Project 3: Car Demand and Home Market Bias

Due: Wednesday December 13th at 22:00

1 Car Demand

A key challenge to any firm is to set prices in an optimizing manner. Firms must balance the tradeoff between maximizing the market share and maximizing the price commanded per sale. To do so, it is crucial to have an accurate model of consumer demand. In this project, you will be using data on market shares to infer the parameters of the utility functions of individuals giving rise to consumer demand.

Let i index *markets* (defined as country-year pairs), and let j index *alternatives* (i.e. types of cars). To motivate the demand model, suppose that each country consists of households $h = 1, \ldots, H$, and that the utility that household h in market i receives from choosing car j among J mutually exclusive options (labelled $1, 2, \ldots, J$) is

$$u_{ijh} = \mathbf{x}_{ij}\boldsymbol{\beta}_o + \varepsilon_{ijh}, \quad j = 1, \ldots, J,$$

where $\mathbf{x}_{ij}$ is a $1 \times K$ vector of observable market-car characteristics, $\boldsymbol{\beta}_o$ an unknown $K \times 1$ vector of parameters, and ε_{ijh} is a (scalar) error term observed by the household, but not the econometrician (and, thus, from the perspective of the econometrician, it is perceived as random). Here the $\{\varepsilon_{ijh}\}$ are presumed (Type-I) *extreme value* distributed independently over both markets, alternatives and households.[1] The households act in a utility-maximizing manner. Let $\mathbf{x}_i$ be the $J \times K$ matrix arising from stacking the $\mathbf{x}_{ij}$ over j. With the particular distributional assumption on the error term, conditional on $\mathbf{x}_i$, the probability that j gives the highest utility can be shown to be

$$\Pr(\text{household } h \text{ chooses car } j|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_{ij}\boldsymbol{\beta}_o)}{\sum_{k=1}^{J}\exp(\mathbf{x}_{ik}\boldsymbol{\beta}_o)} =: s_j(\mathbf{x}_i, \boldsymbol{\beta}_o).$$

In general, the market share of j is given by the average choice probability, averaging over households. However, since the households are identical (from the perspective of the econometrician), we can think of s_j as both the individual household's *choice probability function*

[1] A random variable is distributed (Type-I) *extreme value* if its cumulative distribution function (CDF) is $F(z) = \exp[-\exp(-z)], z \in \mathbb{R}$ (see, e.g., https://en.wikipedia.org/wiki/Generalized_extreme_value_distribution). Some people refer to this distribution as the (standard) *Gumbel* distribution (see, e.g., https://en.wikipedia.org/wiki/Gumbel_distribution). Beware that some programming languages (such as MATLAB) define the extreme value distribution in a different ("mirrored") way (cf. https://se.mathworks.com/help/stats/extreme-value-distribution.html). We here follow the definition and terminology most common to economists.

and the *market share function*. We can naturally extend the definition of the market share function to allow for any candidate parameter $\boldsymbol{\beta} \in \mathbb{R}^K$,

$$s_j(\mathbf{x}, \boldsymbol{\beta}) := \frac{\exp(\mathbf{x}_j \boldsymbol{\beta})}{\sum_{k=1}^{J} \exp(\mathbf{x}_k \boldsymbol{\beta})},$$

where $\mathbf{x}_j$ denotes the jth row of a matrix $\mathbf{x} \in \mathbb{R}^{J \times K}$ (capturing a possible realization of $\mathbf{x}_i$).

In the data, we have *observed market shares* $\mathbf{y}_i := (y_{i1}, \ldots, y_{iJ})'$, where y_{ij} is the share of households in market i that purchased car j. The log-likelihood contribution for market i is

$$\ell_i(\boldsymbol{\beta}) := \sum_{j=1}^{J} y_{ij} \ln s_j(\mathbf{x}_i, \boldsymbol{\beta}),$$

and we obtain an estimator by maximizing the average of log-likelihood contributions from N markets over candidate parameter vectors $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^K} \frac{1}{N} \sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}).$$

When thinking about approximations, we consider many-markets asymptotics ($N \to \infty$). That is, we already consider the number of households H to be large enough for $s_j(\mathbf{x}_i, \boldsymbol{\beta}_o)$ to be a good approximation to (in fact, equal to) y_{ij}.[2]

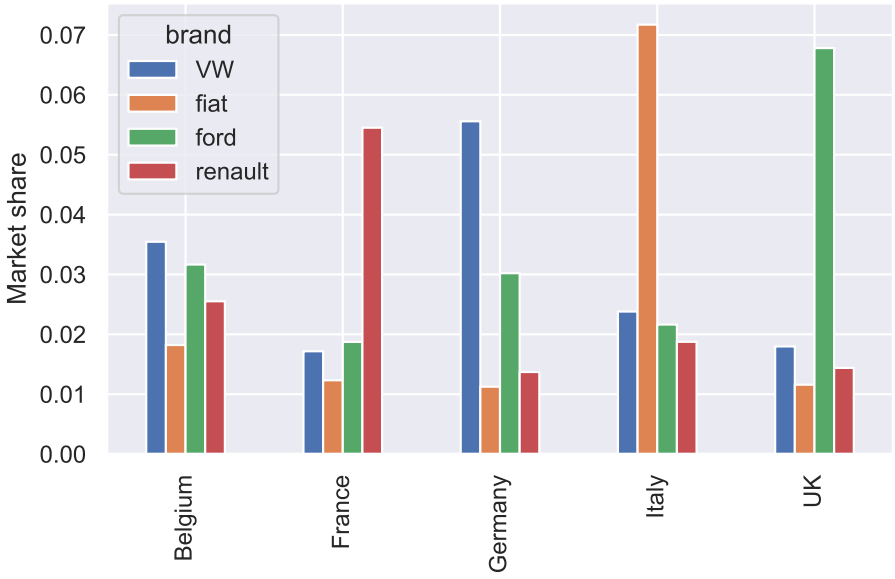
Once we have obtained estimates, using the model we can quantify many interesting aspects of demand. For example, we may be interested in the elasticity of demand of a particular car with respect to own price. Specifically, if P_{ij} (an element of $\mathbf{x}_{ij}$) denotes the price of car j in market i and p_{ij} a possible value thereof, then our interest lies in

$$\mathcal{E}_{jj}(\mathbf{x}) := \frac{\partial s_j(\mathbf{x}, \boldsymbol{\beta}_o)}{\partial p_{ij}} \cdot \frac{p_{ij}}{s_j(\mathbf{x}, \boldsymbol{\beta}_o)}.$$

Note that $\mathcal{E}_{jj}(\mathbf{x})$ generally depends on the point of evaluation ($\mathbf{x}$). Furthermore, we may be interested in whether there is variation in this elasticity across car manufacturers, leading to market power. One specific example might be the notion of a *home bias*, i.e. that consumers may prefer domestic to foreign products. Figure 1 shows a simple descriptive graph that motivates why we might suspect home bias to be strong. The figure shows the market shares of four selected car manufacturers in each of the five countries from the datset. A clear pattern emerges: each manufacturer has by far the strongest sales in its home market. Belgium serves

[2] In practice, N primarily grows as we get more time periods rather than new countries. For this assignment, we are agnostic about that.

Figure 1: Home Bias and Car Market Shares



as a neutral comparison since that country has no domestic car manufacturing (at least not during the sampling period).

2 Data

The dataset `cars.csv` has one row for five countries over 30 years (1970–1999). We continue to refer to a country-year pair as a "market," here indexed by $i = 1, \ldots, N$, where $N = 150 (= 5 \cdot 30)$. From the raw underlying dataset, the $J = 40$ highest-selling cars cars have been selected; that is, there are always precisely $J = 40$ discrete car alternatives available in any market i. Thus, there are $NJ = 6,000$ rows in the dataset `cars.csv`.

There are a total of 85 explanatory variables in the dataset, including several ways of measuring the "price" of a car, a list of technical car attributes (weight, horsepower, etc.), and several possible ways of classifying cars based on the "class" (sedan, SUV, etc.), manufacturer, country of manufacturer, etc.

3 Assignment

Using the dataset `cars.csv`, determine the strength is the "home bias" in the demand for cars, and analyze how it affects the own-price elasticity of demand.

4 Hints

(1) Appendix A contains a list of all available variables. We suggest using the following but encourage independent ideas:

   (1) `np.log(princ)` as the price variable,

   (2) Apart from the price, we recommend including as core car characteristics at least: `we`, `li`, `hp`, `home`,

   (3) we recommend that you consider whether to control for dummies at some level, e.g. the brand (the variable `brd`).

   Regardless of what you choose, argue for your choice. Beware that some of the variables cannot be included in the model because their associated coefficients would not be identified. (You are encouraged to give an example of such a variable and explain, in a few sentences, the reasoning for non-identification).

(2) There are different ways of quantifying the magnitude of an effect. Natually, one can use partial effects. Alternatively, a non-price variable, such as the home-dummy, can be converted to monetary terms. This is possible if there is a price variable in the model: then we can calculate how much the price must change in order to keep constant the observed part of utility of that car, $\mathbf{x}_{ij}\boldsymbol{\beta}$, of course for a specific value of $\mathbf{x}_i$.[3]

(3) Remember to properly define all variables and symbols employed and distinguish between them. For example, you should distinguish between the true parameter and an estimate thereof. Strive to employ the notation used in the course/textbook. Make use of boldface and capitalization to avoid confusing scalars, vectors and matrices. Specify dimensions whenever confusion may arise.

(4) When using an estimation procedure, carefully discuss the assumptions required to derive the estimator and establish properties thereof, and assess whether these assumptions are likely to be satisfied in the current empirical setting. If not, what are the consequences for the estimator in question (and your results)? Strive to provide a real-world example of behavior that might invalidate a given assumption, carefully linking the behavior or mechanisms to the mathematical symbols in the model.

(5) If you come up with several model specifications and associated estimates, discuss which one seems the most appropriate and justify your decision (e.g., based on formal testing).

[3] This can be interpreted as the *compensating variation* associated with changing the non-price attribute.

(6) Be precise about the statistical tests you use for testing various hypotheses. Explain which null hypothesis you are testing and the alternative you are testing against, how the test statistic is constructed, the decision rule you employ, and the conclusion you reach. If a variance (matrix) has been estimated, discuss the assumptions invoked for consistency. If several choices are possible, justify your choice.

5 Formal Requirements

- You must hand in a report that presents the econometric model, presents your estimation results and results of formal statistical tests (including interpretation and statements on economic and statistical significance), and discusses the potential weaknesses of the model, data and approach. If you present many estimates of the same parameters (e.g. estimators based on different asumptions, or varying the controls or sub-sample used), it may be helpful to present the estimates together in one table to facilitate comparison.

- The report must be written in English using an academic language and uploaded to Peergrade via Absalon as a single PDF file.

- The report must be at most five pages of main text (including mathematics) plus at most two pages of output.

  - For the main text (and mathematics), you must use fontsize = 12p, line spacing = 1.5, and 2.5 cm page margins (as used in this document).
  - The ouput can be any (relevant) tables or graphs as long as they properly formatted and labelled. Place the output at the end of your report, starting on a new page, and link to it from the main text when relevant.

- Along with your report, you must upload a compressed zip-folder with all the Python code needed to replicate your results. Make sure that your code is transparent and runs with only minor modifications. (Make sure to use relative paths, i.e. `./input/cars.csv` and not `C:\user\long\path`.) There is no character limit on the submitted Python code.

- You are allowed (and strongly encouraged) to work in groups of up to three people. List all group members on the front page of your report in alphabetical order of surnames.

- The assessment criteria are given on the course website in Absalon.

A Variable Labels

Variable	Label	Mean
ye	year (=first dimension of panel)	84.50
ma	market (=second dimension of panel)	3.00
co	model code (=third dimension of panel)	207.50
zcode	alternative model code (predecessors and succe...	177.76
brd	brand code	16.79
type	name of brand and model	NaN
brand	name of brand	NaN
model	name of model	NaN
org	origin code (demand side, country with which c...	2.72
loc	location code (production side, country where ...	5.17
cla	class or segment code	2.30
home	domestic car dummy (appropriate interaction of...	0.32
frm	firm code	14.50
qu	sales (number of new car registrations)	35606.68
cy	cylinder volume or displacement (in cc)	1337.09
hp	horsepower (in kW)	50.10
we	weight (in kg)	934.49
pl	places (number, not reliable variable)	4.88
do	doors (number, not reliable variable)	3.55
le	length (in cm)	409.24
wi	width (in cm)	163.44
he	height (in cm)	140.46
li1	measure 1 for fuel efficiency (liter per km, a...	6.59
li2	measure 2 for fuel efficiency (liter per km, a...	8.11
li3	measure 3 for fuel efficiency (liter per km, a...	8.92
li	average of li1, li2, li3 (used in papers)	7.87
sp	maximum speed (km/hour)	154.22
ac	time to acceleration (in seconds from 0 to 100...	16.27
pr	price (in destination currency including V.A.T.)	2608988.58
princ	=pr/(ngdp/pop): price relative to per capita i...	0.76
eurpr	=pr/avdexr: price in common currency (in SDR t...	7256.92
exppr	=pr/avexr: price in exporter currency	600384.65
avexr	av. exchange rate of exporter country (exporte...	229.30
avdexr	av. exchange rate of destination country (dest...	319.91
avcpr	av. consumer price index of exporter country	492.53
avppr	av. producer price index of exporter country	671.77
avdcpr	av. consumer price index of destination country	77.02
avdppr	av. producer price index of destination country	87.69
xexr	avdexr/avexr	66.32
tax	percentage VAT	0.21
pop	population	49183800.00
ngdp	nominal gross domestic product of destination ...	178667304825541.97
rgdp	real gross domestic product	216716720230172.44
engdp	=ngdp/avdexr: nominal gdp in common currency (...	504371708122.45
ergdp	=rgdp/avexr	659002565347.56
engdpc	=engdp/pop: nominal gdp per capita in common c...	10015.84
ergdpc	=ergdp/pop	13264.84
s	market share (qu / qu_tot)	0.02
qu_tot	total sales in this market-year	1424267.29
inc	avg. income per capita	26829.74

B Coding Hints

B.1 Max-rescaling

If you compute logit choice probabilities, remember to use the "max rescaling" trick. That is, utilizing that

$$\frac{\exp(v_{ij})}{\sum_{k=1}^{40}\exp(v_{ik})} = \frac{\exp(-K_i)}{\exp(-K_i)}\frac{\exp(v_{ij})}{\sum_{k=1}^{40}\exp(v_{ik})} = \frac{\exp(v_{ij}-K_i)}{\sum_{k=1}^{40}\exp(v_{ik}-K_i)},$$

which holds for any scalar $K_i \in \mathbb{R}$. While the equation holds analytically, the numerical implementation may be a different story, particularly because exponential functions can be prone to roundoff errors. It is therefore useful to set $K_i := \max_{\ell\in\{1,\dots,40\}} v_{i\ell}$ since then $\tilde{v}_{ij} := v_{ij} - K_i \leqslant 0$ for all j so that we can only get downwards roundoff errors, which are more well-behaved than upwards ones. (Here: $v_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}$.)