**Subject:** Demand Estimation Methodology for MGS

**From:** Ryan Cummings and Daniel Posthumus

**To:** Neale Mahoney

**Date:** 1/10/2025

**Summary**

In this memo, we summarize ways to model a demand system based on the gasoline market in California, **with access only to market-level data**. As of now, we have no consumer-level data. We believe that given the data and our objective at this point, the best course of action is to follow a standard procedure for estimating a demand system, that is:

1. Estimate a simple pure logit model to get reduced-form estimates that illustrate the importance of both brand and location.
2. To incorporate systematic heterogeneity in consumers’ preferences, implement a more “structural” model through the classic BLP (1995) model.
   1. That is, we define “markets” as those within a certain geographic region (starting with using zip code, although we will need to apply robustness tests to this) for a given year (i.e., year-market).
   2. We would then calculate market share as the percent of all gasoline sold by a given station within that geographic area.
   3. Next, we would specify a random coefficients model, using standard BLP methodology to pin down the “” (linear parameters of observables and nonlinear parameters of the random coefficients).
   4. Importantly, we want to model distance to a gas station as a feature in consumer preferences. This could be done through including the distance from the center of geographic area into the station attributes and then having a random coefficient for distances.

We discuss other possible demand models as well, discussing their advantages and disadvantages. To guide us in our tour of demand estimation techniques, we relied on a modern (2021) review of demand estimation, see this [*NBER* article](https://www.nber.org/system/files/working_papers/w29305/w29305.pdf) by Berry and Haile. We begin with the simplest possible model, working adding on models with increased richness, and increased complexity, as we go.

1. **Linear Demand Model**

In the linear demand model, we can write that consumer *i*’s utility from good *j* takes the following form:

Here, utility is only a function of price; this is obviously not borne out in our data, where there is the significant and growing gap in branded and unbranded prices. Of course, we can add dummies for brands or other characteristics, but this overly simplistic regression would feature endogeneity and render the coefficients uninstructive.

1. **Pure Logit Model**

The most basic model assumes away systematic heterogeneity in individuals’ preferences, but should still give yield some reduced-form estimates that are instructive. To do this, we express the mean utility of good *j* in market *t* as:

And thus utility (where consumer systematic heterogeneity is zero) is defined as:

Whereis a vector of non-price characteristics of product *j* in market *t* and \xi\_{jt} are unobserved product characteristics.

Then, each market share is:

Then, we can write and estimate:

Where is the market share of the outside good, which in our case would be going outside the geographic area to get gas, since presumably “not getting gas” isn’t a choice if your car needs fuel! This model is fully-identified, as we have observed shares for each good (yearly quantity data), prices (daily price data), and we can develop some gas station-specific characteristics, such as a measure of its proximity to commuter routes (an issue addressed in the subsequent section). We would also incorporate brand into this vector of characteristics.

is of course endogenous, as it is correlated with unobserved product characteristics \xi\_{jt}. Thus, we would have to instrument it—which we also have to do in BLP. Possible instruments include entries from/exits to nothing (a shock to demand) or cost of crude (a cost shifter).

The advantage of this approach is that it’s easy/straightforward and very computationally cheap. This makes sense as a first step/first pass to see if our basic approach is correct. The disadvantage, as stated above, is that it does not allow for systematic heterogeneity in individuals’ preferences, restricting elasticity estimation. This runs counter to our story of branded market power, which is predicated on there being different ‘types’ of consumers preferring different brands.

1. **BLP/Random Coefficients/Mixed Logit**

The next step would be to model for more systematic heterogeneity in consumer preferences. To tackle this, though, we need a more structural model of consumer preferences. Let’s reintroduce a non-zero consumer systematic heterogeneity. Now:

And then we can write

Where represents *all* observed characteristics, including price.

Then, if we want to capture systematic heterogeneity, we need to estimate a new coefficient, which variables at the product-consumer level. In a parametric model, we need to draw this from some distribution, most specified as:

Where shifts preferences according to observed demographics drawn from a data source like the census (the array of which is referred to as ). Then refers to unobserved preferences among the population of consumers.

This allows to decompose \beta\_{it} into two parts: 1) the homogenous component (denoted by \beta) and 2) the heterogeneous component (denoted by \Sigma v\_{it} + \pi y\_{it}) such that:

Then, market shares are determined according to the following equation:

Which we identify using the standard BLP procedure where either GMM or MLE is employed in both an outer loop (for nonlinear parameters) and an inner loop (to find linear parameters), and this estimation is iterated until a given distance criterion is satisfied.

Advantages of BLP:

* Allows for *systematic* heterogeneity in consumer preferences

Disadvantages of BLP:

* Parameters are unknown; estimating them requires an iterative non-linear optimization problem, meaning there is no mathematical guarantee of a solution.
* This computational difficulty has helped to cause a lack of a standardized implementation of the model, resulting in numerical inconsistencies.

1. **Micro BLP**

Micro BLP approach connects microdata on the consumer level to the BLP estimator; given that we don’t have consumer-level data, this approach is likely not applicable/appropriate for our model as of yet. Perhaps at some point we can supplement

1. **Spatial analysis.**

I think we have the following outstanding questions which can help us determine the appropriate model/approach to the spatial component of the gasoline market:

* Do we want a non-parametric or parametric model?
* What kind of data can we get access to?
  + Is it possible to develop some algorithm for coding what services each of the gas stations in our dataset offer? Car wash, mechanic service, food mart, etc.
  + Can we get any data on commuting patterns on a zip-code level?
  + Can we get some kind of measure of the density of businesses/residences on a zip-code level?

One potential solution is to, within the BLP framework, include distance from ‘center’ of market as a gas station characteristic. For example, the center could be the geographic center of a zip code, a position denoted by \bar{l}\_t; then, the distance of gas station *j* from the center would be denoted by (l\_{jt} - \bar{l}\_t) \in x’\_{jt} and would be included in our vector of gas station characteristics. For now, let’s say the other two station characteristics are price (p\_{jt}) and brand (b\_{jt}). Then, the utility equation in random coefficients takes on the following form: