No More Free Choice: An Empirical Analysis of Default Options in the Amazon Marketplace[*]

Guido Deiana

November 29, 2021

Abstract

The purpose of this work is to estimate the impact of products suggestions in an online marketplace on consumer choice. Specifically, we focus on the Amazon marketplace and we study how a platform suggestion (the "Amazon's Choice" badge) affects demand. We collected data on usb cables sold in the five Amazon Europe marketplaces and from numerous sources we obtained the time series of prices, reviews, and of a proxy for sales. Then, we developed a discrete choice model in the characteristics' space. Following the Empirical Industrial Organization literature, the model is then estimated using a Nested Logit and a Random Coefficient Logit (BLP). The models show that the "Amazon's Choice" badge has an extremely large, positive, and statistically significant effect on demand. These results show the potential for distortion of consumers' choice coming from a vertically integrated platform which competes with third party sellers in its own environment.

[*] Note: What follows is a shorter version of my MS Thesis. It reports the main findings and the contributions of this work, while shortening the extensive literature review. I am grateful to Professor Francesco Decarolis (Bocconi University) for the guidance, the comments and the suggestions provided. Without his help, this work would not have been possible.

1 Introduction

In the early 2000s the online platforms were ruled by Ebay, which only acted as an intermediary between sellers and potential buyers. Today, Amazon dominates the online shopping platforms in Europe and in the United States, thanks to a system of customer loyalty (Amazon Prime) and a peculiar positioning. Amazon is the owner of the platform through which numerous of third party sellers work, as well as being a large seller and producer of products that are then directly sold through its own platform. The peculiar positioning of Amazon, together with its large market power, has let to concerns from the European and US Competition authorities regarding the potential abuse of dominance against third party sellers[1]. The platform owner could in fact interfere with the market in at least two ways: (i) use buyers' and sellers' data to strategically enter the market where there is potential profit to be made (ii) distort customers' demand towards certain products directly provided (produced) by itself. The combination of the two leads to effective foreclosure of the market towards third party sellers that try to compete with Amazon itself on its own platform.

This paper focuses on a specific cue that Amazon includes in the products research page and in how effective it can be at distorting demand. More specifically, it is built on the assumption that buyers in online marketplaces are often in a situation called "Choice Overload". As explained by Spiegler (2011): "The larger the set of market alternatives, the greater consumers' tendency to evade the choice problem and adhere to the default". In this setting, the abundance of similar products available to the buyer can be exploited by the platform to suggest a default product which customers tend to buy to avoid having to compare all the others. This paper naturally builds on the literature of the applications of Behavioral Economics models in an Industrial Organization setting (Ellison, 2006; Spiegler, 2011) and more specifically of situations in which consumers inertia is a significant factor when agents chose which product to buy, as in the electricity suppliers' market (Hortaçsu, Madanizadeh, and Puller, 2017) and in the insurance providers' one (Brown and Jeon, 2020).

The nudge analyzed in this paper is the "Amazon's Choice" badge (from now AC) that Amazon started including in almost every product research from 2015, which is algorithmically assigned to a single product for each relevant combination of words used in the search bar.[2] The badge is often seen as a form of quality assurance by the customers and this paper argues that it plays a fundamental role in

[1] On July 17, 2019 the European Commission opened a formal investigation to assess whether the use of sellers' data from Amazon is a breach of EU competition rules. See: European Commission (2019)

[2] The assignment of the Amazon's Choice badge is an opaque process that has led to speculations and articles regarding the potential unfairness of the process (Shifflett, Berzon, and Mattioli, 2019).

determining which product customers will ultimately choose, especially when comparing similar low cost articles. Other works have shown a statistically significant effect on consumers satisfaction of systems of sellers reputation in online markets (Saeedi, 2014; Tadelis, 2016; Hui et al., 2016).

To quantify the impact of the AC badge on consumers' demand, a simple discrete choice demand model is developed, with each customer buying only one product from Amazon according to its own utility maximization function. The AC badge enters the system as a characteristic of the product, with linear effect on the agents' utilities. The demand coefficients are individual varying. To estimate the model, this work uses modern techniques from the Empirical Industrial Organization literature. More specifically the Random Coefficient Logit model (Berry, Levinsohn, and Pakes, 1995; Nevo, 2001) and the Nested Logit model (Goldberg and Verboven, 2001). We also draw extensively from the application of the Random Coefficient Logit model in Decarolis, Polyakova, and Ryan (2020b).

From the five Amazon Europe marketplaces (Italy, France, Spain, Germany, United Kingdom), we derive a sample of more than 1000 electronic cables divided in five sub-categories (dvi, ethernet, hdmi, lightning, usb). Electronic cables are chosen for the following reasons: they have distinct and easily recognizable characteristics, both from the consumer's and from the researcher's standpoint; cables with the same characteristics are very close to being perfect substitutes; the mapping of the products in the Amazon subcategories is consistent with the cables end use, leading to a reliable market size estimation of cables sold on the Amazon marketplace. A web scraper is used to create the sample and gather static information about the products. An external source[3] is then used to gather historical sales and prices data for each of the selected products. From these sources, current data on products characteristics and AC badge is derived, as well as current and historical data on prices, sales, number of reviews and ratings of products.

The AC badge shows a strong and statistically significant effect on consumer demand across all the models. The order of magnitude of the AC badge is also strikingly large, showing a high potential for demand distortion coming from the platform. These results have important policy implications in the fields of Competition Economics and Antitrust. The badge could in fact bring an unfair advantage to Amazon when it competes with other sellers upstream. So far its badge assignment has been an obscure mechanism which could be advantageous for certain sellers and discriminatory for others. This work also shows the potential for more general subtle demand distortions in online platforms, which could also happen in other forms. Ultimately, the evidence presented supports the argument that in marketplaces with numerous similar products, when too many choices are available, customers can fail to optimize and often prefer following a generic default choice.

[3] `https://keepa.com/#!api`

This work contributes to the literature of Competition Economics papers in online platforms and on the exploitation of limited consumers' attention: Prat and Valletti (2021) consider consumer attention as a scarce resource used as a product from online platforms (referred to as "attention brokers") to sell to advertising companies. Also Bordalo, Gennaioli, and Shleifer (2013, 2015) start from the assumption that consumers' attention is limited and can be drawn to a limited set of salient characteristics from the sellers. Other works relate limited consumers' attention to their failure to chose the optimal option (Grubb, 2015; Bronnenberg, Dubé, Gentzkow, and Shapiro, 2015). This paper also builds more generally on other works on competition and market dynamics in online platforms (Decarolis and Rovigatti, 2021; Decarolis, Goldmanis, and Penta, 2020a).

The remainder of this work is structured as follows. Section 2 presents the main data sources and sampling techniques used. Section 3 presents the theoretical models of discrete choice estimated in the following sections. Section 4 and 5 show and explain the empirical models used and the identification strategies implemented. Section 6 covers the results of the empirical models presented in the previous sections. Section 7 draws the conclusions of the work. The appendix at the end covers the steps for the estimation of the Random Coefficient Logit model.

