Project 3: High-Dimensional Models and Cars

Due: Sunday December 6th, 2020 at 22:00

1 A Simple High-Dimensional Model

Let $\{(Y_{i1}, Y_{i2})\}_{i=1}^{N}$ be N independent pairs of random variables with ith pair satisfying

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right), \quad i = 1, \ldots, N,$$

and the μ_i's and $\sigma^2 > 0$ are unknown parameters (i.e., constants) to be estimated.

(1) Derive the probability density function of the ith pair (Y_{i1}, Y_{i2}).

(2) Derive the log-likelihood function of $\{(Y_{i1}, Y_{i2})\}_{i=1}^{N}$ as a function of the μ_i's and σ^2.

(3) Solve analytically for the maximum-likelihood estimators (MLEs) $(\{\widehat{\mu}_i\}_1^N, \widehat{\sigma}^2)$ of the μ_i's and σ^2.

(4) What is the bias[1] of $\widehat{\mu}_i$ in estimating μ_i? What happens to the bias as $N \to \infty$?

(5) Argue that $\widehat{\mu}_i$ converges in probability as $N \to \infty$ and derive its probability limit. Comment on your findings.

(6) What is the bias of $\widehat{\sigma}^2$ in estimating σ^2? What happens to the bias as $N \to \infty$?

(7) Argue that $\widehat{\sigma}^2$ converges in probability as $N \to \infty$ and derive its probability limit. Comment on your findings.

(8) How does consistent estimation of the variance σ^2 relate to the construction of valid standard errors in the context of fixed-effects estimation of the standard linear model for panel data with individual-specific fixed effects discussed at the beginning of the semester? Feel free to make (strong) assumptions with respect to the model error components.

[1] The *bias* of an estimator $\widehat{\theta}$ of a parameter θ is defined as $\mathrm{bias}(\widehat{\theta}) := \mathrm{E}[\widehat{\theta}] - \theta$.

1.1 Hints

You may find the following facts about normally distributed random variables helpful.

- If $X \sim \mathrm{N}(\alpha, \beta)$ and $Y := cX + d$ for constants α, β, c and d, then $Y \sim \mathrm{N}(c\alpha + d, c^2\beta)$.
- If X_1 and X_2 are two independent normal random variables with means α_1 and α_2 and variances β_1 and β_2, respectively, then $X_1 + X_2 \sim \mathrm{N}(\alpha_1 + \alpha_2, \beta_1 + \beta_2)$.
- Normal random variables are independent if and only they are uncorrelated.

2 Car Demand with Market-level Data

The accompanying dataset, `cars.xslx`, contains data on the car market for $N = 5$ major European countries (Belgium, France, Germany, Italy, and the UK) over $T = 30$ years (1970–1999), covering a total of $J = 356$ unique car types. Throughout, let $i = 1, ..., N$ denote countries, $t = 1, ..., T$ denote years, and $j = 1, ..., J$ denote cars. The data are a balanced panel of countries (i, t), but the set of cars available, $\mathcal{J}_{it}$, differs over time and across countries.[2]

A total of 228 million cars were sold during this period, totaling 170.9 billion EUR (nominal) worth of sales. A full description of the variables in the dataset is given in Table 1. The variables come from Frank Verboven's website and have been used in Goldberg and Verboven [2001, 2005], Brenkers and Verboven [2006], with exception of the three last variables, which are from the World Inequality Database, `wid.world`.

Throughout this project, you should focus on *static models* (e.g. no leads or lags of variables), and focus your presentation around the following two key variables: home-market advantage and price. The first, the dummy `home`, is equal to one if the car is domestic, which may help explain why during the sample period, Fiat sold 7 times as many cars per capita in its home country of Italy as in France, whereas Citroen sold 4 times as many cars per capita in its home of France as in Italy. VW on the other hand sold roughly the same number of cars per capita in France and Italy. The second key variable is the (log) price of the car. The dataset contains several price variables, and you should pick one and briefly argue for your choice. We will be primarily interested in the price elasticity of demand and progressing towards question 6: should a car manufacturer charge a higher price in the home market versus abroad.

[2]Some cars are, however, sold very infrequently (i.e. in only a few years and/or markets). You are welcome to restrict the sample based on such a criterion, and you are encouraged to start with this before proceeding.

To start out, consider the following linear model:

$$\log q_{itj} = \mathbf{z}'_{itj}\gamma + \epsilon_{itj}, \tag{1}$$

where q_{itj} denotes the number of cars of type j sold in country i in year t, $\mathbf{z}_{itj}$ is a vector of covariates, and ϵ_{itj} is an idiosyncratic error term.

1. Estimate γ, both using Pooled OLS. Also estimate a version that controls for dummies for i, t, and j (separately and jointly). Choose a sample, what to include in $\mathbf{z}_{itj}$, and discuss identification of the coefficient on price and the home dummy. Briefly comment on your results in light of your insights from Part 1.

