MGS Demand Estimation - Bay Area

Daniel Posthumus

February 1, 2025

1 Introduction

In this memo, I explain and present results from **preliminary** demand estimation in the California gasoline markets. Specifically, I address the following:

- Describing the merged dataset of A15 and OPIS Retail Data, along with external datasources.
- Presenting Pure Logit estimation and results.
- Presenting Nested Logit estimation and results.

2 Data

2.1 Sources of Data

I use the following data series

- Daily, gas station-level retail prices from OPIS.
  - I collapse on the (unweighted) average for each gas station in a given year.
- Annual, gas station-level quantity sold from CDTFA/ PIIRA. This is the A15 data.
- EIA's annual series, "U.S. Gulf Coast Conventional Gasoline Regular Spot Price FOB".
- The National Neighborhood Data Archive's (NaNDA) data series on yearly traffic volume by Zip Code Tabulation Area.

2.2 Defining Market Size

While we have observed gas station-level quantities sold, this doesn't allow us to directly calculate observed market shares; instead, we need the denominator–the market size M_t. This isn't as simple as summing the observed quantities sold, as we need to capture 'consumption' of the outside option in each market, effectively all consumers that did *not* purchase gasoline for a given zip code-year.

In practice, estimates of market size are very rough; in BLP's seminal paper (1995), for example, they simply took the number of registered vehicles in all of the United States as an estimate for the market size M_t of the national automobile market. Clearly, this is only a rough approximation–many people would be happy with their cars and not looking to buy a new one, regardless of price changes.

The key to our defining of market size is to generate some variation between markets, capturing systematic heterogeneity in the density of traffic across California. Our solution to this was to use traffic volume data from the National Neighborhood Data Archive (NaNDA). This data yields the average number of cars observed in a given Zip Code Tabulation Area (ZCTA); then, to estimate the number of possible gallons to be sold in each ZCTA, I multiply this number by 489 gallons of gasoline, estimated by API to be the average gasoline consumed in a given year by each registered vehicle in the United States. This results in an estimate of the *maximum* potential gallons of gasoline consumed in a given geographic year for a given year; the difference between this maximum and the observed quantities of gasoline sold would result in the outside share's quantity/share.

2.3 Descriptive Statistics

Here is a table containing descriptive statistics by possible market identifications, specifically either 1) ZIP-year or 2) ZIP3-year.

Table 1: Summary Statistics by Market Definition

	ZIP-Year	ZIP3-Year
Mean Number of Stations	5.02	80.25
$\bar{s}$_0	0.77	0.31
$\bar{s}$_j	0.057	0.005
Average % of Stations That Are Unbranded	18.7	14.6
Average Within-Market Std. Dev. of Observed Quantity	875371	1139785
Average Within-Market Std. Dev. of Prices	0.13	0.18
Total Number of Markets	5087	318

Here is a table showing summary statistics statistics by store brand (I selected the top 10 most common brands in California by the number of unique gas stations belonging to each brand):

Table 2: Summary Statistics by Store Brand

	Retail Prices								
Store Brand	Min.	25pctile	Median	75pctile	Max.	Total # of Stations	% of Observed Stations	$\bar{s_j}$ (ZIP3)	$\bar{s_j}$ (ZIP)
unbranded	2.14	2.83	3.14	3.71	5.25	1331	15.47	0.01	0.21
chevron	2.29	3.08	3.33	3.85	5.20	1312	15.25	0.01	0.21
shell	2.30	3.02	3.27	3.82	5.04	1045	12.15	0.01	0.20
76	2.16	3.04	3.37	3.84	4.94	935	10.87	0.01	0.16
valero	2.36	2.94	3.20	3.76	5.09	538	6.25	0.01	0.15
ampm	2.27	2.72	2.98	3.56	4.04	497	5.78	0.02	0.24
arco	2.12	2.73	3.00	3.59	4.85	460	5.35	0.02	0.24
mobil	2.38	3.02	3.38	3.86	4.84	386	4.49	0.01	0.15
7_eleven	2.35	2.86	3.18	3.71	4.48	322	3.74	0.01	0.12
circle_k	2.30	2.94	3.32	3.81	4.26	267	3.10	0.01	0.18

3 Pure Logit

3.1 Estimation

As discussed further below in the appendix, by assuming there is no systematic consumer heterogeneity in preferences, we are left with the following equation:

$$\ln(s_j) - \ln(s_0) = \ln(\frac{s_j}{s_0}) = x_j\beta - \alpha p_j + \xi_j \tag{1}$$

This equation is fully identified; market shares, including for the outside good; prices; and product characteristics are all observed. This allows us to estimate the equation using OLS or 2SLS. The motivation for 2SLS is clear; price is endogenous, as it will be correlated with unobserved demand shocks, represented by ξ_{jt}. There are several possible instruments for price; for now, we are using one type–a cost shifter that changes marginal cost in an exogenous fashion–RBOB spot prices on the gulf coast. RBOB represents one input cost for producing CARBOB, the CA-specific blend of gasoline required by law to be sold in the state. However, the marginal barrel of gasoline sold in California originates from Asia, so we can conclude that demand in California would not substantively shift RBOB spot prices on the Gulf Coast.

For now, the only non-price characteristics included are selected brand fixed effects; I selected the top 5 most common brands in the data as well as unbranded gas stations and included fixed effects for these 6 categories of stations. Thus, the estimates of the fixed effects should be interpreted as relative to all gas stations in the market outside of the top 5 most common brands (which appear to align with refiner brands) and unbranded stations. I didn't include a constant.

While α allows us to model and estimate elasticities, it is also not interpretable. Therefore, in regression tables below, I run the following regression to allow us to interpret the price coefficient (labelled as 'log prices' hereafter). However, in the estimated elasticities, I used α estimated according to the above equation, with an un-transformed price regressor. In specifications where I use the RBOB Spot Price as an instrument, I take the natural log of the spot price for consistency.

$$\ln(\frac{s_j}{s_0}) = x_j\beta + \zeta \ln(p_j) + \xi_j \tag{2}$$

Derived in the appendix, the elasticities for this pure logit model take the following forms (own- and cross-price elasticities, respectively):

$$\nu_{jj} = \alpha \cdot p_j \cdot (1 - s_j) \tag{3}$$

$$\nu_{jk} = -\alpha \cdot p_k \cdot s_k \tag{4}$$

The elasticties reported in the results table below are the simple average across all products in all markets.

3.2 Results

First, we start with the simplest model, there is no product differentiation. Thus, we have just two products: the inside good and the outside good. To formalize this market, there are only $j \in 0, 1$ products where j_{0t} is the outside good and j_{1t} is the inside good. We calculate s_{1t} as the sum of the observed shares in the data. We take p_{1t} as the market share-weighted average price of all observations in the data for market t. Since there is no intra-market variation in prices, we are not able to use ZIP fixed effects, and run two specifications: 1) without and 2) with the RBOB Gulf Coast Spot Price as an instrument for retail price.

The results from this estimation are reported below in Panel A, with the first stage coefficient on the RBOB instrument in Panel B, the results of which indicate the instrument is quite strong.

Table 3: Pure Logit Results Without Product Differentiation

	(1)	(2)
Panel A: Estimation Results		
log prices	0.094	-1.347***
	(0.189)	(0.023)
$\bar{\nu}_{jk}$	nan	nan
$\bar{\nu}_{jj}$	-0.075	1.294
N	5087	5087
RBOB Instrument	NO	YES
Panel B: First-Stage Results		
RBOB Instrument	NA	0.789***
	NA	(0.003)

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Excluded brand fixed effects are all brands not falling under either 1) the top 5 most common brands across the entire sample and 2) all unbranded stations.

Next, I estimated the pure logit model *with* product differentiation among the inside goods, introducing brand fixed effects for the top 5 most common brands and unbranded stations. I've included results estimated directly using either OLS or IV in Panel A. I've also included the first stage coefficient on the RBOB instrument in Panel B.

Table 4: Pure Logit Results With Product Differentiation

	(1)	(2)	(3)	(4)
Panel A: Estimation Results				
log prices	1.053***	1.362***	-2.284***	2.135***
	(0.108)	(0.093)	(0.024)	(0.099)
$\bar{\nu_{jk}}$	0.037	0.047	-0.098	0.073
$\bar{\nu_{jj}}$	-1.059	-1.352	2.795	-2.076
N	25959	25959	25959	25959
ZIP CODE FE	NO	YES	NO	YES
RBOB Instrument	NO	NO	YES	YES
Panel B: First-Stage Results				
RBOB Instrument	NA	NA	0.785***	0.784***
	NA	NA	(0.002)	(0.002)

Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Excluded brand fixed effects are all brands not falling under either 1) the top 5 most common brands across the entire sample and 2) all unbranded stations.

3.3 Elasticity Matrix

3.4 The Extensive Margin

Of particular importance is substitution patterns between, broadly, the inside good and the outside good–here referred to as the 'extensive margin'.

4 Nested Logit

As discussed in the appendix, cross-price elasticities estimated by the pure logit model have a serious shortcoming: they only depend on one of the two products' price. We can counter this shortcoming, called the Independence of Irrelevant Alternatives (IIA) property, through a nested logit which has looser restrictions on the elasticity patterns. There are generally three approaches we can take to nested logit based on different nesting schemes:

1. **One nest**: 1) all stations
2. **Two nests**: 1) unbranded and 2) branded stations
3. **Three nests**: 1) unbranded, 2) hypermarket, and 3) non-hypermarket branded stations

4.1 Estimation

Based on the derivation in the appendix, we can use the following linear estimating equation for nested logit (for nest h):

$$\ln(\frac{s_j}{s_0}) = x_j\beta - \alpha p_j + \rho \ln s_{j|h(j)} + \xi_j \tag{5}$$

where $s_{j|h(j)}$ is the within-nest market share (i.e., market share conditional on nest-market) for product j in nest h. Note that estimating this equation requires two different instruments, one for price (we'll continue using the gulf coast spot price for RBOB) and the other for the within-group market share, which is also correlated with unobserved idiosyncratic demand shocks.

Some ideas for instruments for within-nest market shares:

- Number of stations in a nest-market

- Sum of characteristics of other firms in market, i.e. a station's competitors

Then, once again, to estimate an *interpretable* coefficient on price, I also report the price coefficient estimated from this equation:

$$\ln(\frac{s_j}{s_0}) = x_j\beta^{\ln} + \zeta_1 \ln(p_j) + \zeta_2 \ln s_{j|h(j)} + \xi_j^{\ln} \tag{6}$$

In the results tables, below, I report the coefficients on $\ln(p_j)$ (referred to as 'log prices') and $\ln s_{j|h(j)}$ (referred to as 'log conditional shares'); however, when deriving the elasticities, I use α and ρ per the above linear equation.

Derived in the appendix, the elasticities for the nested logit take the following forms for 1) own-price elasticity (ν_{jj}), 2) cross-price elasticity between two products in the same nest (ν_{jk}), and 3) cross-price elasticity between three products in different nests (ν_{jf}).

$$\nu_{jj} = \frac{\partial s_j}{\partial p_j}\frac{p_j}{s_j} = \frac{-\alpha \cdot p_j}{1-\rho}(1 - \rho s_{j|h} - (1-\rho)s_j) \tag{7}$$

$$\nu_{jk} = \frac{\partial s_j}{\partial p_k}\frac{p_k}{s_j} = \alpha \cdot p_k s_k (1 + \frac{\rho}{1-\rho} s_{j|h}) \tag{8}$$

$$\nu_{jf} = \frac{\partial s_j}{\partial p_f}\frac{p_f}{s_j} = \alpha \cdot p_f s_f \tag{9}$$

What I report in the tables below are the simple average of each of these elasticities across all products.

4.2 Results

Here are the results from Nested Logit, manually estimated using IV methods, with number of firms per nest-market as the instrument for conditional share in Panel A, and the first stage results for both RBOB and the conditional share in Panel B. Also reported are the mean elasticities.

Note that $\bar{\nu_{jf}}$ reports the elasticity of good j to the price of good f, which *is not* contained within the nest of good j. On the other hand, $\bar{\nu_{jk}}$ reports the elasticity of different goods in the same nest.

	(1)	(2)	(3)
Panel A: IV Estimation Results			
log prices	-1.199***	-1.081***	-1.100***
	(0.033)	(0.035)	(0.033)
log conditional shares	0.666***	0.761***	0.786***
	(0.014)	(0.016)	(0.015)
$\bar{\nu_{jj}}$	-3.573	-3.816	-4.088
$\bar{\nu_{jk}}$	0.050	0.041	0.037
$\bar{\nu_{jf}}$	nan	0.053	0.063
N	25959	25959	25959
Panel B: First-Stage Results			
Number of Firms In Nest-Market	-0.152***	-0.172***	-0.178***
	(0.003)	(0.003)	(0.003)
NESTS	INSIDE	BRANDED – UNBRANDED	NON-HYPER BRANDED – HYPERMARKET – UNBRANDED

5 Appendix

5.1 'Cheat Sheet' of Market Share and Elasticity Equations

5.1.1 Pure Logit Model

$$\ln(\frac{s_j}{s_0}) = x_j - \alpha p_j + \xi_j \tag{10}$$

$$\nu_j j = \alpha \cdot p_j \cdot (1 - s_j) \tag{11}$$

$$\nu_j k = -\alpha \cdot p_k \cdot s_k \tag{12}$$

5.1.2 Nested Logit Model

For nested logit, we have three elasticities: 1) own-price elasticity (ν_{jj}), 2) cross-price elasticity between two different products in the same nest (ν_{jk}), and 3) cross-price elasticity between two products in different nests (ν_{jf}).

$$\ln(\frac{s_j}{s_0}) = x_j\beta - \alpha p_j + \rho \ln s_{j|h(j)} + \xi_j \tag{13}$$

$$\nu_{jj} = \frac{\partial s_j}{\partial p_j}\frac{p_j}{s_j} = \frac{-\alpha \cdot p_j}{1-\rho}(1 - \rho s_{j|h} - (1-\rho)s_j) \tag{14}$$

$$\nu_{jk} = \frac{\partial s_j}{\partial p_k}\frac{p_k}{s_j} = \alpha \cdot p_k s_k (1 + \frac{\rho}{1-\rho}s_{j|h}) \tag{15}$$

$$\nu_{jf} = \frac{\partial s_j}{\partial p_f}\frac{p_f}{s_j} = \alpha \cdot p_f s_f \tag{16}$$

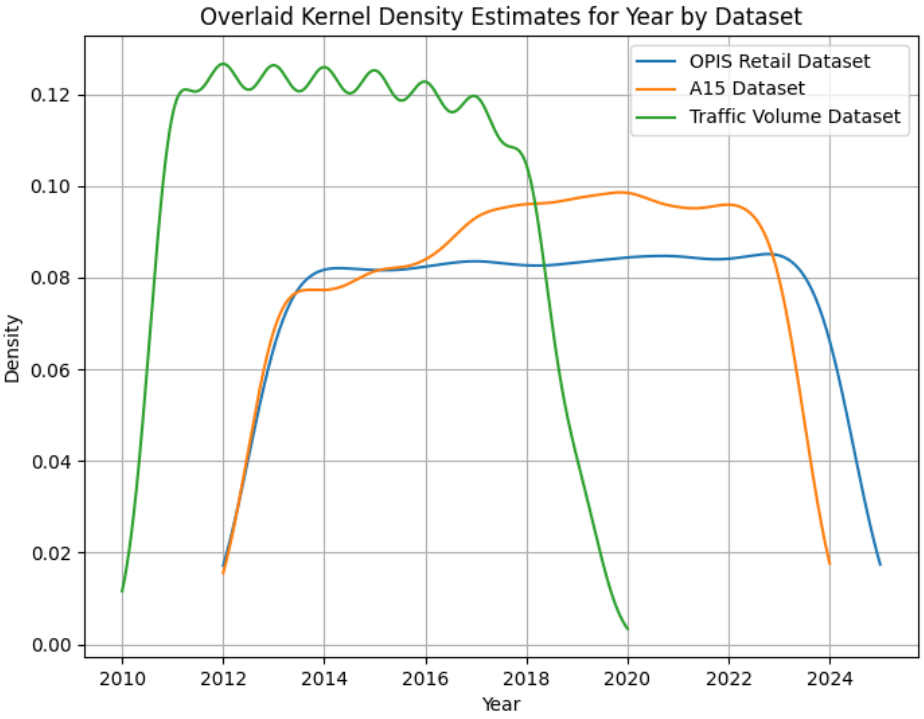
6 Data Build Notes

Building the data required different sources of data and upon merging these sources of data, some observations were lost. To ensure this didn't introduce bias into the data, I've compiled the following table comparing summary statistics for different stages of the data build:

Table 5: Summary Statistics Across Steps of Data Build

	OPIS Retail Dataset	A15 Data	Present in Both OPIS and A15 Data	OPIS/A15/Traffic Data	$s_0 > 0$ Condition Met
Number of Observations	117955	84926	75040	39625	33416
Observations as % of OPIS Retail Data	100.0	72.0	63.6	33.6	28.3
Number of Unique Stations	15835	13663	12426	10265	9630
Unique Stations as % of OPIS Retail Data	100.0	86.3	78.5	64.8	60.8
% of Observations From the Bay Area	17.3	17.2	17.6	15.9	15.7
% of Observations From Los Angeles County	20.7	20.4	20.6	19.7	20.7
% of Observations From Unbranded Stations	16.2	NA	13.6	14.6	13.9
Average Retail Price	3.87	NA	3.80	3.33	3.31
Average Quantity Sold	NA	1138163	1160774	1228538	1214994
Average Year	2018.5	2018.2	2018.2	2015.3	2015.2

Our different sources of data have differing intervals of time covered, as the kernel density plot estimate below illustrates:



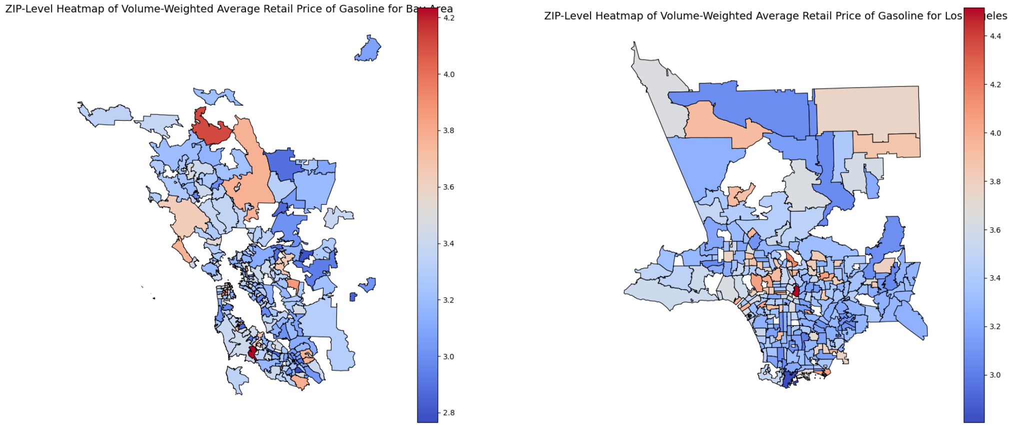
Thus, I can recreate the summary statistics table, *first restricting* the OPIS and A15 data by date (to 2013-2019):

Table 6: Summary Statistics Across Steps of Data Build

	OPIS Retail Dataset	A15 Data	Present in Both OPIS and A15 Data	OPIS/A15/Traffic Data	$s_0 > 0$ Condition Met
Number of Observations	68117	84926	46056	32106	25959
Observations as % of OPIS Retail Data	100.0	124.7	67.6	47.1	38.1
Number of Unique Stations	13048	13663	10415	9192	8604
Unique Stations as % of OPIS Retail Data	100.0	104.7	79.8	70.4	65.9
% of Observations From the Bay Area	17.6	17.2	17.6	16.5	16.2
% of Observations From Los Angeles County	20.8	20.4	20.5	19.2	21.0
% of Observations From Unbranded Stations	17.6	NA	14.5	14.4	14.1
Average Retail Price	3.39	NA	3.40	3.32	3.31
Average Quantity Sold	NA	1138163	1248617	1252944	1243409
Average Year	2016.0	2018.2	2016.2	2015.3	2015.2

7 State-Wide Geographic Visualizations

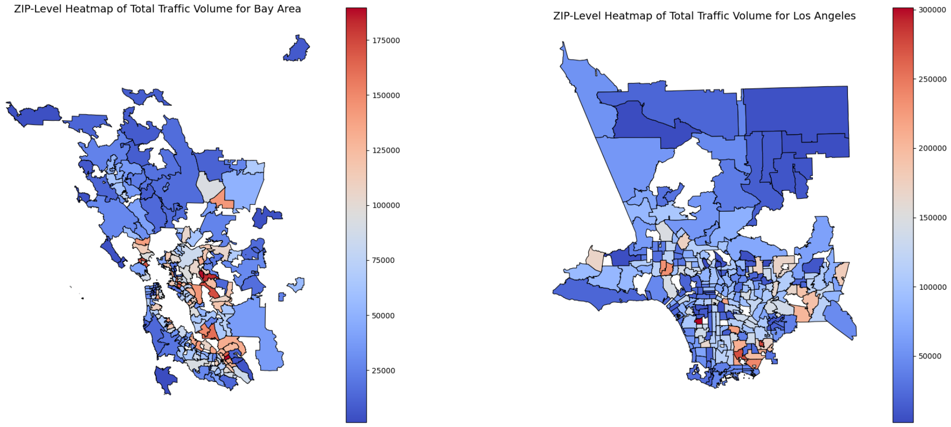
Figure 1: Heatmaps of Volume-Weighted Average Retail Prices on ZIP Level


ZIP-Level Heatmap of Volume-Weighted Average Retail Price of Gasoline for Bay Area
4.2
4.0
3.8
3.6
3.4
3.2
3.0
2.8
ZIP-Level Heatmap of Volume-Weighted Average Retail Price of Gasoline for Los Angeles
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0

(a) Bay Area

(b) Los Angeles County

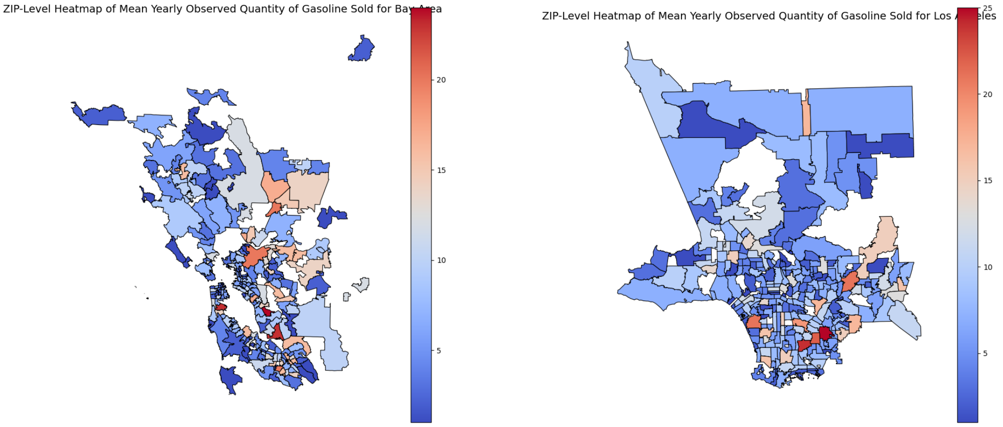
Figure 2: Heatmaps of Average Daily Traffic Volume on ZIP Level


ZIP-Level Heatmap of Total Traffic Volume for Bay Area
175000
150000
125000
100000
75000
50000
25000
ZIP-Level Heatmap of Total Traffic Volume for Los Angeles
300000
250000
200000
150000
100000
50000

(a) Bay Area

(b) Los Angeles County

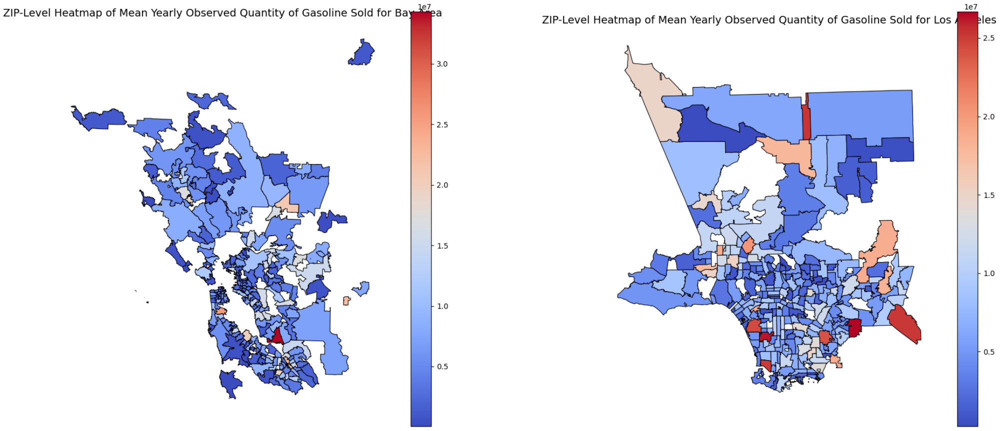
Figure 3: Heatmaps of Uniquely Identified Gasoline Stations on ZIP Level


ZIP-Level Heatmap of Mean Yearly Observed Quantity of Gasoline Sold for Bay Area
20
15
10
5
ZIP-Level Heatmap of Mean Yearly Observed Quantity of Gasoline Sold for Los Angeles
25
20
15
10
5

(a) Bay Area

(b) Los Angeles County

Figure 4: Heatmaps of Average Annual Observed Quantity Sold of Gasoline on ZIP Level


ZIP-Level Heatmap of Mean Yearly Observed Quantity of Gasoline Sold for Bay Area
1e7
3.0
2.5
2.0
1.5
1.0
0.5
ZIP-Level Heatmap of Mean Yearly Observed Quantity of Gasoline Sold for Los Angeles
1e7
2.5
2.0
1.5
1.0
0.5

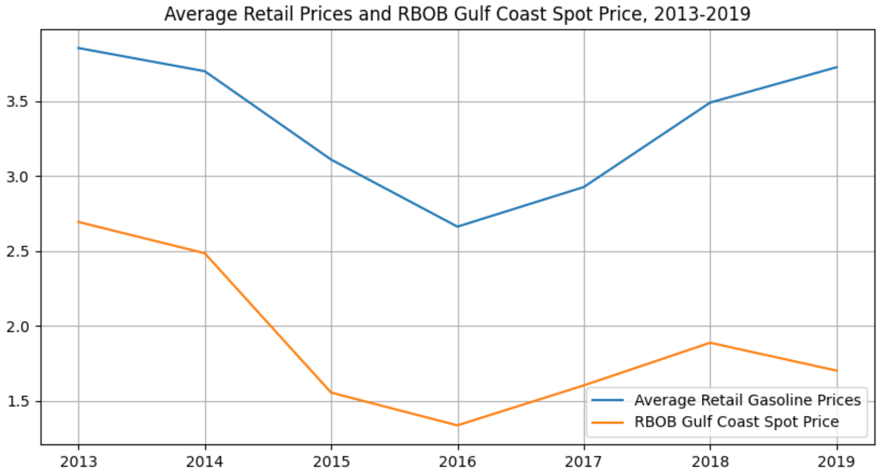
(a) Bay Area

(b) Los Angeles County

8 Instruments Descriptives

8.1 RBOB Gulf Coast

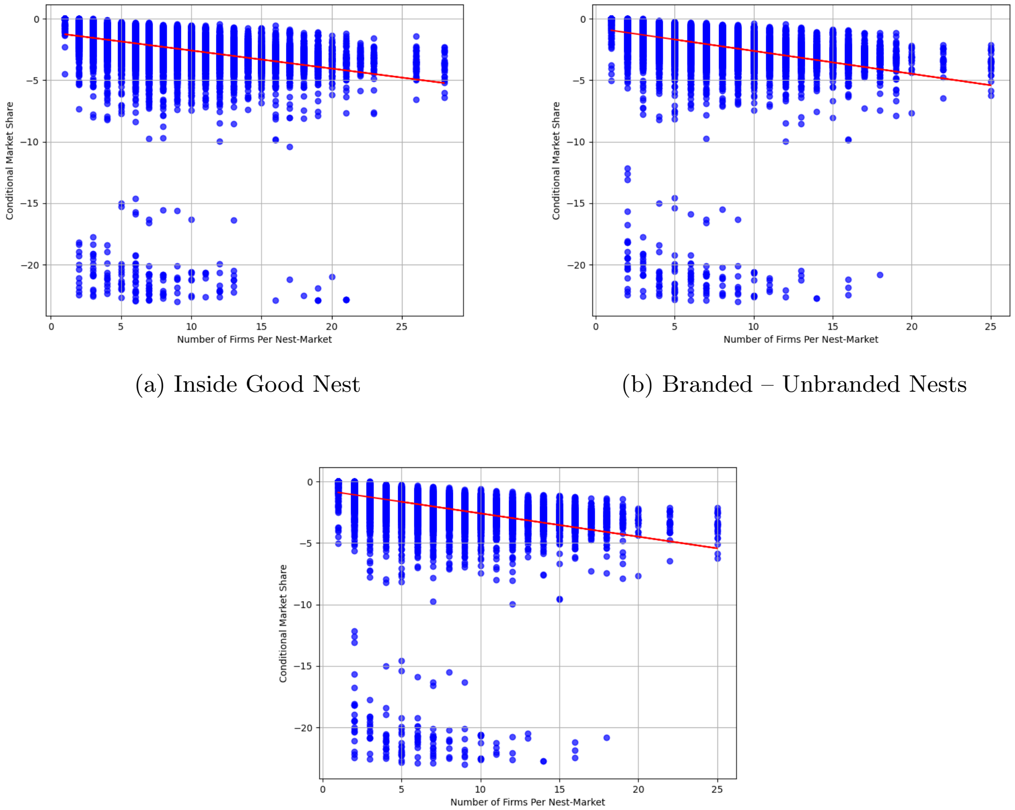
Here is a time-series, with the volume-weighted average gasoline prices in the sample plotted alongside the RBOB Gulf Coast Spot Price (all nominal) over time:



8.2 Number of Firms in Each Nest-Market

Below, I've included a scatterplot showing the relationship between the number of products in each nest-market (by nest designation) and log conditional market share. This provides suggestive evidence that there is a positive correlation here, suggesting the number of products in nest-market has some strength as an instrument.

Figure 5: Scatterplots of Number of Products In Nest-Market and Market Share Conditional on Nest-Market



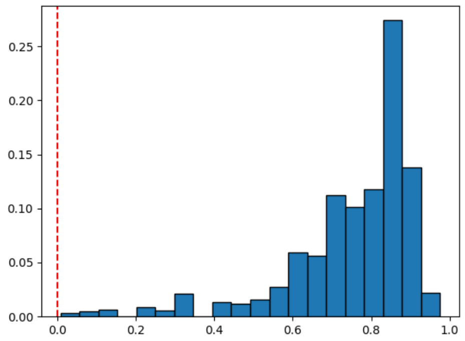
(a) Inside Good Nest

(b) Branded – Unbranded Nests

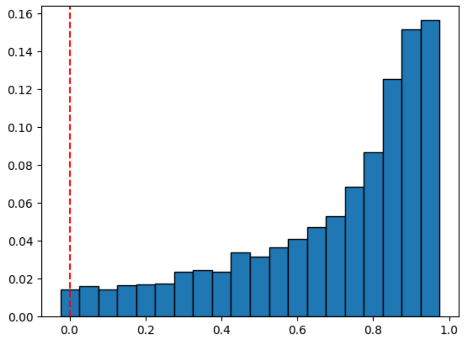
(c) Non-Hypermarket Branded – Hypermarket – Unbranded Nests

9 Market Shares Descriptives

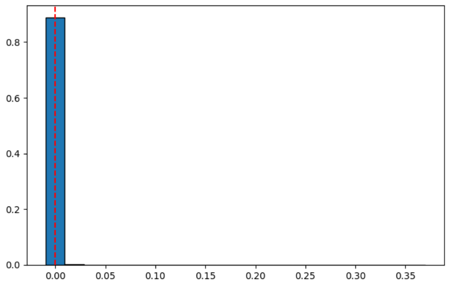
There were some outliers where it appears the NaNDA traffic data was off by an order of 10; after dropping instances where the resulting outside good market share (s_{0t}) was negative (this amounted to), we have the following distributions for the outside good and inside good shares:



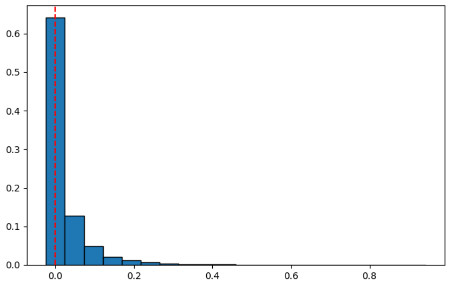
(a) Zip3-Level Market Outside Good Shares (s_{0t})



(b) Zip-Level Markets Outside Good Shares (s_{0t})



(c) Zip3-Level Market Inside Good Shares (s_{jt})



(d) Zip-Level Markets Inside Good Shares (s_{jt})

Although very large, these outside shares appear reasonable and, most importantly, incorporate the significant heterogeneity in California in commuting and travel patterns.