting from the most basic problem of demand estimation, will be useful to those, who like myself, are trying to understand modern structural estimation.

II. The Problem

Suppose we are trying to estimate a market demand curve. Our data is the quantity sold of a product, q_t, and the price, p_t in twenty towns $t = 1, ...20$. Our theory is that demand is linear, with this equation:

$$q_t(p_t) = \alpha - \beta p_t + \varepsilon_t. \tag{1}$$

Let's start with an industry subject to price controls. A mad dictator sets the price in each town, changing it from town to town for entirely whimsical reasons. The result will be data that looks like Figure 1a. That data nicely fits our model in equation (1), as Figure 1b shows.



(a) Observation (b) Estimation

Figure 1: Supply and Demand with Price Controls

Next, suppose we do not have price controls. Instead, we have a situation of normal supply and demand. The problem is that now we might observe data like that in Figure 2a. Quantity rises with price; the demand curve seems to slope the wrong way, if we use ordinary least squares (OLS) as in Figure 2b.

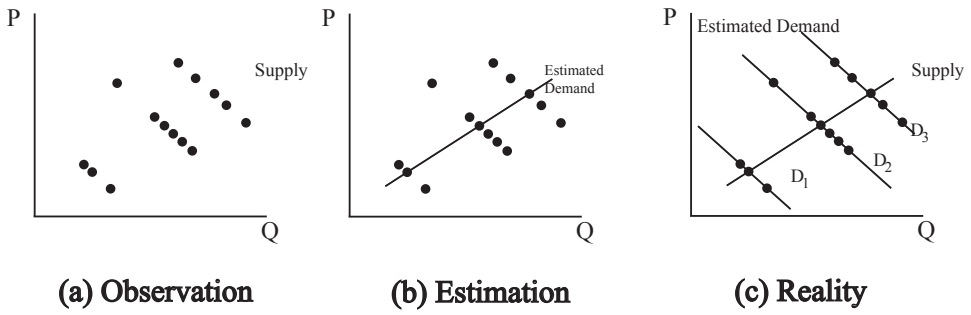


Figure 2: Supply and Demand without Price Controls

The solution to the paradox is shown in Figure 2c: OLS estimates the supply curve, not the demand curve. This is what Working (1927) pointed out.

It could be that the unobservable variables ε_t are what are shifting the demand curve in Figure 2. Or, it could be that it is some observable variable that we have left out of our theory. So perhaps we should add income, y_t:

$$q_t(p_t) = \alpha - \beta p_t + \gamma y_t + \varepsilon_t. \tag{2}$$

This will help make the estimation of β more accurate if p_t and y_t are correlated, but it does not help our basic problem of untangling supply shifts

from demand shifts. We need to include some observed variable that shifts supply.

Another approach would be to try to use a different kind of data. What if we could get data on individual purchases of product j? Our theory for the individual is different from our theory of the market. The demand curve looks much the same, except now we have a subscript for consumer i.

$$q_{it}(p_t) = \alpha_i - \beta_i p_t + \gamma_i y_{it} + \varepsilon_{it}. \tag{3}$$

But what about the supply curve? From the point of view of any one small consumer, the supply curve is flat. He has only a trivial influence on the quantity supplied and the market price, an influence we ignore in theories of perfect competition, where the consumer is a price-taker. Thus, we are back to something like Figure 1.

There is indeed one important difference. Now, all we've done is estimate the demand curve for one consumer. That is enough, if we are willing to simplify our theory to assume that all consumers have identical demand functions:

$$q_{it}(p_t) = \alpha - \beta p_t + \gamma y_{it} + \varepsilon_{it}. \tag{4}$$

This is not the same as assuming that all consumers have the same quantity demanded, since they will still differ in income y_{it} and unobserved influences, ε_{it}, but it does say that the effect of an increase in price on quantity is the same for all consumers. If we are willing to accept that, however, we can estimate the demand curve for one consumer, and we can use our estimate $\hat{\beta}$ for the market demand curve. Or, we can use data on n different consumers, to get more variance in income, and estimate $\hat{\beta}$ that way.

But we would not have to use the theory that consumers are identical. One alternative is to use the demand of one or n consumers anyway, arriving at the same estimate $\hat{\beta}$ that we would under the theory (1). The interpretation would be different, though–it would be that we have estimated the *average* value of β, and interpreting the standard errors would be harder, since they would be affected by the amount of heterogeneity in β_i as well as in ε_{it}. (The estimates would be unbiased, though, unlike the estimates I criticize in Rasmusen (1989a, b), since p_t is not under the control of the consumer.) Or, we could estimate the β_i for all n consumers and then average *the n estimates* to get a market β, as opposed to running one big regression.

Under either alternative, if we had reason to believe that the n consumers in our sample were not representative of the entire population, we would want to weight each consumer by the likely frequency of his type of preferences in the population.

Individual consumer data, however, is no panacea. For one thing, it is hard to get– especially since it is important to get a representative sample of the population. For another, its freedom from the endogeneity problem is deceptive. Recall that we assumed that each individual's demand had no effect on the market price. That is not literally correct, of course— every one of 900, 000 buyers of toothbrushes has some positive if trivial effect on market sales. If one of them decides not to buy a toothbrush, sales fall to 899,999. That effect is so small that the one consumer can ignore it, and the econometrician could not possibly estimate it given even a small amount of noise in the data. The problem is that changes in individuals' demand are unlikely to be statistically independent of each other. When the unobservable variable $\epsilon_{900000t}$ for consumer $i = 900,000$ is unusually negative, so too in all likelihood is the unobservable variable $\epsilon_{899999t}$ for consumer $i = 899,999$. Thus, they will both reduce their purchases at the same time, which will move the equilibrium to a new point on the supply curve, reducing the market price. Price is endogenous for the purposes of estimation, even though it is exogenous from the point of view of any one consumer.

So we are left with a big problem—identification— for demand estimation. The analyst needs to use instrumental variables, finding some variable that is correlated with the price but not with anything else in the demand equation, or else he must find a situation like our initial price control example where prices are exogenous.

In fact, even price controls might not lead to exogenous prices. A mad dictator is much more satisfactory, at least if he is truly mad. Suppose we have a political process setting the price controls, either a democratic one or a sane dictator who is making decisions with an eye to everything in the economy and public opinion too. When is politics going to result in a higher regulated price? Probably when demand is stronger and quantity is greater. If the supply curve would naturally slope up, both buyers and sellers will complain more if the demand curve shifts out and the regulated price does not change. Thus, even the regulated price will have a supply curve.

All of these problems arise in any method used to estimate a demand curve, whether it be the reduced-form methods just described or the BLP structural method. One thing about structural methods is that they force us to think more about the econometric problems. Your structural model will say what the demand disturbance is— unobserved variables that influence demand. If you must build a maximizing model of where regulated prices come from, you will realize that they might depend on those unobserved variables.

What does all this have to do with industrial organization? Isn't it just price theory, or even just consumer theory? Where are the firms? The reason this is so important in industrial organization is that price theory underlies it. Production starts because somebody demands something. An entrepreneur then discovers supply. That entrepreneur needs to organize the supply, and so we have the firm. Other entrepreneurs compete, and we have an industry. How consumers react to price changes is fundamental to this. Natural extensions to this problem bring in most of how firms behave. Demand for a product depends on the prices of all firms in the industry, and so we bring in the theory of oligopoly. Demand depends on product characteristics, and so we have monopolistic competition and location theory. Demand depends on consumer information, and so we have search theory, advertising, adverse selection, and moral hazard. As we have seen, estimating demand inevitably brings in supply. Or, if you like, you could think of starting with the problem of estimating the supply curve in a perfectly competitive industry, a problem that can be approached in the same way as we approach demand here.

III. The Structural Approach

Let us now start again, but with a structural approach. We will not begin with a demand curve this time. Instead, we will start with consumer utility functions. The standard approach in microeconomic theory is to start with the primitives of utility functions (or even preference orderings) and production functions and then see how maximizing choices of consumers and firms result in observed behavior. Or, in game theory terms, we begin with players, actions, information, and payoffs, and see what equilibrium strategies result from players choosing actions to try to maximize their payoffs given their information.

Suppose we are trying to estimate a demand elasticity— how quantity demanded responds to price. We have observations from 20 towns of cereal market data, the same 50 brands of cereal for each town (50 "products"), which makes a total of 1,000 data points. We also have data on 6 characteristics of each cereal brand and we have demographic data on how 4 consumer characteristics are distributed in the population in each town.

The older structural approach is to use a model of consumer preferences over the products and estimate 50 interconnected demand curves, as in the "linear expenditure model" of Stone (1954) (for the descendants of that approach, see the references in Hosken, Daniel, Daniel OBrien, David Scheffman & Michael Vita [2002] and Gould [2006]). A problem with that approach is that if there are 50 demand curves, and demand for each product depends on the prices of the others there are 2,500 parameters in the model, and we only have 1,000 data points with which to estimate them. This is the "curse of dimensionality." The symmetry of the cross-elasticities and the adding-up restrictions required by the common budget constraint reduce the number of free parameters, but estimation is still impractical. Even if we had more observations, we also would need sufficient variation in the data to sort out all the different interactions— we would need sufficient differences in the patterns of all 50 prices, for example. Thus, we will use a different approach, making consumer utility a function of product characteristics instead of the products themselves, and making the individual consumer's problem one of the probability of buying a particular product rather than how much to buy.

Each type of consumer will decide which product to buy, buying either one or zero units. We do not observe individual decisions, but we will model them so we can aggregate them to obtain the product market shares that we do observe. The fact that the frequency of different consumer types is different in different towns is the variance in the data that allows us to estimate how each of the 4 consumer characteristics affects demand. We can see if more sweet cereal is bought in Smallville, with its many children, than in Littleton, with its aging population.

At this point, we could decide to estimate the elasticity of demand for each product and all the cross-elasticities directly, but with 50 products that would require estimating 2,500 numbers. Instead, we will focus on the product characteristics. There are only 6 of these, and there would only be 6

even if there were 500 brands of cereal instead of 50. In effect, we will be estimating cross-elasticities between cereal characteristics, but if know those numbers, we can calculate the cross-elasticities between products by combining the characteristic effect parameter estimates with the characteristic level for each product.

The Consumer Decision

The utility of consumer i if he were to buy product j in town t is given by the following equation, denoted equation (1n) because it is equation (1) in Nevo (2000):

$$\begin{aligned} u_{ijt} &= \alpha_i(y_i - p_{jt}) + \mathbf{x_{jt}}\boldsymbol{\beta_i} + \xi_{jt} + \epsilon_{ijt} \\ & i = 1, ..., 400, \quad j = 1, ..., 50, \quad t = 1, ..., 20, \end{aligned} \tag{1n}$$

where y_i is the income of consumer i (which is unobserved and which we will assume does not vary across time), p_{jt} is the observed price of product j in town t, $\mathbf{x_{jt}}$ is a 6- dimensional vector of observed characteristics of product j in town t, ξ_{jt} (the letter "xi") is a disturbance scalar summarizing unobserved characteristics of product j in town t, and ϵ_{ijt} is the usual unobserved disturbance with mean zero. The parameters to be estimated are consumer i's marginal utility of income, α_i, and his marginal utility of product characteristics, the 6-vector $\boldsymbol{\beta}_i$. I have boldfaced the symbols for vectors and matrices above, and will do so throughout the article.

Consumer i also has the choice to not buy any product at all. We will model this outside option as buying "product 0" and normalize by setting the $j = 0$ parameters equal to zero (or, if you like, by assuming that it has a zero price and zero values of the characteristics):

$$u_{i0t} \equiv \alpha_i y_i + \epsilon_{i0t} \tag{5}$$

Equation (1n) is an indirect utility function, depending on income y_i and price p_{jt} as well as on the real variables x_{jt}, ξ_{jt}, and ϵ_{ijt}. It is easily derived from a quasilinear utility function, however, in which the consumer's utility is the utility from his consumption of one (or zero) of the 50 products, plus utility which is linear in his consumption of the outside good. Our consumers

are solving the problem,

$$\begin{aligned} \underset{q_0, q_1, \ldots q_{50}}{Maximize} \quad & \{q_0 + u_1(q_1) + u_2(q_2) + \cdots u_{50}(q_{50})\} \\ & \text{such that} \\ & (a) \quad \textstyle\sum_{j=0}^{50} \{q_j p_j\} \leq y \\ & (b) \quad \textstyle\sum_{j=0}^{50} \{q_j\} = 1 \end{aligned} \tag{6}$$

Constraint (a) is the usual budget constraint. Constraint (b) requires that only one unit of one of the 50 goods, or no units of any of them (so $q_0 = 1$ instead), be bought.

Quasilinear utility is linear, not concave, in income, so it lacks income effects.As a result, we do not need to observe consumer incomes and we can denote income as y_i, constant across towns, rather than y_{it}, varying across towns. If income effects are important, they can be modelled by indirect utility that is a function not of $(y_i - p_{jt})$ but of some concave function such as $log(y_{it} - p_{jt})$, as in BLP (1995). Consumer i's income does appear, but it does not change across towns, which may seem strange. That is because we are implicitly assuming that utility is linear in money, so different income levels would make no difference to choices.

A consumer's utility depends on a product's characteristics ($\mathbf{x_{jt}}$), directly on the product in a way that is the same for all consumers (ξ_{jt}), and on unobserved effects special to that consumer, product, and town (ϵ_{ijt}). Why is there no fixed consumer effect v_{it} analogous to the fixed product effect ξ_{jt}— an effect special to consumer i in town t, but independent of product j? We could add a v_{it} variable without harming the model, but it would have no effect, because when a consumer comes to compare the utilities from different products that variable would be the same for all of them and make no difference. The Smallsville consumer may be an unusually happy person, but that does not affect his choice between cereals.

Consumer characteristics do not appear directly in the utility function. They play a role later in the model, in determining $\boldsymbol{\beta}_i$, the marginal utility of product characteristics. Note that consumer income could appear there if we observe it varying across towns, as affecting the utility of different

characteristics even if we assume that utility is linear in money and the price elasticity does not depend on income.

We will assume that ϵ_{ijt} follows the Type I extreme-value distribution, which if it has mean zero and scale parameter one has the density and cumulative distribution

$$f(x) = e^x e^{e^{-x}} \qquad F(x) = e^{-e^{-x}}. \tag{7}$$

This is the limiting distribution of the maximum value of a series of draws of independent identically distributed random variables with support over the entire real line. The more draws, the bigger the expected maximum, which is why the mean and scale parameter can take various values. Figure 3 illustrates the density, which is not dissimilar to the normal distribution, except that it is slightly asymmetric and has thicker tails. Note that it has infinite support, so the unobserved effect ϵ_{ijt} always has some probability of outweighing product characteristics, price, and everything else. This distribution is standardly used in logit models because its cumulative distribution is related to the probability of x being larger than any other of a number of draws, which is like the random utility from one choice being higher than that from a number of other choices. This leads to a convenient formula for the probability that a consumer makes a particular choice, and thus for a product's market share, as we will see below.

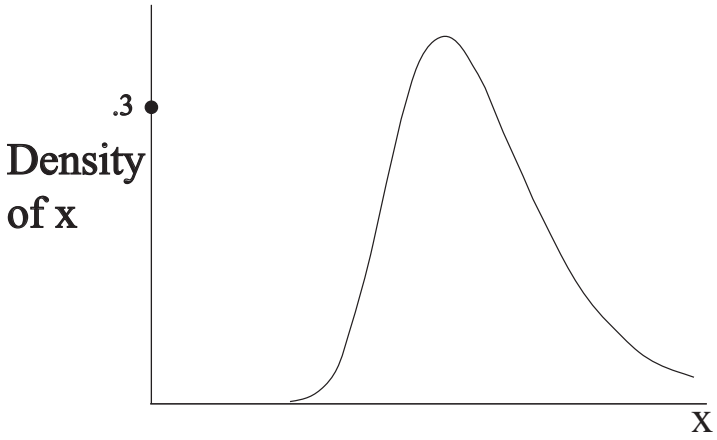


Figure 3: The Type I Extreme-Value Distribution

Consumer i buys product j in town t if it yields him the highest utility of any product or of not buying at all. What we observe, though, is not consumer i's decision, but the market share of product j. Also, though we do not observe $\boldsymbol{\beta}_i$, consumer i's marginal utility of product characteristics, we do observe a sample of observable characteristics. Even if we did observe his decision, we would still have to choose between regular logit and BLP's random- coefficients logit, depending on whether we assumed that every consumer had the same marginal utility of characteristics $\boldsymbol{\beta}$ or whether $\boldsymbol{\beta}_i$ depended on consumer characteristics.