Next, consider a discrete choice model of car purchases. Assume that consumers (denoted c) derive (unobserved) utility from choosing car j on the form

$$u_{citj} = \mathbf{x}'_{itj}\beta_c + \varepsilon_{citj}, \quad \varepsilon_{citj} \sim \text{IID Extreme Value Type I}, \tag{2}$$

where $\mathbf{x}_{ijt}$ are attributes of car j in country i at time t. Let $\mathcal{J}_{it} \subset \{1, 2, ..., J\}$ denote the subset of cars that are available for purchase in country i at time t. You are allowed to assume that coefficients are homogenous in the populations, $\beta_c = \beta$. Further assume that consumer c chooses the car $y_{citj} = \arg\max_{j \in \mathcal{J}_{it}} u_{citj}$. Given the assumption on the distribution of the error term, the probability that a given car $j \in \mathcal{J}_{it}$ gets chosen is

$$\Pr(y_{citj} = j | \mathbf{X}_{it}; \beta) = \begin{cases} \frac{\exp(\mathbf{x}'_{itj}\beta)}{\sum_{k \in \mathcal{J}_{it}} \exp(\mathbf{x}'_{itk}\beta)} & \text{if } j \in \mathcal{J}_{it} \\ 0 & \text{if } j \notin \mathcal{J}_{it}. \end{cases} \tag{3}$$

where $\mathbf{X}_{it}$ is the stacked matrix of the attributes of all cars available, $\{\mathbf{x}_{itj}\}_{j \in \mathcal{J}_{it}}$. Note that the choiceset, $\mathcal{J}_{it}$, is varying over time and countries, and in particular there is no car j available for all i, t.[3] Furthermore, a specific car attribute may vary over time (e.g. if manufacturers make minor improvements to the engine over time) and/or across markets (e.g. the price).

2. Compare and contrast (1) with (2): what do the models assume about the own- and cross-price sensitivity of demand? How does total quantity demanded depend on price (and other factors)? What type of variables can be included in (1) and (2)?

3. Motivate an OLS estimator that allows you to consistently estimate β and compute it.

[3] In fact, only two cars, the Ford Escort and the Rover Mini, have positive sales in all years (and only in the UK). Most other car types are only available for a few years after they are invented, before being replaced by better models.

[Hint: Try to take log of choice probabilities and take differences between two cars $j, k \in \mathcal{J}_{it}$. Does it matter how the two cars, j and k, are chosen?]

In the subsequent questions, it may be easiest to start by estimating a model with a quite low-dimensional vector of attributes, $\mathbf{x}_{itj}$, which makes numerical optimization much simpler. Hence, you can keep any discussion about dummies at the i, t, or j level to the linear specifications and focus on other aspects of the non-linear model in the following.

Next, define the market share of car j in country i at time t as

$$s_{ijt} \equiv \frac{1}{C_{it}} \sum_{c=1}^{C_{it}} \mathbf{1}\{\text{car } j \text{ was purchased}\},$$

where there were C_{it} is the total number of consumers that bought a car in country i in year t. Consider the following estimator:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^K} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j \in \mathcal{J}_{it}} \left[s_{ijt} - \Pr(j|\mathbf{X}_{it}; \beta) \right]^2 . \tag{4}$$

4. Present an argument for consistency of $\hat{\beta}$ and compute the estimator and, if possible, standard errors.

   [Hint: If your optimizer gets stuck, try switching between `fminsearch` *and* `fminunc`*. Make sure that you are "max rescaling" your utility values before taking exponential values to ensure numerical stability (as in the exercises for Week 8). If you have trouble inverting your estimated Variance matrix, it may be because your optimizer has not actually found a stationary point but has terminated prematurely (giving the final message,* `local optimum possible`*).]*

We now turn to the **own-price elasticity** of demand. The logit model gives rise to conditional choice probabilities for each car, $\Pr(j|\mathbf{X}_{it})$, so the typical measure of the own-price elasticity of a car is

$$e_{itj} \equiv \frac{\partial \Pr(j|\mathbf{X}_{it})}{\partial x_{itjk}} \frac{x_{itjk}}{\Pr(j|\mathbf{X}_{it})}, \tag{5}$$

where the k'th variable in $\mathbf{x}_{itj}$ is the price. The elasticity in e_{itj} can be computed by taking the derivative of the choice probability in (3), or it can, perhaps more easily, be computed numerically. To compute it numerically, simply evaluate the choice probability at two different values of the covariates, $\mathbf{X}_{it}$ (the baseline) and $\tilde{\mathbf{X}}_{it}$, where $\tilde{\mathbf{X}}_{it}$ is equal to $\mathbf{X}_{it}$ everywhere, except that $\tilde{x}_{itjk} = 1.01 x_{itjk}$. That is, the price has been increased by 1%. The numerical approximation of the elasticity is then simply the percentage change in probability caused

by this one percent change in price:

$$e_{itj}^{\text{num}} = \left(\frac{\Pr(j|\tilde{\mathbf{X}}_{it})}{\Pr(j|\mathbf{X}_{it})} - 1 \right) \Big/ 0.01.$$

5. What is the own-price elasticity of demand? And how is it related to home-market advantage?

   [Hint: Note that even without heterogeneity in the coefficient on price, the elasticity varies depending on the value of $x_{itj\ell}$ for non-price attributes $\ell \neq k$. You are encouraged (but not required) to try to estimate a model with heterogeneity in the price coefficient, either with interactions of variables or in a random coefficient model.]

6. Do firms charge a higher or lower price on cars in their home market? How do you judge their pricing decision?

   [Hint: The second part of the question is intended as an opportunity to be creative. You are welcome to make bold assumptions, or to draw parallels to overly simplistic microeconomic models (e.g. how a monopolist would approach pricing depending on the price elasticity of demand).]

7. What is the greatest limitation of your analysis, and how might it be remedied?

Formal Requirements

- You must hand in a report containing your answers to all parts of the assignment.
- The report must be written in English and uploaded to Peergrade via Absalon as one single PDF file.
- You must obey the following page constraints:
  - Part 1: No constraints,
  - Part 2: Maximum four normal pages of but no limit on the amount of output in the form of tables and/or figures. [4]
- You are allowed to work in groups of up to three people (not necessarily in the same exercise class as yours). List all group members on the front page of your report.

[4] One normal page of text is defined as: Font size = 12p, 1.5 line spacing, and margins of 2.5 cm.

Table 1: Variables in the dataset `cars.csv`

Variable	Description
ye	year (=first dimension of panel)
ma	market (=second dimension of panel)
co	model code (=third dimension of panel)
zcode	alternative model code (predecessors and successors get same number)
brd	brand code
type	name of brand and model
brand	name of brand
model	name of model
org	origin code (demand side, country with which consumers associate model)
loc	location code (production side, country where producer produce model)
cla	class or segment code
home	domestic car dummy (appropriate interaction of org and ma)
frm	firm code
qu	sales (number of new car registrations)
cy	cylinder volume or displacement (in cc)
hp	horsepower (in kW)
we	weight (in kg)
pl	places (number, not reliable variable)
do	doors (number, not reliable variable)
le	length (in cm)
wi	width (in cm)
he	height (in cm)
li1	measure 1 for fuel efficiency (liter per km, at 90 km/h)
li2	measure 2 for fuel efficiency (liter per km, at 120 km/h)
li3	measure 3 for fuel efficiency (liter per km, at city speed)
li	average of li1, li2, li3 (used in papers)
sp	maximum speed (km/hour)
ac	time to acceleration (in seconds from 0 to 100 km/h, some from 0 to 96 km/h)
pr	price (in destination currency including V.A.T.)
princ	=pr/(ngdp/pop): price relative to per capita income (often used in demand model)
eurpr	=pr/avdexr: price in common currency (in SDR times 1.2956 to interpret in Euros)
exppr	=pr/avexr: price in exporter currency
avexr	av. exchange rate of exporter country (exporter 'loc' currency per SDR)
avdexr	av. exchange rate of destination country (destination 'ma' currency per SDR)
avcpr	av. consumer price index of exporter country
avppr	av. producer price index of exporter country
avdcpr	av. consumer price index of destination country
avdppr	av. producer price index of destination country
xexr	avdexr/avexr
tax	percentage VAT
pop	population
ngdp	nominal gross domestic product of destination country (destination currency)
rgdp	real gross domestic product
engdp	=ngdp/avdexr: nominal gdp in common currency (SDR)
ergdp	=rgdp/avexr
engdpc	=engdp/pop: nominal gdp per capita in common currency
ergdpc	=ergdp/pop
IncShare_p0p50	Income share for the percentiles [0;50)
IncShare_p50p90	Income share for the percentiles [50;90)
IncShare_p90p100	Income share for the percentiles [90;100]
NationalInc	National income per capita

- The assessment criteria are posted on the course page on Absalon.

References

Randy Brenkers and Frank Verboven. Liberalizing a distribution system: the european car market. *Journal of the European Economic Association*, 4(1):216–251, 2006.

Pinelopi K Goldberg and Frank Verboven. Market integration and convergence to the law of one price: evidence from the european car market. *Journal of international Economics*, 65(1):49–73, 2005.

Pinelopi Koujianou Goldberg and Frank Verboven. The evolution of price dispersion in the european car market. *The Review of Economic Studies*, 68(4):811–848, 2001.