2 Data

2.1 Sources

The dataset used in this work is comprised of a subsample of Amazon's products. The main goal in samlpling the correct subset of products was finding categories of products which were largely substitutable and with few easily identifiable characteristics. To do so, the market for electronic cables is selected and 5 categories of cables (usb, hdmi, vga, lightning, dvi) are analyzed. Amazon products are all assigned to a macro-category and numerous sub-categories. For this work, the 5 most populated subcategories of cables, all of which belong to the Electronics or Informatics macro-categories. The macro-categories are useful in the dataset construction as each product is assigned a "Best Seller's Rank" (BSR) which ranks all the products in a category and in a geographical market according to how much quantity they are selling in the current month. The BSR is of crucial importance in trying to derive a reliable estimate of sales. From each category the data on the 100 most sold products are gathered for the same 5 categories in the 5 Amazon Europe markets (Italy, France, Germany, Spain, United Kingdom). There are some products that are repeated among the different markets, but sometimes they display different characteristics (like the AC badge which is market specific) and prices. To collect the sample URLs and static information, such as the AC badge and other product characteristics, we coded and ran a web scraper on the five European Amazon marketplaces. We collected data on the 100 most sold products in each sub-category and in each geographical market.

For the time-series data collection, this work relies on the keepa api database[4]. Keepa is a website which tracks prices, BSR, reviews and average score of all products available on Amazon Marketplace as well as Ebay prices for comparison (when an ebay alternative seller is available). From Keepa, the historical data of all the analyzed products is downloaded and then the BSRs and the prices are averaged on a monthly basis to mitigate the impact of isolated data collection errors, which sometimes were found in less popular products.

In order to estimate the correct market shares, the BSRs of products are not enough as there isn't a linear relationship between BSR and monthly sales. However, since BSR is a measure of products' popularity, there are estimates of monthly sales available which are constructed from BSR, category and market. The ones used here come from the website Junglescout.[5] Other estimators tend to give similar estimates, but this work refers only to Junglescout as it is the quickest and most reliable to work with. Amazon does not directly reveal how the BSR is assigned, as for the AC badge, however it is broadly defined as an index of the most sold products within a category. The BSR is calculated on the basis of recent and historical sales, with more weight on the recent. The approach used is to average the BSR on a monthly basis, and then calculate sales according to the average BSR, this should give a rough estimate of the sales of a product, without being influenced by sudden spikes in the sales. This approach has the potential issue of not estimating the correct market shares. This could lead to major problems for the Logit approach as the Logit model does not allow errors in the market share definition and can be highly biased when products display market shares close to zero. To mitigate this issue, all products with a small market share (threshold set at 0.05%) are dropped from the sample. These are the products more likely to create biases due to low market shares as well as being the ones more likely to have a relatively higher estimation error in their market share. With this approach, the aim is to eliminate the small market share bias, as well as mitigate the errors in market shares calculation.[6]

2.2 Descriptive Statistics

To assess the correlations between the AC badge and the product popularity, we ran a series of OLS regressions of Estimated Monthly Sales on products characteristics. The purpose is not to assess any causal relationship between characteristics and

[4] `https://keepa.com/#!api`

[5] `https://www.junglescout.com/estimator/`

[6] An alternative solution for the issue of market shares calculated with an error term is proposed by Gandhi, Lu, and Shi (2020). But we are not aware of any application of it to real world data.

sales, but to simply assess the magnitude of the correlation between the two. The equation estimated is:

$$Sales_j = -\alpha p_j + \mathbf{x}_j' \boldsymbol{\beta} + \varepsilon_j \tag{1}$$

Two approaches are proposed to assess the correlations between variables (Table 1). In the first one, the sample is divided in the five different markets analyzed. In the second one, the sample is divided in five different nests of products analyzed.

Table 1: OLS Regression of Monthly Sales

			By Countries:		
	UK	Germany	Spain	France	Italy
AC	218.159***	164.978***	174.841***	220.540***	326.573***
	(39.760)	(38.688)	(35.131)	(34.546)	(38.156)
Bundle	50.708	95.266*	229.845***	20.277	164.088***
	(48.709)	(53.495)	(67.473)	(72.561)	(44.702)
Price	-5.703**	-0.835	-7.438***	-6.799***	-4.818**
	(2.851)	(1.944)	(2.653)	(2.177)	(1.986)
Rating*Reviews	0.030***	0.016***	0.000	-0.001	0.002*
	(0.004)	(0.004)	(0.001)	(0.001)	(0.001)
Constant	202.000***	212.949***	293.943***	278.024***	223.992***
	(43.394)	(40.151)	(36.685)	(36.498)	(36.065)
Observations	409	361	355	362	427
R-squared	0.203	0.095	0.123	0.129	0.236
			By Nests:		
	Dvi	Ethernet	Hdmi	Lightning	Usb
AC	-27.473	11.546	21.226	176.818***	367.718***
	(31.463)	(16.644)	(24.771)	(47.407)	(27.045)
Bundle	-15.871	-2.177	27.663	60.478	161.267***
	(81.261)	(23.609)	(44.749)	(62.052)	(34.794)
Price	4.612	-0.820	1.316	-5.289	-5.247***
	(2.841)	(0.964)	(1.510)	(4.394)	(1.411)
Rating*Reviews	0.055***	0.006**	0.000	0.001	0.016***
	(0.014)	(0.003)	(0.001)	(0.002)	(0.003)
Constant	122.603***	223.991***	239.241***	485.077***	150.840***
	(37.967)	(20.074)	(24.312)	(61.591)	(24.317)
Observations	61	308	324	366	855
R-squared	0.244	0.019	0.007	0.043	0.305

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: the table reports correlations of the product characteristics and their monthly sales. The correlations are calculated by nests (categories) and by countries.

The results show a persistent large positive correlation between AC badge and quantities of products sold in a month. The impact tends to fluctuate a lot between countries as well as between different groups of products. This approach does not allow to draw any causal relationship between the variables, it however shows that AC products are consistently more popular than the rest of the products. This was the expected result indeed and is useful in testing the quality of the sample proposed.

3 Theoretical Model

The theoretical model presented is a standard discrete choice model, in which agents choose the product which gives them the highest utility, accounting for agent specific variations in the coefficients of the products characteristics. The utility function is linear in products characteristics, with individual varying coefficients. The default choice bias is therefore integrated in this traditional, utility maximizing specification, by considering the AC badge as a standard characteristic.

The Utility function of agent i, who chooses among a set of J products is expressed as:

$$\mu_{i,j} = \alpha_i(y_i - p_j) + \mathbf{x}_j'\boldsymbol{\beta}_i \quad \forall j \in J \tag{2}$$

Where y_i is consumer i's income, p_j is product j price and $\mathbf{x}_j$ the vector of product j characteristics, with dimensions Kx1. The α_i and $\boldsymbol{\beta}_i$ coefficients are individual specific. To simplify the analysis and to make the model tractable in the empirical part, the average utility is separated from the individual specific portion of utility.

$$\mu_{i,j} = \alpha(y_i - p_j) + \mathbf{x}_j'\boldsymbol{\beta} + \epsilon_{i,j} \tag{3}$$

Where $\boldsymbol{\beta}$ is the average $\boldsymbol{\beta}_i$ coefficient. and α is the average α_i coefficient. So that the average utility of product j is $\alpha(y_i - p_j) + \mathbf{x}_j'\boldsymbol{\beta}$, while $\epsilon_{i,j}$ is the individual deviation from the mean utility of product j.

Agent i will chose which product to buy according to the utility maximization problem:

$$\begin{aligned}
\max_{j\in\{0,1,...J\}} \quad & U_{i,j} = \mu_{i,j} - \mu_{i,0} \\
\text{sub:} \quad & \mu_{i,j} = \alpha(y_i - p_j) + \mathbf{x}_j'\boldsymbol{\beta} + A_j\beta_a + \epsilon_{i,j} \qquad \forall j \in \{1, 2, ...J\} \\
& \mu_{i,0} = \mu_0 + \epsilon_{i,0} \\
& \mu_0 \geq 0
\end{aligned} \tag{4}$$

Where A_j is the dummy that indicates whether the product is AC or not. Heterogeneity of consumers is accounted for by the term $\epsilon_{i,j}$. $\mu_{i,0}$ represents the utility of the outside option.

In the next section, the coefficients of the utility functions of the consumers are estimated. The previous utility function can be simplified for a cleaner notation.

$$\begin{aligned}
U_{i,j} &= \mu_{i,j} - \mu_{i,0} \\
&= \alpha(y_i - p_j) + \mathbf{x}_j'\boldsymbol{\beta} + A_j\beta_a + \epsilon_{i,j} - \mu_0 - \epsilon_{i,0}
\end{aligned} \tag{5}$$

μ_0 is accounted for in the constant term of our econometric equation and a single ϵ error term is considered for simplicity, so the variables μ_0 and $\epsilon_{i,0}$ are removed.

Finally, the utility of the agent buying product j is simply:

$$U_{i,j} = \mu_{i,j} = \alpha(y_i - p_j) + \mathbf{x}_j'\boldsymbol{\beta} + \epsilon_{i,j} \tag{6}$$

Where $\mathbf{x}_j$ includes the A_j AC dummy, and the utility of consumer buying the outside good can be simplified to $U_{i,0} = \mu_{i,0} = 0$.

4 Empirical Approach

The empirical approach for the paper follows the work of Berry (1994), Berry, Levinsohn, and Pakes (1995), and the more recent works on the matter of Nevo (2001) and Rasmusen et al. (2007).

A demand equation is estimated in which all consumers are considered to be perfectly rational and the AC dummy A_j is part of the characteristics set.

$$\mu_{i,j} = \alpha(y_i - p_j) + \mathbf{x}_j'\boldsymbol{\beta} + \beta_a A_j + \epsilon_{i,j} \tag{7}$$

To estimate the model, this work resorts to a two levels nested logit, and then a more complex Random Coefficient approach is also used to allow for interaction with demographic variables in the estimation of the demand coefficients. (Berry, Levinsohn, and Pakes, 1995).

4.1 Multinomial Logit

The simple multinomial logit approach is the starting point for all the models used in this paper. It starts with the utility of agent i buying product j from equation (7).

Assuming agents are fully rational, agent i will buy product j if and only if $\mu_{i,j} \geq \mu_{i,k} \forall k \neq j$. This allows to simplify equation (7) by noting that y_i is eliminated in considering the optimal choice for agent i. The term ε_j represents the mean deviation of agent i from the predicted utility of good j, so then $\epsilon_{i,j}$ now represents the individual deviation from the mean deviation from the predicted utility.

$$\mu_{i,j} = \mathbf{x}_j'\boldsymbol{\beta} - \alpha p_j + \varepsilon_j + \epsilon_{i,j} \tag{8}$$

Assuming $\epsilon_{i,j}$ i.i.d. and distributed as EV type 1 (Cardell, 1997), it is possible to find a closed form solution for the market shares of product j

$$\ln(S_j) - \ln(S_0) = \mathbf{x}_j'\boldsymbol{\beta} - \alpha p_j + \varepsilon_j \tag{9}$$

Unfortunately, due to the property of Independence of Irrelevant Alternatives

(IIA) [7] (Train, 2009), model 9 is only useful for some initial estimates, but is not fully robust and it is not possible to draw relevant conclusions from it.

4.2 Nested Logit

In the nested Logit approach, the products are divided in different groups (“Nests”), and the choice of consumers is modelled in the sequential manner that follows:

1. Consumers choose which Nest of products to buy.

2. Inside each nest, consumers choose by maximizing their utility function.

This approach requires the researcher to know the different categories in which the products are divided according to the customers. While in some cases it can be a limitation, in this setting the Amazon marketplace already divides its products in reliable subcategories. For this work, the 5 subcategories of usb cables sold on Amazon with the most products in them (usb, ethernet, dvi, lightning, hdmi) are used as the five nests of the analysis.

The separation of different products into nests is fundamental to model with more precision the consumer heterogeneity and to restrict the span of the IIA restrictions posed earlier. IIA property in fact now holds only within the single nests and not across the different nests.

The utility of agent i can now be expressed as:

$$\mu_{i,j} = \mathbf{x}_j'\boldsymbol{\beta} - \alpha p_j + \zeta_g + (1-\sigma)\epsilon_{i,j} + \varepsilon_j \tag{10}$$

Where ζ_g is the mean utility of all products belonging to the same group (or nest) $g \in G$. The probability that agent i buys product j now needs to be expressed in a sequential manner that reflects the nested form:

$$d_j = d_g * d_{j|g} \tag{11}$$

According to Berry (1994), by assuming that $\zeta_g + (1-\delta)\epsilon_{i,j}$ follows an i.i.d., EV1 distribution, d_g (the average probability that agents chooses nest g) and $d_{j|g}$ (the average probability that agents choose product j inside nest g) can be expressed as:

$$\begin{aligned} d_{j|g} &= \frac{exp\{\delta_j/1-\sigma\}}{D_g} \\ d_g &= \frac{D_g^{1-\sigma}}{\sum\limits_{g\in G} D_g^{1-\sigma}} \end{aligned} \tag{12}$$

[7] By construction, in the nested logit model, cross elasticities of demand are homogeneous. Meaning that when a characteristic of product k changes (say price), the quantities of all the other products $j \neq k$ react in the same manner: $\frac{\partial q_k}{\partial p_j} = -\alpha p_k d_k$. For this work, this assumption is unrealistic as different products have different purposes for the agents

Where:

$$D_g = \sum_{j \in g} exp\{\delta_j/(1-\sigma)\} \tag{13}$$

By expressing the outside good as a separate group, its predicted market share is:

$$d_0 = \frac{1}{\sum\limits_{g \in G} D_g^{1-\sigma}} \tag{14}$$

From (11) it is easy to derive that:

$$\ln(d_j) - \ln(d_0) = \frac{\delta_j}{1-\sigma} - \sigma \ln(D_g) \tag{15}$$

By adding $\sigma d_{j|g}$ and substituting the predicted with the observed market shares:

$$\ln(S_j) - \ln(S_0) = \sigma \ln(S_{j|g}) + \mathbf{x}_j'\boldsymbol{\beta} - \alpha p_j + \varepsilon_j \tag{16}$$

Where the coefficient σ is restricted between 0 and 1 and represents the relative importance of the division in nests against the products characteristics. The closer σ is to 1, the more it is important for customers the division in nests; the more it is close to 0, the more the individual products characteristics are important for consumers. With the limit case $\sigma = 0$, the two levels nested logit equation (16) becomes identical to the simple logit equation (9).

4.3 Random Coefficient Logit

The Random Coefficient Logit model allows to specify more precisely the individual specific portion of the Utility, resulting in more accurate estimates of the mean utility function δ_j, together with a second set of coefficients of products characteristics, interacted with individual random and demographic variables.

This setting starts again with the usual individual utility equation, but the coefficients α_i and β_i are individual specific this time, and the notation t indicates market $t \in T$.

$$\mu_{i,j,t} = \mathbf{x}_{j,t}'\boldsymbol{\beta}_i - \alpha_i p_{j,t} + \epsilon_{i,j} + \varepsilon_j \tag{17}$$

Instead of aggregating the individual specific portion of the demand in the variable $\epsilon_{i.j}$, or specifying group specific utilities as in the nested logit, now the individual portion of the coefficients is considered explicitly:

$$\begin{pmatrix} \alpha_i \\ \boldsymbol{\beta}_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \sum_\alpha \\ \sum_\beta \end{pmatrix} \boldsymbol{v}_i + \begin{pmatrix} \prod_\alpha \\ \prod_{\boldsymbol{\beta}} \end{pmatrix} \boldsymbol{D}_i \tag{18}$$

Where v is a kx1 vector of normal and randomly distributed variables and D_i

is the mx1 vector of demographic variables for agent i. $\sum$ is the kxk matrix of interactions between random variables v and products characteristics X, and $\prod$ is kxm matrix of interactions between demographic variablels D and products characteristics X. The inclusion of demographic variables was proposed in Nevo (2001) and has since become the standard for Random Coefficient Logit estimation.
From 18 it is possible to derive the following utility equation:

$$\mu_{i,j,t} = \delta_{j,t} + u_{i,j,t} + \epsilon_{i,j,t} \tag{19}$$

Where:

$$\begin{aligned} \delta_{j,t} &= \mathbf{x}'_{j,t}\boldsymbol{\beta} - \alpha p_{j,t} + \epsilon_{j,t} \\ u_{i,j,t} &= (-p_{j,t} \quad \mathbf{x}'_{j,t}) \sum \boldsymbol{v}_i + (-p_{j,t} \quad \mathbf{x}'_{j,t}) \prod \boldsymbol{D}_i \end{aligned} \tag{20}$$

And $\varepsilon_{i,j,t}$ is still assumed to IID with EV Type 1 distribution. This setting offers no closed form solution to extract the mean utility, like it was possible to do in the Logit approach. To estimate equation (19) it is therefore necessary to procede with an optimization method explained in the appendix.

5 Identification Strategy

5.1 The Endogenous Prices Problem

The problem of identification in a demand estimation setting rises immediately due to the nature of competition in free markets. Prices are in fact never exogenous, unless they are predetermined from a regulatory source, or they are randomly assigned.

This work exploits the availability past prices to instrument the current price levels. This approach is therefore a variation of what was proposed in Arellano and Bond (1991). Prices in every period are determined by the costs of production of the product (c), plus a reaction to the current demand (δ_t): when demand is high prices tend to increase, when demand is low prices tend to decrease. Since, at least in the short term, costs can be considered fixed, then past prices and current prices are highly correlated.

$$p_t = c + \gamma\delta_t + \varepsilon_t \tag{21}$$

This leads to past prices being a strong instrument for current prices.

It is however necessary to verify that past prices are exogenous and affect current demand only through the current prices. This last assumption (better known as the "Exclusion Restriction") is easy to verify as customers only see current prices and are not considering past prices when making a purchase decision.[8] Therefore, the

[8] A case can be made that explicit discounts that show both past prices and current prices

Exclusion Restriction is satisfied by construction of our model and there is no need to assume that past prices can be considered as a relevant characteristic, when customers decide to buy. Since past demand is correlated with current demand, the exclusion restriction could be invalidated if low prices lead to high demand in the past and then current demand stays high regardless of current prices. The only channel through which this could be possible is through reviews and AC selection: as products become more popular they receive better reviews and therefore the demand stays high even when prices go up again. Something similar could happen to a product that becomes popular due to low prices and then is given the AC badge, once the prices go up demand could stay high anyways due to the badge. These cases are treated by including all time varying characteristics in the econometric equation. By including current characteristics in the second stage of the instrumental variable approach, past prices can only affect current demand through current prices and current characteristics.

The last, and most important, caveat of this approach is the need to assume that errors in the demand function are not serially correlated. The main problem that drives the endogeneity of prices is reverse causality, prices, however, could also be correlated with omitted variables in the demand equation. it is therefore necessary to assume that the error term ε_j in the average demand equation is not serially correlated:

$$\delta_j = \mathbf{x}_j'\boldsymbol{\beta} - \alpha p_j + \varepsilon_j \tag{22}$$

5.2 Extending the Framework to All Endogenous Characteristics

In this framework, prices are not the only endogenous variable, unlike the standard applied Industrial Organization problems. In fact, in the Amazon online market, agents are exposed to other three endogenous characteristics: number of reviews, average rating of the reviews and AC badge.

The first two are interacted to construct the rating*reviews coefficient. The reason for this choice is that the rating is not particularly explicit in the Amazon Marketplace (being indicated only as stars) and overall not a high variance in ratings was found. Since agents see average rating and number of reviews one next to the other it can be assumed that they consider rating and reviews only in conjunction one with the other (a single 5 stars review is probably worth less than one hundred reviews with 4.5 stars on average). Rating*reviews is still endogenous and is

would indeed render the assumption false as the customer could be influenced by the amount of discount applied to the product, transforming the average portion of the demand equation in: $\delta_t = -\alpha_1 p_t + -\alpha_2(p_t - p_{t-1}) + \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t$. In this work however agents are considered to be mostly rational in their utility function and a discount applied to the product does not induce direct utility gains on a rational consumer therefore instrumented with a number of past lags of itself, starting from the second lag.

AC is the most difficult to instrument as there is no clear indication from Amazon on how it is assigned. However, it is reasonable to assume that more popular and better performing products are more likely to be assigned the AC badge than less popular and worse rated products. For that reason, AC is instrumented with the same lags of rating*reviews explained in the previous paragraph. More than one lag is included for this reason as well: at least one lag is necessary to instrument rating*reviews, at least another lag is necessary to instrument AC.

All the previous considerations on the validity of the price instruments hold here as well for the identification of the characteristics coefficients

6 Results

This analysis led to interesting and sometimes unexpected results on the demand estimation of Amazon products, what follows is an overview and comment on the main results of the model.

All the following models include the same characteristics X and the same set of instruments Z. The dataset includes the 2500 products which come from the five different Amazon EU markets (Italy, France, Spain, Germany, United Kingdom). From these five markets, the 100 most sold products in each of these electronics cables subcategories are taken (usb, hdmi, vga, lightning, ethernet). USB cables are considered the outside good, this is necessary to find a simple closed form solution for our nested Logit. The standard category of "usb" cables was chosen because it contains the most standardized products and is very likely that an agent buying in the "usb" nest is looking for any product that could fulfill his needs, without looking at the differentiation among products. The utility of the nest "usb" is then standardized to zero, while the mean utility of all other products is estimated. All products with a market share of less than 0.05% are then added to outside good. After excluding the outside goods, the sample used contains 1059 products.

6.1 Multinomial Logit

The first technique implemented is the simple Multinomial Logit, thanks to its easy closed form solution, it provides a valuable starting point for the analysis.

The equations that follow are estimated with a standard Generalized Method of Moments procedure, in which the independent variables are instrumented with the two months lag of Price and the two months, three months and four months lags of Rating*Reviews. The reason for this is to include one instrument for Price, one instrument for Rating*Reviews, at least one instrument for the AC badge and the

Table 2: Logit Estimates

	(1) Simple	(2) Simple	(3) Nested	(4) Nested
AC	0.0516	1.098*	-0.0224	1.174**
	(0.545)	(0.657)	(0.440)	(0.465)
Price	-0.00465	0.000541	-0.00449*	-0.000378
	(0.00321)	(0.00730)	(0.00242)	(0.00564)
Rating*Reviews	3.72e-06	-2.41e-06	4.95e-06**	-6.99e-07
	(2.46e-06)	(3.58e-06)	(1.99e-06)	(2.47e-06)
Bundle	0.135	0.116	0.222***	0.220**
	(0.0870)	(0.104)	(0.0703)	(0.0910)
σ			0.493***	0.631***
			(0.0299)	(0.0496)
Constant	-5.442***	-6.135***	-3.258***	-3.396***
	(0.289)	(0.471)	(0.271)	(0.289)
Observations	1,059	1,059	1,059	1,059
R-squared	0.015		0.291	
Hansen's J	1.922	0	4.218	0
Hansen p_value	0.166	1	0.0400	1
Brand FE		Yes		Yes

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

*Note: The table reports the results of the Simple (Multinomial) Logit and of the Nested Logit Models. Both models are estimated with the GMM approach using Arellano-Bond instruments for the variables AC, Price and Rating*Reviews*

extra lags of the Rating*Reviews variables are included in order to allow for the Hansen and Sargan test for over identifying restrictions (Sargan, 1958). In order for these estimates to be unbiased, it is necessary to assume that the error terms of equations (16) and (9) are not serially correlated for more than two months. Under this restriction, then the three assumptions of the GMM should all be satisfied.[9]

As shown in model (1) of table 2, without the brand fixed effects none of the coefficients is statistically significant, and they are all positive and small, apart from price which remains negative. Including brand fixed effects in model (2), AC seems to have a much larger impact than before and is statistically significant at the 10% level. However none of the other relevant variables is statistically significant and price even displays a positive coefficient.

These estimates may be unreliable also due to the IIA property which is unrealistic given that in this setting the sample includes electronic cables made for different purposes. In fact the IIA property leads to the conclusion that a change in a single characteristic of one of the products will lead to the same percentage change in market share in all other products. For example this leads to assume that if an “hdmi” cable increases in price, then both a direct competitor “hdmi” cable and a “lightning” cable will gain the same relative percentage of market share. This is not possible in this setting as these cables have specific different purposes, even

[9] See section 5.1

if they are under the same market of products. It is necessary therefore to turn to a more robust method (the Nested Logit) which should allow for a better model of consumer heterogeneity.

6.2 Nested Logit

The Nested Logit approach allows to better model the heterogeneity of individual consumers and it should give more precise estimates for the demand coefficients. Note here that the IIA property holds only between products of the same group and therefore is a more reasonable assumption given our dataset of highly substitutable products inside each category.

The results of the two-levels Nested Logit are reported in models (3) and (4) of table 2. It is easy to see that the coefficients are for the majority larger than before and statistically significant. One important different is that in model (4) the impact on consumers' utility of the AC badge and of the bundle is much larger than before. Price looses statistical significance in model (4) as well as Rating*Reviews. Nevertheless, the signs of the coefficients are consistent with what was shown in the Multinomial Logit. Note that in model (3), which does not include the Brand fixed effects, the AC coefficient is not statistically significant and even negative. The validity of this model is, however, challenged by a low p-value in the Hansen and Sargan test for overidentifying restrictions, showing that by not controlling for brand fixed effects there might be biases in the GMM estimation. Therefore the most robust model is considered model (4).

The coefficient σ represents how much consumers value the division in nests compared to products characteristics. σ is bounded between 0 and 1 and it is zero when nests are not important, while it is 1 when characteristics are not important. The values of σ in models (3) and (4) are reasonable and show a good balance in the analysis.

These preliminary analyses show that the importance of AC goes beyond the one of a simple characteristic but is relevant enough to systematically shift consumers buying patterns

6.3 Random Coefficient Logit

The Random Coefficient Logit allows to account in a more precise way for consumer heterogeneity. Here consumers choose from the full set of products, like in the multinomial logit. Unlike the previous models, however, a set of interactions of products characteristics with demographics or standard normal variables is constructed in order to simulate a virtual sample of individuals. In the following model, two forms of heterogeneity are presented: in the first one consumers are simulated from a simple random variable v i.i.d. and normally distributed. In the second model, agents are simulated starting from a full set of demographic variables, as well as the variable v which are both interacted with products characteristics. The results of the first model are expressed in table 3 and two specifications of the model are presented: model (1) without brand fixed effects and model (2) with brand fixed effects.

In both models the AC coefficient is large and statistically significant, but much smaller than our previous Nested Logit estimates. However, it is much more interesting to look at the relative size of the coefficients here. The effect of the AC coefficient is in fact two to three times the effect of offering the product in a bundle or more than thirty times the effect of a decrease in price of one euro (even though the price coefficient loses significance here). Rating*Reviews, while remaining positive, also loses statistical significance with this approach.

Table 3: Random Coefficient Logit Estimates

	θ_1		θ_2	
	(1)	(2)	(1)	(2)
AC	0.182***	0.2978**	-8.37e-4	-7.60e-4
	(0.0188)	(0.118)	(147.121)	(180.413)
Price	-0.00454	-0.00800	-0.00859	-0.0103
	(0.0258)	(0.0120)	(47.684)	(21.385)
Rating*Reviews	2.483e-06	8.29e-07	0.00296	0.00321
	(4.61e-06)	(2.25e-06)	(0.0133)	(0.00964)
Bundle	0.116***	0.0906	-5.11e-4	-3.21e-4
	(0.0434)	(0.0617)	(70.872)	(104.501)
Constant	-5.634***	-5.622***		
	(0.147)	(0.124)		
Observations	1,059	1,059	1,059	1,059
Brand FE		Yes		Yes

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

*Note: the table reports the mean coefficient estimates of the BLP model (Θ_1) and the interactions of the coefficients with the random element v (Θ_2). Two models are estimated (1) wihtout brand fixed effects and (2) with brand fixed effects. Both models are estimated with the GMM approach using Arellano-Bond instruments for the variables AC, Price and Rating*Reviews*

6.4 Full Model Estimates

The most interesting results come from the full model estimation, which includes demographic variables to simulate individuals. The set of demographics include $Log(Income)$, $Log(Income)^2$ Age and $Mobile$. The interactions estimated are present for all the variables, excluding the constant, which is left out of the random component of demand. This is due to the fact that the demographic variables and the random component are not expected to have a direct impact on consumers' utility. These variables, in fact, can only affect the utility of buying a product when interacted with products characteristics.

As in the previous Sections, two models are presented, one with brand fixed effects and one without. In the estimation without brand fixed effects reported in

Table 4, it is easy to see that the mean coefficients β are similar to the ones estimated in the simpler Random Coefficient model without demographic variables. However, the more precise consumer specification through demographics results in much lower standard errors for the mean effects β. The interactions with demographics are less interesting and generally not statistically significant. It is difficult to propose an explanation for the occasionally significant coefficients as there seems to be no structural meaning to these isolated cases of significant coefficients. The lack of significance in the demographics coefficients may also be due to the little number of markets considered in this work, as it is not possible to consider more than the five European Amazon markets. In the seminal papers that use the BLP integrated with demographic variables, such as Nevo (2001), the number of markets is much larger (in that case 94 markets were considered), allowing for more differentiation among the distributions of demographics in different markets. Unfortunately, Amazon Europe only allows us to study five separate geographical markets.

Table 4: Full Model

	Mean Effects	Random	Demographics Interactions			
	β	v	$Log(Income)$	$log(Income^2)$	Age	$Mobile$
No Brand FE						
AC	0.183***	-0.00172	1.29e-4	0.00129	0.0208**	0.0156
	(3.15e-6)	(0.00242)	(0.0129)	(0.0129)	(0.00940)	(0.0597)
Price	-0.00454**	-0.00151	8.91e-4	-0.0238	0.00489	-0.00133
	(0.00198)	(0.915)	(0.836)	(0.836)	(0.471)	(1.41)
Rating*Reviews	2.52e-6	0.00293	5.68e-4	0.0408***	8.58e-4	0.0309*
	(3.623e-5)	(0.101)	(0.0113)	(0.0113)	(0.0382)	(0.0176)
Bundle	0.116***	3.53e-5	-0.00104	3.26e-4	7.12e-4	-0.0361
	(6.08e-5)	(0.0417)	(0.0295)	(0.0295)	(0.00771)	(0.0400)
Constant	-5.635***					
	(1.93e-6)					
Brand FE						
AC	0.299***	-0.00165	-1.43e-4	0.00133	0.0205	0.0164
	(0.0563)	(73.325)	(73.969)	(73.969)	(39.222)	(80.302)
Price	-0.00800	-0.00161	-7.35e-4	-0.0256	0.00877	-0.00139
	(0.0275)	(56.469)	(3.884)	(3.884)	(20.883)	(13.945)
Rating*Reviews	8.82e-7	0.00316	5.24e-4	0.0429***	8.36e-4	0.0327*
	(2.30e-5)	(0.0255)	(0.00293)	(0.00293)	(0.0314)	(0.0181)
Bundle	0.0906*	-6.91e-5	1.52e-4***	2.33e-4	6.43e-4	-0.0339
	(0.0563)	(37.612)	(65.361)	(65.360)	(25.547)	(58.907)
Constant	-5.622***					
	(0.105)					

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

*Note: The table reports the full BLP model estimates with mean effects of the products characteristics and interactions with the demographic variables: Income, Age and Mobile. Two models are estimated, the first one with brand fixed effects and the second one with brand fixed effects. Both models are estimated with the GMM approach using Arellano-Bond instruments for the variables AC, Price and Rating*Reviews*

In Table 4 we also report the results of the full model with brand Fixed Effects. Here as well most of the interactions are not statistically significant, and the mean coefficients are estimated with higher precision than in the previous model without demographics (table 3). All the previous conclusions on mean effects hold here as well, with AC having a strong positive and statistically significant effect on Utility of the agents and likely driving a large portion of the demand. It is important to note a strong and statistically significant interaction of Rating*Reviews with $Log(Income)^2$ and $Mobile$ which shows that Rating*Reviews is likely more relevant for high income groups and for people using mobile phone to shop online. Bundle also shows a positive and statistically significant interaction with $Log(Income)$ even though the size of the interaction is small when compared with the mean effect of Bundle on consumers utility. This suggests that higher income agents are slightly more likely to buy a bundle of products to store for later use than to buy a single product for a one-time use.

As in the previous results, more interesting conclusions can be drawn on the relative magnitude of the coefficients. With the AC badge being always more than thirty times larger (in absolute terms) than the price coefficient and up to three times as large as the bundle coefficient. The consistency of the relative size of the AC coefficient across different models shows that the relative magnitudes are not dependent on model specific assumptions, such as the nests chosen for the nested logit or the way in which we simulated individuals in the BLP models.

6.5 Discussion

The presence of such large coefficients across different models and estimation techniques suggests that the AC badge has a substantial effect on consumers' demand. Understanding the exact process through which the badge acts is out of the scope of this paper, but it is recognized here that there are multiple interpretations of the AC coefficient.

The model presented follows the work of Hui et al. (2016) and Tadelis (2016) in interpreting the AC badge as a signal of seller reputation and product quality. In this case, consumers consider all possible options, but value more the AC one as it is directly suggested by the platform. This perception is not necessarily correct as it is unclear to customers how the AC badge is assigned. This approach also follows the work of Bordalo, Gennaioli, and Shleifer (2013) in which a model of firms competition with salient characteristics is presented. The AC badge can in fact easily capture consumers' attention in a situation of choice overload, in which simpler options are often preferred by customers, as shown in the experiment proposed in Iyengar and Kamenica (2010).

An alternative interpretation of the coefficient could be that the time required for the transaction is not null and therefore search costs are not equal to zero. Consumers may be choosing under time constraints and the loss caused by choosing a non optimal predetermined option may be less than the value of time needed to compare all possible alternatives. Some customers may in fact prefer a predetermined option in a market such as the one analyzed, where products are similar in price and characteristics. The small benefits of a complete search for the best product may be out weighted by the time saved in purchasing the suggested option. This interpretation is consistent with the latest research on consumer's choice under time pressure. Reutskaja et al. (2011) shows that customers, while still optimizing under choice overload and time pressure, often display behavioral biases like systematically putting more attention to products displayed in the central area of their vision. Also the more recent Coey, Larsen, and Platt (2020) shows that under time constraints consumers may settle for non-optimal choices and more expensive products.

The relevance of the interaction of AC and customers' age in the full BLP model without fixed effects shows a potential heterogeneous effect of the AC badge on consumer's demand. This heterogeneity is insignificant when brand fixed effects are included. However, future research with a larger number of markets and more precise customer data could show more relevant interactions of the AC badge with individual consumer's characteristics. This small heterogeneity in the effect of the AC badge could be due to the heterogeneous effects of the badge on consumers attention (which could vary with age); alternatively it could be due to heterogeneity in consumers' value of time. The differential effects of the AC badge on time should however be reflected in the interaction with income: agents with larger income should value more the AC option as the value of time saved during the choice should be larger. However, the interaction with income doesn't show any relevant effects on consumer demand, showing that this type of heterogeneous effect is unlikely to be present in the proposed setting.

7 Conclusions

This study aims at showing the impact of a specific Amazon Marketplace suggestion on consumer demand. To do so a simple theoretical model of discrete choice is constructed and the Amazon's Choice badge is analyzed as a characteristic of the product. The scope of this work is to find robust evidence of the influence of Amazon's Choice badge on consumers demand.

A set of structural empirical models is presented to estimate the demand function of consumers and strong evidence of the influence of Amazon's Choice badge on consumers' utility and demand is found. The structural empirical techniques used are taken from the literature of applied industrial organization and followed mainly the work of Berry (1994), Berry, Levinsohn, and Pakes (1995) and Nevo (2001). The techniques used are Multinomial Logit, Nested Logit and Random Coefficient Logit. These three have a gradually more precise characterization of simulated individual agents across the five European Amazon Marketplaces (Italy, United Kingdom, Spain, France, Germany). The estimates of Amazon's Choice impact for the rational coefficient range from a more conservative 0.299 in the Random Coefficient model with brand Fixed Effects, to a more extreme 1.174 in the Nested Logit with brand Fixed Effects.

These results show not only a strong and statistically significant impact of the Amazon's Choice badge on consumer utility, but also a very large impact when compared to the other products' characteristics. Among the characteristics included in the analysis (Price, Rating*Reviews and Bundle), none of them is able to rival the impact of Amazon's Choice, with Bundle having the largest effect ranging from 0.0906 to 0.22. These results suggest that the effect of Amazon's Choice on these particular set of products (highly substitutable with generally low variance in prices) is more than offering the same product in a bundle, which likely drastically increases the costs of production for the product.

All these estimates show how strong the effect of a suggestion badge can be in online platforms. The purpose of the work is in fact not to draw conclusions on the channels through which the suggested choices act in consumers' decisions. Instead this works proves that the Amazon's Choice badge has a strong and statistically significant effect on consumers' decisions, through the use of different theoretical models and empirical techniques.

This work contributes to the existing literature of Empirical Industrial Organization in showing that even in an environment with almost null search costs (such as a common online platform with similar products), the role of suggestions and attention grabbers seems to be fundamental in consumers' choices. The results proposed here are likely driven by the peculiarity of the market presented, in which single products characteristics may not be as relevant as in markets with highly differentiated products. Nonetheless the results shown here show that the platform itself is capable of driving demand to the desired products with very little effort (simply suggesting a product over others in a noninvasive and subtle way). The implications of these results are even more relevant due to the fact that Amazon does not reveal how the Amazon's Choice badge is assigned, leaving sellers to speculate how to reach it. A non-neutral environment for the assignment of the Amazon's Choice badge, which could favour certain sellers over others, is likely to have a dramatic impact on the relevant market for such product, driving demand away or towards certain sellers (among which also Amazon competes either producing products under the "Amazon Basics" brand or buying products from producers and selling them directly).

This work suggests for more transparency on the assignment of the Amazon's Choice badge and other instruments similar to this one. Under the current European legislation (Council of the European Union, 2019), online platforms are required to disclose the main determinants of ranking algorithms, among which, by article 2.8, Amazon's Choice should be considered. It remains to be seen whether a caveat on consumer protection and personalization of results (article 5.6) is used to both protect both the consumers data and the intellectual property of the algorithm determining the Amazon's Choice products. The more recent "New Competition Tool" (European Commission, 2020b) and the "Digital Service Act Package" (European Commission, 2020a) initiative, which both aim at being adopted by the end of 2020, propose specific tools for the control of large internet platforms that act as gatekeepers and intermediaries. Among the objectives expressed in section B of the "Digital Service Act Package", at point 2 is written:

> "*[...] further horizontal rules could be envisaged with a purpose to enable collection of information from large online platforms acting as gatekeepers by a dedicated regulatory body at the EU level to gain, for example, further insights into their business practices and their impact on these platforms' users and consumers.*"

It remains to be seen whether the proposed tools will have the power to assess the neutrality of the assignment process of the Amazon's Choice badge and similar platform suggestions.

References

Manuel Arellano and Stephen Bond. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297, 1991.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–90, 1995.

Steven T. Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 25(2):242–262, 1994.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843, 2013.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Competition for Attention. *The Review of Economic Studies*, 83(2):481–513, 2015.

Bart J. Bronnenberg, Jean-Pierre Dubé, Matthew Gentzkow, and Jesse M. Shapiro. Do Pharmacists Buy Bayer? Informed Shoppers and the Brand Premium. *The Quarterly Journal of Economics*, 130(4):1669–1726, 2015.

Zach Y Brown and Jihye Jeon. Endogenous information and simplifying insurance choice. Mimeo, University of Michigan, 2020.

N. Scott Cardell. Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. *Econometric Theory*, 13(2):185–213, 1997.

Dominic Coey, Bradley J. Larsen, and Brennan C. Platt. Discounts and deadlines in consumer search. *American Economic Review*, 110(12):3748–85, 2020.

Council of the European Union. Council regulation (EU) no 1150/2019, 2019. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R1150&from=EN`. Last access 10 November, 2020.

Francesco Decarolis and Gabriele Rovigatti. From mad men to maths men: Concentration and buyer power in online advertising. *American Economic Review*, 111(10):3299–3327, 2021.