Standard Multinomial Logit: Identical Consumers

One way to proceed would be to assume that all consumers are identical in their taste parameters; i.e., that $\alpha_i = \alpha$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, and that the ϵ_{ijt} disturbances are uncorrelated across i's. Then we have the standard multinomial logit model (multinomial because there are multiple choices, not just two). The utility function reduces to the following form, which is (1n) except that the parameters are no longer i-specific.

$$\begin{aligned} u_{ijt} &= \alpha(y_i - p_{jt}) + \mathbf{x_{jt}}\boldsymbol{\beta} + \xi_{jt} + \epsilon_{ijt} \\ & \\ & i = 1, ..., 400, \quad j = 1, ..., 50, \quad t = 1, ..., 20. \end{aligned} \tag{5n}$$

Now the coefficients are the same for all consumers, but incomes differ. We can aggregate by adding up the incomes, however, since the coefficient on each consumer has the same value, α. Thus we obtain an aggregate utility function,

$$u_{jt} = \alpha(y - p_{jt}) + \mathbf{x_{jt}}\boldsymbol{\beta} + \xi_{jt} + \epsilon_{jt}, \quad j = 1, ..., 50, \quad t = 1, ..., 20. \tag{8}$$

If we assume that ϵ_{jt} follows the Type I extreme-value distribution, then this is the multinomial logit model.

Since ϵ_{jt} follows the Type I extreme value distribution by assumption, the market share of product j under our utility function is

$$s_{jt} = \frac{e^{\mathbf{x_{jt}}\boldsymbol{\beta} - \alpha p_{jt} + \xi_{jt}}}{1 + \sum_{k=1}^{50} e^{\mathbf{x_{kt}}\boldsymbol{\beta} - \alpha p_{kt} + \xi_{kt}}} \tag{6n}$$

Equation (6n) is by no means obvious. The market share of product j is the probability that j has the highest utility, which occurs if ϵ_{jt} is high enough relative to the other disturbances. The probability that product 1 has a higher utility than the other 49 products and the outside good (which has a utility normalized to zero) is thus

$$\begin{aligned} &Prob(\alpha(y - p_{1t}) + \mathbf{x_{1t}}\boldsymbol{\beta} + \xi_{1t} + \epsilon_{1t} > \alpha(y - p_{2t}) + \mathbf{x_{2t}}\boldsymbol{\beta} + \xi_{2t} + \epsilon_{2t}) * \\ &Prob(\alpha(y - p_{1t}) + \mathbf{x_{1t}}\boldsymbol{\beta} + \xi_{1t} + \epsilon_{1t} > \alpha(y - p_{3t}) + \mathbf{x_{3t}}\boldsymbol{\beta} + \xi_{3t} + \epsilon_{3t}) * \cdots \\ &* Prob(\alpha(y - p_{1t}) + \mathbf{x_{1t}}\boldsymbol{\beta} + \xi_{1t} + \epsilon_{1t} > \alpha(y - p_{50t}) + \mathbf{x_{50t}}\boldsymbol{\beta} + \xi_{50t} + \epsilon_{50t}) * \\ &Prob(\alpha(y - p_{1t}) + \mathbf{x_{1t}}\boldsymbol{\beta} + \xi_{1t} + \epsilon_{1t} > \alpha y + \epsilon_{0t}) \end{aligned} \tag{9}$$

Substituting for the Type I extreme value distribution into equation (9) and solving this out yields, after much algebra, equation (6n). (For a step-by-step derivation, see chapter 3.1 and its appendix in Train (2003) on the Web.) Since αy appears on both sides of each inequality, it drops out. The 1 in equation (6n) appears because $e^0 = 1$ and the outside good adds 0 to utility.

Elasticities of Demand

To find the elasticity of demand, we need to calculate $\frac{\partial s_{jt}}{\partial p_{kt}}$ for products $k = 1, ..., 50$. It is helpful to rewrite equation (6n) by defining M_j as

$$M_j \equiv e^{\mathbf{x_{jt}}\boldsymbol{\beta} - \alpha p_{jt} + \xi_{jt}}. \tag{10}$$

so

$$s_{jt} = \frac{M_j}{1 + \sum_{k=1}^{50} M_k}. \tag{11}$$

Then

$$\frac{\partial s_{jt}}{\partial p_{kt}} = \frac{\frac{\partial M_j}{\partial p_{kt}}}{1 + \sum_{k=1}^{50} M_k} + \left(\frac{-M_j}{(1 + \sum_{k=1}^{50} M_k)^2} \right) \left(\frac{\partial M_k}{\partial p_{kt}} \right) \tag{12}$$

First, suppose $k \neq j$. Then

$$\begin{aligned} \tfrac{\partial s_{jt}}{\partial p_{kt}} &= \frac{0}{1 + \sum_{k=1}^{50} M_k} + \left(\frac{-M_j}{(1 + \sum_{k=1}^{50} M_k)^2} \right) (-\alpha M_k) \\ &= \alpha \left(\frac{M_j}{1 + \sum_{k=1}^{50} M_k} \right) \left(\frac{M_k}{1 + \sum_{k=1}^{50} M_k} \right) \\ &= \alpha s_{jt} s_{kt} \end{aligned} \tag{13}$$

Second, suppose $k = j$. Then

$$\begin{aligned} \tfrac{\partial s_{jt}}{\partial p_{jt}} &= \frac{-\alpha M_j}{1 + \sum_{k=1}^{50} M_k} + \left(\frac{-M_j}{(1 + \sum_{k=1}^{50} M_k)^2} \right) (-\alpha M_j) \\ &= -\alpha s_{jt} + \alpha s_{jt}^2 \\ &= -\alpha s_{jt}(1 - s_{jt}) \end{aligned} \tag{14}$$

We can now calculate the elasticity of the market share: the percentage change in the market share of product j when the price of product k goes up:

$$\eta_{jkt} \equiv \frac{\%\Delta s_{jt}}{\%\Delta p_{kt}} = \frac{\partial s_{jt}}{\partial p_{kt}} \cdot \frac{p_{kt}}{s_{jt}} = \begin{cases} -\alpha p_{jt}(1 - s_{jt}) & if\ j = k \\ \alpha p_{kt} s_{kt} & \text{otherwise.} \end{cases} \tag{15}$$

Why Multinomial Logit Is Unsatisfactory for Demand Estimation

The theoretical structure of the elasticities in equation (15) is unrealistic in two ways.

1. If market shares are small, as is frequently the case, then $\alpha(1 - s_{jt})$ is close to α, so that own-price elasticities are close to $-\alpha p_{jt}$. This says that if the price is lower, demand is less elastic, less responsive to price, which in turn implies that the seller will charge a higher markup on products with low marginal cost. There is no particular reason why we want to assume this, and in reality we often see that markups are higher on products with higher marginal cost, e.g. luxury cars compared to cheap cars.

2. The cross-price elasticity of product j with respect to the price of product k is $\alpha p_{kt} s_{kt}$, which only depends on features of product k— its price and market share. If product k raises its price, it loses customers equally to each other product. [1] This is a standard defect of multinomial logit, which McFadden's famous red-bus/blue-bus example illustrates. If you can choose among going to work by riding the red bus, the blue bus, or a bicycle, and the price of riding the red bus rises, are you equally likely to shift to the blue bus and to riding a bicycle? Of course not, but the multinomial logit model says that you will.

Another way to proceed is to use nested logit. In the red-bus/blue- bus example, you would first decide whether the blue bus or the red bus had the highest utility, and then decide whether the best bus's utility was greater than the bicycle's. In such a model, if the price of the red bus rose, you might switch to the blue bus, but you would not switch to the bicycle. A problem with nested logit, however, is that you need to use prior information to decide how to construct the nesting. In the case of automobiles, we might want to make the first choice to be between a large car and small car, but it is not clear that this makes more sense than making the first choice to be between a high-quality car and a low-quality car, especially if we are forcing all consumers to use the same nesting.

Random Coefficients Logit: Heterogeneous Consumers

An alternative to simple logit or nested logit is to assume that the

[1] An apparent third feature, that a high price of product k means it has higher cross-elasticities with every other product, is actually the same problem as problem (1).

parameters— the marginal utilities of the product characteristics— are different across consumers, and are determined by the consumer characteristics. Random coefficients is the name used for this approach, though it is a somewhat misleading name. The approach does not say that the consumer behaves randomly. Rather, each consumer has fixed coefficients in his utility function, but these coefficients are a function both of fixed parameters that multiply his observed characteristics and on unobservable characteristics that might as well be random. Thus, "random coefficients" really means "individual coefficients". We will denote the average values of the parameters α_i and $\boldsymbol{\beta}_i$ across consumers as α and $\boldsymbol{\beta}$, and assume the following specification:

$$\begin{aligned}\begin{pmatrix} \alpha_i \\ \boldsymbol{\beta}_i \end{pmatrix} &= \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\Pi}\mathbf{D_i} + \boldsymbol{\Sigma}\boldsymbol{\nu_i} \\ &= \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Pi}_\alpha \\ \boldsymbol{\Pi}_{\boldsymbol{\beta}} \end{pmatrix} \mathbf{D_i} + \begin{pmatrix} \boldsymbol{\Sigma}_\alpha \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\nu}_{i\alpha}, & \boldsymbol{\nu}_{i\boldsymbol{\beta}} \end{pmatrix}\end{aligned} \tag{2n}$$

where $\mathbf{D_i}$ is a 4×1 vector of consumer i's observable characteristics, $\boldsymbol{\nu}_i$ is a 7×1 vector of the effect of consumer i's unobservable characteristics on his α_i and $\boldsymbol{\beta}_i$ parameters; $\boldsymbol{\Pi}$ is a 7×4 matrix of how parameters (the α_i and the 6 elements of $\boldsymbol{\beta}_i$) depend on consumer observables, $\boldsymbol{\Sigma}$ is a 7×7 matrix of how those 7 parameters depend on the unobservables, and $(\boldsymbol{\nu}_{i\alpha}, \quad \boldsymbol{\nu}_{\mathbf{i}\beta})$, ($\boldsymbol{\Pi}_\alpha, \boldsymbol{\Pi}_\beta$) and ($\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta$) just split each vector or matrix into two parts.

We will denote the distributions of $\mathbf{D}$ and $\boldsymbol{\nu}$ by $\mathbf{P}^*_{\mathbf{D}}(\mathbf{D})$ and $\mathbf{P}^*_{\boldsymbol{\nu}}(\boldsymbol{\nu})$. Since we'll be estimating the distribution of the consumer characteristics $\mathbf{D}$, you will see the notation $\hat{\mathbf{P}}^*_{\mathbf{D}}(\mathbf{D})$ show up too. We will assume that $\mathbf{P}^*_{\boldsymbol{\nu}}(\boldsymbol{\nu})$ is multivariate normal.[2]

Utility in the Random Coefficients Logit Model

[2] Nevo variously uses $\mathbf{P}^*_{\mathbf{D}}(\mathbf{D})$, $\hat{\mathbf{P}}^*_{\mathbf{D}}(\mathbf{D})$, $\mathbf{P}^*_{\boldsymbol{\nu}}(\boldsymbol{\nu})$, and $\hat{\mathbf{P}}^*_{\boldsymbol{\nu}}(\boldsymbol{\nu})$ in his exposition. I have tried to follow him, but I may simply have misunderstood what he is doing.

Equation (1n) becomes[3]

$$\begin{aligned}
u_{ijt} &= \alpha_i(y_i - p_{jt}) + \mathbf{x_{jt}}\boldsymbol{\beta_i} + \xi_{jt} + \epsilon_{ijt} \\
&= \alpha_i y_i - (\alpha + \boldsymbol{\Pi}_\alpha \mathbf{D_i} + \boldsymbol{\Sigma}_\alpha \boldsymbol{\nu}_{i\alpha}) p_{jt} + \mathbf{x_{jt}}(\boldsymbol{\beta} + \boldsymbol{\Pi}_\beta \mathbf{D_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{\nu}_{\mathbf{i}\boldsymbol{\beta}}) + \xi_{jt} + \epsilon_{ijt} \\
&= \alpha_i y_i + (-\alpha p_{jt} + \mathbf{x_{jt}}\boldsymbol{\beta} + \xi_{jt}) - (\boldsymbol{\Pi}_\alpha \mathbf{D_i} + \boldsymbol{\Sigma}_\alpha \boldsymbol{\nu}_{i\alpha}) p_{jt} + \mathbf{x_{jt}}\left(\boldsymbol{\Pi}_{\boldsymbol{\beta}} \mathbf{D_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{\nu}_{\mathbf{i}\boldsymbol{\beta}}\right) + \epsilon_{ijt} \\
&= \alpha_i y_i + (-\alpha p_{jt} + \mathbf{x_{jt}}\boldsymbol{\beta} + \xi_{jt}) + (-p_{jt}, \mathbf{x_{jt}}) (\boldsymbol{\Pi} \mathbf{D_i} + \boldsymbol{\Sigma} \boldsymbol{\nu_i}) + \epsilon_{ijt} \\
&= \alpha_i y_i + \delta_{jt} + \mu_{ijt} + \epsilon_{ijt} \\
&\quad j = 1, ..., 50, \quad t = 1, ..., 20.
\end{aligned} \tag{16}$$

What I have done above is to reorganize the terms to separate them into four parts. First, there is the utility from income, $\alpha_i y_i$. This plays no part in the consumer's choice, so it will drop out.

Second, there is the "mean utility", δ_{jt}, which is the component of utility from a consumer's choice of product j that is the same across all consumers.

$$\delta_{jt} \equiv -\alpha p_{jt} + \mathbf{x_{jt}}\boldsymbol{\beta} + \xi_{jt} \tag{17}$$

Third and fourth, there is a heteroskedastic disturbance, μ_{ijt}, and a homoskedastic i.i.d. disturbance, ϵ_{ijt}.

$$\mu_{ijt} \equiv (-p_{jt}, \mathbf{x_{jt}}) (\boldsymbol{\Pi} \mathbf{D_i} + \boldsymbol{\Sigma} \boldsymbol{\nu_i}) \tag{18}$$

Market Shares and Elasticities in the BLP Model

If we use the Type I extreme value distribution for ϵ_{ijt}, then the market share of product j for a consumer of type i turns out to be

$$s_{ijt} = \frac{e^{\delta_{jt} + \mu_{ijt}}}{1 + \sum_{k=1}^{50} e^{\delta_{kt} + \mu_{ikt}}}. \tag{19}$$

[3]There is a typographical error in the Nevo paper on p. 520 in equation (3n): u instead of μ in each line.

Recall that we denote the distributions of $\mathbf{D}$ and $\boldsymbol{\nu}$ by $\mathbf{P}^*_{\mathbf{D}}(\mathbf{D})$ and $P^*_{\boldsymbol{\nu}}(\boldsymbol{\nu})$. Since we will be estimating the distribution of the consumer characteristics $\mathbf{D}$, you will see the notation $\hat{\mathbf{P}}^*_{\mathbf{D}}(\mathbf{D})$ show up too.

The overall market share of product j in town t is found by integrating the market shares picked by each consumer in equation (19) across the individual types, weighting each type by its probability in the population:

$$\begin{aligned} s_{jt} &= \int_{\boldsymbol{\nu}} \int_{\mathbf{D}} s_{ijt} \mathbf{d}\hat{\mathbf{P}}^*_{\mathbf{D}}(\mathbf{D}) \mathbf{d}\mathbf{P}^*_{\boldsymbol{\nu}}(\boldsymbol{\nu}) \\ &= \int_{\boldsymbol{\nu}} \int_{\mathbf{D}} \left[\frac{e^{\delta_{jt}+\mu_{ijt}}}{1+\sum_{k=1}^{50} e^{\delta_{kt}+\mu_{ikt}}} \right] \mathbf{d}\hat{\mathbf{P}}^*_{\mathbf{D}}(\mathbf{D}) \mathbf{d}\mathbf{P}^*_{\boldsymbol{\nu}}(\boldsymbol{\nu}) \end{aligned} \tag{20}$$

Equation (20) adds up the market shares of different types i of consumers based on how common that type i is in town t.

The price elasticity of the market share of product j with respect to the price of product k is

$$\begin{aligned} \eta_{jkt} &\equiv \frac{\partial s_{jt}}{\partial p_{kt}} \cdot \frac{p_{kt}}{s_{jt}} \\ &= \begin{cases} -\dfrac{p_{jt}}{s_{jt}} \displaystyle\int_{\boldsymbol{\nu}} \int_{\mathbf{D}} \alpha_i s_{ijt}(1-s_{ijt}) \mathbf{d}\hat{\mathbf{P}}^*_{\mathbf{D}}(\mathbf{D}) \mathbf{d}\mathbf{P}^*_{\boldsymbol{\nu}}(\boldsymbol{\nu}) & if\ j=k \\ \dfrac{p_{kt}}{s_{jt}} \displaystyle\int_{\boldsymbol{\nu}} \int_{\mathbf{D}} \alpha_i s_{ijt} s_{ikt} \mathbf{d}\hat{\mathbf{P}}^*_{\mathbf{D}}(\mathbf{D}) \mathbf{d}\mathbf{P}^*_{\boldsymbol{\nu}}(\boldsymbol{\nu}) & \text{otherwise.} \end{cases} \end{aligned} \tag{21}$$

This is harder to estimate than the ordinary logit model, whose analog of equation (20) is equation (6n), repeated below.

$$s_{jt} = \frac{e^{\mathbf{x_{jt}}\boldsymbol{\beta} - \alpha p_{jt} + \xi_{jt}}}{1+\sum_{k=1}^{50} e^{\mathbf{x_{kt}}\boldsymbol{\beta} - \alpha p_{kt} + \xi_{kt}}} \tag{6n}$$

The difficulty comes from the integrals in (20). Usually these need to be calculated by simulation, starting with our real-world knowledge of the distribution of consumer types i in a given town t and the characteristics of product j, and combining that with estimates of how different consumer types value

different product characteristics. This suggests that we might begin with an initial set of parameter estimates, calculate what market shares that generates for each town, see how those match the observed market shares, and then pick a new set of parameter estimates to try to get a closer match.

What is special about random-coefficients logit is not that it allows for interactions between product and consumer characteristics, but that it does so in a structural model. One non-structural approach would have been to use ordinary least squares to estimate the following equation, including product dummies to account for the ξ_{jt} product fixed effects.

$$s_{jt} = \mathbf{x_{jt}}\boldsymbol{\beta} - \alpha p_{jt} + \xi_{jt} \tag{22}$$

Like simple logit, the method of equation (22) implies that the market share depends on the product characteristics and prices but not on any interaction between those things and consumer characteristics. We can incorporate consumer characteristics by creating new variables in a vector d_t that represents the mean value of the 4 consumer characteristics in town t, and then interacting the 4 consumer variables in $\mathbf{d_t}$ with the 6 product variables in $\mathbf{x_t}$ to create a 1×24 variable $\mathbf{w_t}$. Then we could use least squares with product dummies to estimate

$$s_{jt} = \mathbf{x_{jt}}\boldsymbol{\beta} - \alpha p_{jt} + \mathbf{d_t}\boldsymbol{\theta_1} + \mathbf{w_t}\boldsymbol{\theta_w} + \xi_{jt} \tag{23}$$

where $\boldsymbol{\theta}_1$ is the coefficient vector for the 4 direct effects of the consumer variables and $\boldsymbol{\theta}_w$ is the coefficient vector for the 24 interactions between consumer and product variables.

Equation (23) adds market shares directly from the 4 consumer characteristics and less directly via the 24 consumer-product characteristic interactions. Thus, it has some of the flexibility of the BLP model. It is not the result of a consistent theoretical model. Even aside from whether rational consumer behavior could result in a reduced form like (**??**) or (23), those equations do not take account of the relationships between the market shares of the different products— the sum of all the predicted market shares should not add up to more than one, for example. The random-coefficient logit model avoids this kind of internal contradition.

Before going on to estimate the random coefficients model, however, recall that regardless of how we specify the utility function— logit, nested logit,

random-coefficients logit, or something else— we face the basic simultaneity problem of estimating demand and supply functions. Market shares depend on prices and disturbance terms, but prices will also depend on the disturbance terms. If demand is unusually strong for a product because of some unobservable variable, the price of that product will be higher too. This calls for some kind of instrumental variables estimation. What we will use here is the generalized method of moments.

The Generalized Method of Moments

The generalized method of moments (GMM) of Hansen (1982) will combine aspects of overidentified instrumental variables, generalized least squares, and a nonlinear specification in our demand setting. It helps to separate these things out. We will go one step at a time, before circling back to the BLP method's use of GMM.

Suppose we want to estimate

$$y = x_1\beta_1 + x_2\beta_2 + \epsilon, \tag{24}$$

where we observe y, x_1, and x_2, but not ϵ, though we know that ϵ has a mean of zero. We assume that the x's and the unobservable disturbances $\boldsymbol{\epsilon}$ are uncorrelated, which is the definition of the x's being exogenous (if we included a constant term, it would be uncorrelated with $\boldsymbol{\epsilon}$ too— another way to include the assumption that $\boldsymbol{\epsilon}$ has a mean of zero). This lack of correlation makes up the two "moment conditions" that we can write as

$$M_1 : E(\mathbf{x_1'}\boldsymbol{\epsilon}) = 0, \qquad M_2 : E(\mathbf{x_2'}\boldsymbol{\epsilon}) = 0. \tag{25}$$

or

$$E(M_1) = 0, \qquad E(M_2) = 0. \tag{26}$$

Note that if there are T observations, $\mathbf{x_1}$ is a $T \times 1$ vector, but $M_1 = \mathbf{x_1'}\boldsymbol{\epsilon}$ is 1×1.

In the generalized method of moments, we choose $\hat{\beta}$ to make a weighted sum of squares of the sample moment expressions (which are the M's that equal zero in the moment condition) as small as possible. Let's put aside the question of what weights to use for now. The sum of squares of the moment expressions is

$$(M_1 \;\; M_2)'(M_1 \;\; M_2) \tag{27}$$

Think of M_1 as a random variable, since it incorporates those T random variables $\boldsymbol{\epsilon}$ in the T observations. The expected value of M_1 is zero, by assumption, but in our sample its realization might be positive or negative, because its variance is not zero. The matrix $(M_1 \;\; M_2)$ is 2×1, so the sum of squared moment expressons is 1×1.

Another way to write the problem is to choose $\hat{\boldsymbol{\beta}}$ to minimize $\mathbf{M}'\mathbf{M}$. If $\mathbf{M} = \mathbf{X}'\boldsymbol{\epsilon}$, we will find

$$\hat{\boldsymbol{\beta}} \;\; = \overset{argmin}{\boldsymbol{\beta}} \;\; \hat{\boldsymbol{\epsilon}}'\mathbf{X}\mathbf{X}'\hat{\boldsymbol{\epsilon}} \tag{28}$$

Thus, we maximize the function $f(\hat{\boldsymbol{\beta}})$:

$$\begin{aligned} f(\hat{\boldsymbol{\beta}}) &= \hat{\boldsymbol{\epsilon}}'\mathbf{X}\mathbf{X}'\hat{\boldsymbol{\epsilon}} \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{X}\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad (29) \\ &= \mathbf{y}'\mathbf{X}\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

We can differentiate equation (29) with respect to $\hat{\boldsymbol{\beta}}$ to get the first order condition,

$$\begin{aligned} f'(\hat{\boldsymbol{\beta}}) &= -\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{X}'\mathbf{X} + 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{X} \\ &= -2\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{y} + 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{X} \qquad (30) \\ &= 2\mathbf{X}'\mathbf{X}(-\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}) = 0 \end{aligned}$$

in which case

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{31}$$

and we have the OLS estimator.

We might also know that the x's and disturbances are *independent*:

$$E(\epsilon|x_1, x_2) = 0. \tag{32}$$

This is different from being uncorrelated. For example, suppose ϵ's value is either -1 or $+1$ with equal probability, and x_1 equals 0 if $\epsilon = -1$ and 3 or -3

with equal probability if $\epsilon = 1$. Then $E(\boldsymbol{\epsilon}'\mathbf{x_1}) = 0$, but $E(\epsilon|x_1 = 0) = -1$. Moreover, various functions of $\mathbf{x_1}$ are correlated with ϵ even though $\mathbf{x_1}$ is not. $E(\boldsymbol{\epsilon}'(\mathbf{x_1^2})) = .5(-1)(0) + .5(.5 * 3^2 * 1) + .5 * (-3)^2 * 1 = 4.5,$. Thus, ϵ and x_1 are uncorrelated but not independent.

We want to use all available information, for efficient estimation, so we would like to use that independence information. It will turn out to be useful information if the variance depends on $\mathbf{X}$, though not otherwise. (That "not otherwise" will show up as that if the variance does not depend on $\mathbf{X}$, then the moment variance covariance matrix entry is zero, and if we know ex ante the information that it equals zero then we should use that information.)

Independence gives us lots of other potential moment conditions. Here are a couple:

$$E((\mathbf{x_1^2})'\boldsymbol{\epsilon}) = E(M_3) = 0, \qquad E((\mathbf{x_2} * \mathbf{x_1})'\boldsymbol{\epsilon}) = E(M_4) = 0. \tag{33}$$

Some of these conditions are more reliable than others. So we'd like to weight them when we use them. Since M_3 and M_4 are random variables, they have variances. So let's weight them by the the inverse of their variances—more precisely, by the inverse of their variance-covariance matrix, since they have cross-correlations. Call the variance-covariance matrix of all the moment conditions $\boldsymbol{\Phi}(\mathbf{M})$. We can estimate that matrix consistently by running a preliminary consistent regression such as OLS and making use of the residuals.

The GMM estimator uses the inverse of the variance-covariance matrix of the moment conditions, $\boldsymbol{\Phi}(\mathbf{M})^{-\mathbf{1}}$, to weight them in the estimation, a weighting scheme that has been shown to be optimal (see Hansen [1982]). We minimize the weighted square of the moment conditions by choice of the parameters $\hat{\boldsymbol{\beta}}$.

$$(M_1 \;\; M_2 \;\; M_3 \;\; M_4)'(\boldsymbol{\Phi}(\mathbf{M})^{-\mathbf{1}})(M_1 \;\; M_2 \;\; M_3 \;\; M_4) \tag{34}$$

The weighting matrix is crucial. OLS uses the most obviously useful information. We can throw in lots and lots of other moment conditions using the independence assumption, but they will contain less and less new information. Adding extra information is always good in itself, but in finite

samples, the new information, the result of random chance, could well cause more harm than good. In such a case, we wouldn't want to weight the less important moment conditions, which might have higher variance, as much as the basic exogeneity ones. Consider the moment condition M_5:

$$E((\mathbf{x_2^3} * \mathbf{x_1^5})'\boldsymbol{\epsilon}) = E(M_5) = 0. \tag{35}$$

That moment condition doesn't add a lot of information, and it could have a big variance not reflected in the consistent estimate of $\mathbf{\Phi}(\mathbf{M})$ that we happen to obtain from our finite sample.

We have now gotten something like generalized least squares, GLS, from the generalized method of moments. I did not demonstrate it, but $\mathbf{\Phi}(\mathbf{M})$ will turn out to be an estimate of the variance covariance matrix of $\boldsymbol{\epsilon}$. It is not the same as other estimates used in GLS, because it depends on exactly which moment conditions are used, but it is consistent. We have a correction for heteroskedasticity, which is something we need for estimation of the BLP problem. Notice that this means that GMM can be useful even though:

(a) This is a linear estimation problem, not nonlinear.

(b) No explanatory variables are endogenous, so this is not an instrumental variables problem.

There are other ways to correct for heteroskedasticity, but the GMM estimator shows its true strength when used as a form of instrumental variables estimation.[4] Suppose that one of our basic moment conditions fails. $E(\mathbf{x_2}\boldsymbol{\epsilon}) \neq 0$, because $\mathbf{x_2}$ is endogenous, and we have lost our moment conditions M_2 and M_4. What we need is a new basic moment condition that will enable us to estimate β_2— that is, we need an instrument correlated with $\mathbf{x_2}$ but not with $\boldsymbol{\epsilon}$. Suppose we do have a number of such conditions, a set of variables $\mathbf{z_1}$ and $\mathbf{z_2}$. We can use our old conditions M_1 and M_3, and we'll add a couple others too, ending up with this set:

$$E(\mathbf{x_1}\boldsymbol{\epsilon}) = 0 \quad E((\mathbf{x_1^2})'\boldsymbol{\epsilon}) = 0, \quad E(\mathbf{z_1}\boldsymbol{\epsilon}) = 0 \quad E(\mathbf{z_2}\boldsymbol{\epsilon}) = 0 \quad E((\mathbf{z_1} * \mathbf{x_1})'\boldsymbol{\epsilon}) = 0 \quad E((\mathbf{z_1} * \mathbf{z_2})'\boldsymbol{\epsilon}) = 0. \tag{36}$$

We will abbreviate these six moment conditions as

$$E(\mathbf{Z}'\boldsymbol{\epsilon}) = E(\mathbf{M}) = 0, \tag{37}$$

[4]The following exposition is based on Hall(1996).

where the matrix $\mathbf{Z}$ includes separate columns for the original variable $\mathbf{x_1}$, the simple instruments $\mathbf{z_1}$ and $\mathbf{z_2}$, and the interaction instruments, $\mathbf{z_1} * \mathbf{x_1}$ and $\mathbf{z_1} * \mathbf{z_2}$.

Let's suppose also, for the moment, that we have the ex ante information that the disturbances are independent of each other and $\mathbf{Z}$, so there is no heteroskedasticity. Then we can derive the weighting matrix thus (noting that a variance is calculated using the deviation from the mean, which here is zero):

$$\mathbf{\Phi}(\mathbf{M}) = \mathbf{Var}(\mathbf{M}) = \mathbf{Var}(\mathbf{Z}'\boldsymbol{\epsilon}) = E(\mathbf{Z}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{Z}) - E(\mathbf{Z}'\boldsymbol{\epsilon})E(\boldsymbol{\epsilon}'\mathbf{Z}) = E(\mathbf{Z}'(\mathbf{I}\sigma^{\mathbf{2}})\mathbf{Z}) - \mathbf{0^2} = \sigma^{\mathbf{2}}\mathbf{Z}'\mathbf{Z}. \tag{38}$$

The GMM estimator uses that variance-covariance matrix in the weighting matrix that puts different weight on the six moment conditions. It solves the problem of choosing the parameters $\hat{\boldsymbol{\beta}}_{\mathbf{2SLS}}$ to minimize

$$\begin{aligned} f(\hat{\boldsymbol{\beta}}_{\mathbf{2SLS}}) &= \hat{\boldsymbol{\epsilon}}'_{\mathbf{2SLS}}\mathbf{Z}(\sigma^{\mathbf{2}}\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\hat{\boldsymbol{\epsilon}}_{\mathbf{2SLS}} \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{2SLS}})'\mathbf{Z}(\sigma^{\mathbf{2}}\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{2SLS}}) \end{aligned} \tag{39}$$

We can differentiate this with respect to $\hat{\boldsymbol{\beta}}_{\mathbf{2SLS}}$ to get the first order condition

$$f'(\hat{\boldsymbol{\beta}}_{\mathbf{2SLS}}) = -\mathbf{X}'\mathbf{Z}(\sigma^{\mathbf{2}}\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{2SLS}}) = 0, \tag{40}$$

which solves to

$$\hat{\boldsymbol{\beta}}_{\mathbf{2SLS}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\mathbf{X}]^{-\mathbf{1}}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\mathbf{y} \tag{41}$$

This estimator is both the GMM estimator and the 2SLS (two-stage least squares) estimator. They are equivalent when the disturbances are independently distributed, though if there were heteroskedasticity they would become different because GMM would use the weighting matrix $(\mathbf{\Phi}(\mathbf{M}))^{-\mathbf{1}}$, which would not be the same as $(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}$. 2SLS could be improved upon with heteroskedasticity corrections, however, in the same way that OLS can be improved.

Notice that this is the 2SLS estimator, rather than the simpler instrumental variables (IV) estimator that is computed by calculating IV directly:

$$\hat{\boldsymbol{\beta}}_{\mathbf{IV}} = (\mathbf{X}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\mathbf{y} \tag{42}$$

Two-stage least squares and IV are the same if the number of instruments is the same as the number of parameters to be estimated, but otherwise the formula in equation (42) cannot be used, because when $\mathbf{X}$ is $T \times J$ and $\mathbf{Z}$ is $T \times K$, $\mathbf{X'Z}$ is $J \times K$, which is not square and cannot be inverted . What 2SLS is doing differently from IV is projecting $\mathbf{X}$ onto $\mathbf{Z}$ with the projection matrix, $\mathbf{Z}(\mathbf{Z'Z})^{-\mathbf{1}}\mathbf{Z'}$, to generate a square matrix that can be inverted. GMM does something similar, but with $\mathbf{\Phi}(\mathbf{M})$ instead of $\mathbf{Z}(\mathbf{Z'Z})^{-\mathbf{1}}\mathbf{Z'}$.

We have so far solved for $\hat{\boldsymbol{\beta}}$ analytically, but that is not an essential part of GMM. The parameters $\boldsymbol{\beta}$ might enter the problem nonlinearly, in which case minimizing the moment expression could be done using some kind of search algorithm. For example, suppose our theory is that

$$y = x_1^{\beta_1} + \beta_1\beta_2 x_2 + \epsilon, \tag{43}$$

and our moment conditions are

$$E(\mathbf{x_1'}\boldsymbol{\epsilon}) = M_1 = 0, \quad E(\mathbf{x_2'}\boldsymbol{\epsilon}) = M_2 = 0 \quad E(\mathbf{x1} * \mathbf{x_2'}\boldsymbol{\epsilon}) = M_3 = 0. \tag{44}$$

We could then search over values of β_1 and β_2 to minimize the moment expression,

$$(\mathbf{y} - \mathbf{x_1}^{\beta_1} + \beta_\mathbf{1} * \beta_\mathbf{2} * \mathbf{x_2})'\mathbf{M}(\mathbf{\Phi}(\mathbf{M}))^{-\mathbf{1}}\mathbf{M'}(\mathbf{y} - \mathbf{x_1}^{\beta_1} + \beta_\mathbf{1} * \beta_\mathbf{2} * \mathbf{x_2}), \tag{45}$$

where we would have to also estimate $\mathbf{\Phi}(\mathbf{M})$ during some part of the search.

Combining Logit and GMM

Now let us return to random coefficients logit. If our assumption on the population is that[5]

$$\mathbf{E}(\mathbf{Z_m}\boldsymbol{\omega}(\boldsymbol{\theta}^*)) = 0, \quad m = 1, \ldots, M, \tag{8n}$$

then the GMM estimator is

$$\hat{\boldsymbol{\theta}} \overset{argmin}{=} \boldsymbol{\theta} \; \boldsymbol{\omega}(\boldsymbol{\theta})'\mathbf{Z}\mathbf{\Phi}^{-\mathbf{1}}\mathbf{Z'}\boldsymbol{\omega}(\boldsymbol{\theta}), \tag{9n}$$

[5]I think there is a typo in Nevo here, on page 531, and z_m' should replace Z_m in equation (8n).

where $\mathbf{\Phi}$ is a consistent estimator of $\mathbf{E}(\mathbf{Z}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{Z})$. The instrument matrix $\mathbf{Z}$ consists both of single- variable instruments and multi-variable instruments consisting of squares and interactions of the single variables.

As in ordinary least squares,but unlike in maximum likelihood, does not require us to know the distribution of the disturbances. In our demand estimation, though, we will still have to use the assumption that the ϵ_{ijt} follow the extreme value distribution, because we need it to calculate the market shares aggregated across consumer types, whether by plain logit or random-coefficients logit.

In more detail, here is how the estimation proceeds. First, the modeller must set up his theoretical model, which at a minimum means choosing lists of observation units (towns, or months, or town/months, etc.), products, product characteristics, and consumer characteristics. He may also choose to include additional variables besides the prices which are common to all consumers: advertising, hot weather, a time trend, and so forth (in which case the parameter α would become a vector of parameters). Following the Appendix to Nevo (2000), denote by $\mathbf{X_1}$ the matrix of explanatory variables common to all consumers, including not only the prices, advertising, and so forth but a dummy variable for each product. These are separated out because they are the only variables that enter into the simpler, linear part of the estimation needed to obtain the average parameter values, $(\alpha, \boldsymbol{\beta})$. Denote by $\mathbf{X_2}$ the matrix of directly observed explanatory variables that enter into the nonlinear part of the estimation, the estimation of $(\mathbf{\Pi}, \mathbf{\Sigma})$ for the random coefficients (i.e. the individuals' deviations from the average coefficients): the prices and any other common variables which affect consumers with different characteristics differently, and the product characteristics, but not product dummies. Also, neither $\mathbf{X_1}$ nor $\mathbf{X_2}$ includes the consumer characteristics, since they enter the estimation separately, at the point where the market shares are estimated by aggregating across consumers. The modeller must also choose instruments for price and perhaps for other common variables. All this data must be collected, as must the market shares for each product and a distribution or actual sample of consumers with their characteristics. The estimation then take the following steps (the labelling of the steps is adapted from Nevo [2000], Appendix, p. 1).

(-1) Select arbitrary values for $\boldsymbol{\delta}$ and $(\mathbf{\Pi}, \mathbf{\Sigma})$ (for step 1) and for $(\alpha, \boldsymbol{\beta})$ (for

step (3)) as starting points. Recall that $\boldsymbol{\delta}$ from (17) is the vector of the mean utility from each of the products, that $(\boldsymbol{\Pi}, \boldsymbol{\Sigma})$ is the matrix of parameters showing how observed and unobserved consumer characteristics and product characteristics interact to generate utility, and that $(\alpha, \boldsymbol{\beta})$ is the average value of the parameters across consumers.

(0) Draw random values for $(\boldsymbol{\nu_i}, \mathbf{D_i})$ for $i = 1, ...n_s$ from the distributions $\mathbf{P}^*_{\boldsymbol{\nu}}(\boldsymbol{\nu})$ and $\hat{\mathbf{P}}^*_{\mathbf{D}}(\mathbf{D})$ for a sample of size n_s, where the bigger you pick n_s the more accurate your estimate will be.

(1) Using the starting values and the random values, and using the assumption that the ϵ_{ijt} follow the extreme-value distribution, approximate the integral for market share that results from aggregating across i by

$$\begin{aligned} s_{jt} &= \left(\frac{1}{n_s}\right) \sum_{i=1}^{n_s} s_{ijt} \\ &= \left(\frac{1}{n_s}\right) \sum_{i=1}^{n_s} \left[\frac{e^{[\delta_{jt} + \Sigma_{k=1}^{6} x_{jt}^k (\sigma_k \nu_i^k + \pi_{k1} D_{i1} + \cdots + \pi_{k4} D_{i4})]}}{1 + \sum_{m=1}^{50} e^{[\delta_{mt} + \sum_{k=1}^{6} x_{mt}^k (\sigma_k \nu_i^k + \pi_{k1} D_{i1} + \cdots + \pi_{k4} D_{i4})]}} \right], \end{aligned} \tag{11n}$$

where $(\nu_i^1, \ldots, \nu_i^6)$ and $(D_{i1}, \ldots, D_{i4})$ for $i = 1, \ldots n_s$ are those random draws from the previous step.

Thus, in step (1) we obtain predicted market shares for given values of the individual consumer parameters $(\boldsymbol{\Pi}, \boldsymbol{\Sigma})$ and for given values of the mean utilities $\boldsymbol{\delta}$.

(2) Use the following contraction mapping, which, a bit surprisingly, converges. Keeping $(\boldsymbol{\Pi}, \boldsymbol{\Sigma})$ fixed at their starting points, find values of $\boldsymbol{\delta}$ by the following iterative process.

$$\boldsymbol{\delta}_{\cdot t}^{h+1} = \boldsymbol{\delta}_{\cdot t}^{h} + (\mathbf{ln}(\mathbf{S_{\cdot t}}) - \mathbf{ln}(\mathbf{s_{\cdot t}})), \tag{12n}$$

where $\mathbf{S_{\cdot t}}$ is the observed market share. and $\mathbf{s_{\cdot t}}$ is the predicted market share from step (1) that uses $\boldsymbol{\delta}_{\cdot t}^{h+1}$ as its starting point. Start with the arbitrary $\boldsymbol{\delta}^0$ of step (-1).

If the observed and predicted market shares are equal, then $\boldsymbol{\delta}_{\cdot t}^{h+1} = \boldsymbol{\delta}_{\cdot t}^{h}$ and the series has converged. In practice, keep iterating until $(\mathbf{ln}(\mathbf{S_{\cdot t}}) - \mathbf{ln}(\mathbf{s_{\cdot t}}))$ is small enough for you to be satisfied with its accuracy.

Thus, in step (2) we come out with values for $\boldsymbol{\delta}$.

(3) Figure out the value of the moment expression, using the starting values for $(\alpha, \boldsymbol{\beta})$ from step (0) and the $\boldsymbol{\delta}$ estimate from step (2).

(3a) Calculate the error term $\boldsymbol{\omega}_{jt}$.

$$\boldsymbol{\omega}_{jt} \quad = \boldsymbol{\delta}_{jt} - (\alpha \mathbf{p_{jt}} + \mathbf{x_{jt}} \boldsymbol{\beta}) \tag{13n}$$

(3b) Calculate the value of the moment expression,

$$\boldsymbol{\omega}' \mathbf{Z} \boldsymbol{\Phi}^{-1} \mathbf{Z}' \boldsymbol{\omega} \tag{46}$$

You need a weighting matrix $\boldsymbol{\Phi}^{-1}$ to do this, which ideally is

$$\boldsymbol{\Phi}^{-1} = (E(\mathbf{Z}' \boldsymbol{\omega} \boldsymbol{\omega}' \mathbf{Z}))^{-1}. \tag{47}$$

In practice, we use a consistent estimator of $\boldsymbol{\Phi}^{-1}$. Until step (4c), just use $\boldsymbol{\Phi}^{-1} = (\mathbf{Z}'\mathbf{Z})^{-1}$ as a starting point.

(4) Compute better estimates of all the parameters:the common parameters $(\alpha, \boldsymbol{\beta})$, the individual parameters $(\boldsymbol{\Pi}, \boldsymbol{\Sigma})$, and the weighting matrix $\boldsymbol{\Phi}$.

(4a) Find an estimate of the parameters that are common to all consumers, $(\alpha, \boldsymbol{\beta})$, using the GMM estimator,

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = (\mathbf{X}' \mathbf{Z} \boldsymbol{\Phi}^{-1} \mathbf{Z}' \mathbf{Z} \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}' \mathbf{Z} \boldsymbol{\Phi}^{-1} \mathbf{Z}' \boldsymbol{\delta} \tag{48}$$

This is a linear estimator that can be found analytically by multiplying matrices without any need for numerically minimizing something with search algorithm. Separating out the parameters that can be linearly estimated from the parameters that require a search algorithm is why we use all these steps instead of simply setting up the moment expression and then using a minimization algorithm to find parameter values that minimize it. Searching takes the computer longer than multiplying matrices, and is less reliable in finding the true minimum, or, indeed, in converging to any solution.

(4b) Estimate the value of the error term in (13n), $\hat{\boldsymbol{\omega}}$ and then the moment expression in (46), using the improved estimates of $(\alpha, \boldsymbol{\beta})$ from equation (48).

(4c) Estimate the value of the weighting matrix $\mathbf{\Phi} = \mathbf{Z}'\boldsymbol{\omega}\boldsymbol{\omega}'\mathbf{Z}$ using the $\hat{\boldsymbol{\omega}}$ just calculated.

$$\hat{\boldsymbol{\omega}}_{jt} = \boldsymbol{\delta}_{jt} - (\hat{\alpha}\mathbf{p_{jt}} + \mathbf{x_{jt}}\hat{\boldsymbol{\beta}}) \tag{49}$$

(4d) Use a search algorithm to find new values for $(\mathbf{\Pi}, \mathbf{\Sigma})$. Take the new values and return to step (1). Keep iterating, searching for parameter estimates that minimize the moment expression (46), until the value of the moment expression is close enough to zero.

Nevo notes that you could then iterate between estimating parameters (step 4a) and estimating the weighting matrix (step 4c). Both methods are consistent, and neither has more attractive theoretical properties, so it is acceptable to skip over step (4c) after the first iteration.

Conclusion

Now that we have gone through the entire procedure, it may be helpful to list some of the ideas we have used.

1. *Instrumental variables.* We use instruments to correct for the endogeneity of prices, the classic problem in estimating supply and demand.

2. *Product characteristics.* We look at the effect of characteristics on demand, and then build up to products that have particular levels of the characteristics. Going from 50 products to 6 characteristics drastically reduces the number of parameters to be estimated.

3. *Consumer and product characteristics interact.* This is what is going on when consumer marginal utilities are allowed to depend on consumer characteristics. This makes the pattern of consumer purchases substituting from one product to another more sensible.

4. *Structural estimation.* We do not just look at conditional correlations of relevant variables with a disturbance term tacked on to account for the imperfect fit of the regression equation. Instead, we start with a model in which individuals maximize their payoffs by choice of actions, and the model includes the disturbance term which will later show up in the regression.

5. *The contraction mapping.* A contraction mapping is used to estimate the parameters that are averaged across consumers, an otherwise difficult optimization problem.

6. *Separating linear and nonlinear estimation problems.* The estimation is divided into one part that uses a search algorithm to numerically estimate parameters that enter nonlinearly and a second part that uses an analytic formula to estimate the parameters that enter linearly.

7. *The generalized method of moments.* The generalized method of moments is used to estimate the other parameters.

Not all of these are special to the BLP method. Ideas (1), (2), and (3) can all be used with least squares (which itself is a simplified version of (7)). Idea (4) is used in standard logit. Ideas (5) and (6) are special to BLP, but of course BLP is a combination of all seven ideas, which is why it is so complex to explain.

The BLP method has been widely used because it is general enough to use for a variety of estimation problems in industrial organization, not just for simple demand problems. It is attractive compared to older methods because it imposes relatively little structure on the theoretical model, and so allows many different kinds of firm and consumer behavior to be tested. This flexibility, however, is achieved at the cost of considerable intricacy. The BLP method is made up of a modelling part and an estimation part. The modelling part is a logit model of a maximizing consumer's choice of product depending on consumer and product characteristics. This is a structural model, and really any structural model of maximizing choice, by consumer, government, or firm, could be used in its place. The estimation part estimates the importance of the product characteristics, consumer characteristics, and prices using the generalized method of moments. This is a highly flexible method, requiring weaker assumptions than maximium likelihood but like that procedure requiring a large number of observations and much computing power. I hope in this summary I have made clearer how the economist would go about combining this modelling and estimation that forms the BLP model.

References

Berry, Steven (1994) "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 25(2): 242-262 (Summer 1994).

Berry, Steven, James Levinsohn & Ariel Pakes (1995) "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4): 841-890 (July 1995).

Bertrand, J. (1883) "Theorie Mathematique de la Richesse Sociale," *Journal des Savants*, 67: 499-508 (1883).

Gould, Brian (2006) syllabus for "Agricultural and Applied Economics 743, Applied Consumption Analysis, Spring, 2006,"http://www.aae.wisc.edu/students/courses/syllabi/2006%5C743-Spring.pdf (2006).

Hall, Bronwyn H. (1996) "Notes on Generalized Method of Moments Estimation,"http://emlab.berkeley.edu/users/bhhall/e244/gmmnotes.pdf, (March 1996, revised February 1999).

Hall, Bronwyn H. (2005) "Computer Code for Problem Set 3 (Effects of Horizontal Merger),"http://emlab.berkeley.edu/users/bhhall/e220c/rc_dc_code.htm.

Hansen, L. P. (1982) "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4): 1029-1054 (July 1982).

Hicks, John (1937) "Mr Keynes and the Classics: A Suggested Simplification," *Econometrica*, 5(2): 147-159 (April 1937).

Hosken, Daniel, Daniel Brien, David Scheffman, and Michael Vita (2002) "Demand System Estimation and its Application To Horizontal Merger Analysis," forthcoming in *The Use of Econometrics in Antitrust*, American Bar Association Section on Antitrust, J. Harkrider, ed., FTC Bureau of Economics Working Paper, http://www.ftc.gov/be/workpapers/wp246.pdf (April 2002).

Hosoe, Moriki & Eric Rasmusen, eds. (1997) *Public Policy and Economic Analysis*, Fukuoka, Japan: Kyushu University Press (1997).

Nevo, Aviv (2000) "A Practitioner's Guide to Estimation of Random-

Coefficients Logit Models of Demand," *Journal of Economic and Management Strategy,* 9(4): 513-548 (Winter 2000).

Nevo, Aviv "Appendix to 'A Practitioner's Guide to Estimation of Random Coefficients Logit Models of Demand Estimation: The Nitty- Gritty'," http://www.faculty.econ.northwestern.edu/faculty/nevo/supplements/Ras_guide_appendix.pdf.

Rasmusen, Eric (1998a) "Observed Choice, Estimation, and Optimism about Policy Changes," *Public Choice*, 97(1-2): 65-91 (October 1998).

Rasmusen, Eric (1998b) "The Observed Choice Problem in Estimating the Cost of Policies," *Economics Letters*, 61(1): 13-15 (1998).

Stone, Richard (1954) "Linear Expenditure Systems and Demand Analysis: An Application to the Pattern of British Demand," *Economic Journal*, 64 (255): 511-527 (September 1954).

Train, Kenneth (2003) *Discrete Choice Methods with Simulation*, Cambridge University Press, http://elsa.berkeley.edu/books/choice2.html (2003).

Wooldridge, Jeffrey M. (2001) "Applications of Generalized Method of Moments Estimation," *Journal of Economic Perspectives*, 15(4): 87-100 (Fall 2001).

Working, E. J. (1927) "What Do Statistical Demand Curves Show?" *Quarterly Journal of Economics*, 41(2): 212-235 (February 1927).

"1.3.6.6.16. Extreme Value Type I Distribution," *The Engineering Statistics Handbook*, http://www.itl.nist.gov/div898/handbook/eda/section3/eda366g.htm.