Francesco Decarolis, Maris Goldmanis, and Antonio Penta. Marketing agencies and collusive bidding in online ad auctions. *Management Science*, 66(10):4433–4454, 2020a.

Francesco Decarolis, Maria Polyakova, and Stephen P Ryan. Subsidy design in privately provided social insurance: Lessons from medicare part d. *Journal of Political Economy*, 128(5):1712–1752, 2020b.

Glenn Ellison. *Bounded Rationality in Industrial Organization*, volume 2 of *Econometric Society Monographs*, page 142–174. Cambridge University Press, 2006.

European Commission. Antitrust: Commission opens investigation into possible anti-competitive conduct of amazon, 2019. URL `https://ec.europa.eu/commission/presscorner/detail/en/IP_19_4291`. Last access 10 November, 2020.

European Commission. Digital service act package ex ante regulatory instrument of very large online platforms acting as gatekeepers, 2020a. URL `https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12418-Digital-Services-Act-package-ex-ante-regulatory-instrument-of-very-large-online-platforms-acting-as-gatekeepers`. Last access 10 November, 2020. Ref: Ares(2020)2877647 - 04/06/2020.

European Commission. New competition tool - inception impact assessment, 2020b. URL `https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12416-New-competition-tool`. Last access 10 November, 2020. Ref: Ares(2020)2877634 - 04/06/2020.

Amit Gandhi, Zhentong Lu, and Xiaoxia Shi. Estimating demand for differentiated products with zeroes in market share data. Available at SSRN: `https://ssrn.com/abstract=3503565`, 2020.

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.

Michael D. Grubb. Failing to choose the best price: Theory, evidence, and policy. *Review of Industrial Organization*, 47(3):303–340, 2015.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

Ali Hortaçsu, Seyed Ali Madanizadeh, and Steven L Puller. Power to choose? an analysis of consumer inertia in the residential electricity market. *American Economic Journal: Economic Policy*, 9(4):192–226, 2017.

Xiang Hui, Maryam Saeedi, Zeqian Shen, and Neel Sundaresan. Reputation and regulations: evidence from ebay. *Management Science*, 62(12):3604–3616, 2016.

Sheena S Iyengar and Emir Kamenica. Choice proliferation, simplicity seeking, and asset allocation. *Journal of Public Economics*, 94(7-8):530–539, 2010.

Aviv Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.

Andrea Prat and Tommaso M Valletti. Attention oligopoly. *American Economic Review: Microeconomics*, 2021. Forthcoming.

Eric Rasmusen et al. The blp method of demand curve estimation in industrial organization. Last access 10 November, 2020, 2007. URL `https://www.rasmusen.org/published/blp-rasmusen.pdf`.

Elena Reutskaja, Rosemarie Nagel, Colin F Camerer, and Antonio Rangel. Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review*, 101(2):900–926, 2011.

Maryam Saeedi. Reputation and adverse selection, theory and evidence from ebay. Available at SSRN: `https://ssrn.com/abstract=2102948`, 2014.

J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958.

Shane Shifflett, Alexandra Berzon, and Dana Mattioli. Amazon's choice isn't the endorsement it appears. *The Wall Street Journal*, 2019. URL `https://www.wsj.com/articles/amazons-choice-isnt-the-endorsement-it-appears-11577035151`. Last access 10 November, 2020.

Ran Spiegler. *Bounded rationality and industrial organization.* Oxford University Press, 2011.

Steven Tadelis. Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8:321–340, 2016.

Kenneth E Train. *Discrete choice methods with simulation.* Cambridge university press, 2009.

Appendix: BLP Method

This appendix describes the process of the BLP optimization procedure employed in this work. It is fully based on the works of Berry, Levinsohn, and Pakes (1995) and Nevo (2001). Before starting with the optimization algorithm, we specify the starting values of the coefficients α and $\boldsymbol{\beta}$ in $\delta_{j,t}$ and for the coefficient $\sum$ and $\prod$ in $u_{i,j,t}$. In this work, α and $\boldsymbol{\beta}$ starting points are derived by applying the simlple logit regression (not presented in this paper).

0. Start by making random draws of the v_i's and D_i's. The distribution of v_i are assumed standard normal, while the distributions of D_i's are taken from the real distributions of the demographics in the different markets. This step will only be done at the beginning of the estimation and will not be part of the optimization routine.

1. Calculate the predicted market shares:

$$d_{jt} = \frac{1}{n}\sum_{i=1}^{n} \frac{exp\{\delta_{j,t} + u_{i,j,t}\}}{1 + \sum\limits_{j=1}^{J} exp\{\delta_{j,t} + u_{i,j,t}\}} \tag{23}$$

2. Use the following contraction mapping to get a better estimate of δ

$$\boldsymbol{\delta}^{h+1} = \boldsymbol{\delta}^{h} + (\ln(\mathbf{S}_t) - \ln(\mathbf{d_t})) \tag{24}$$

   Where $\mathbf{S}_t$ are the real market shares and $\mathbf{d}_t$ are the predicted market shares in market t calculated in (1). When the contraction mapping is concluded, the result is a more precise estimate for the mean utility $\boldsymbol{\delta}_t$, called $\hat{\boldsymbol{\delta}}_t$

3. Calculate the error term

$$w_{j,t} = \hat{\delta}_{j,t} - (-\alpha p_{j,t} + \mathbf{x}'_{j,t}\boldsymbol{\beta}) \tag{25}$$

   And obtain the moment condition:[10] $\boldsymbol{w'Z\Omega^{-1}Z'w}$.

4. Compute better estimates for $\alpha, \beta, \sum, \Omega$.

   - Find $(\hat{\alpha}, \hat{\beta})$ such that:

$$(\hat{\alpha}\ \hat{\boldsymbol{\beta}})_{GMM} = (\boldsymbol{X'Z\Omega^{-1}Z'X})^{-1}\boldsymbol{X'Z\Omega^{-1}Z'\delta} \tag{26}$$

[10] Z is the matrix of exogenous variables, used as instruments for price and for the other endogenous characteristics in X (if present)

- With the new estimates $\hat{\alpha}$ and $\hat{\beta}$ get the new error term:

$$\hat{w}_{j,t} = \hat{\delta}_{j,t} - (-\hat{\alpha} p_{j,t} + \mathbf{x}'_{j,t}\hat{\boldsymbol{\beta}}) \tag{27}$$

- Get the new moment condition $\hat{\boldsymbol{w}}'\boldsymbol{Z\Omega^{-1}Z'}\hat{\boldsymbol{w}}$ with the new $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ just calculated
- Estimate the more precise Hansen (1982) weighting matrix with the new error term $\boldsymbol{w}$: $\boldsymbol{\Omega = (E(Z'ww'Z))}$.
- Use a search algorithm, iterating from 1 to 4, to find the best estimate for $\sum$ amd $\prod$, which is the one that minimizes the GMM objective function[11].

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} \boldsymbol{w'Z\Omega^{-1}Z'w} \tag{28}$$

Note that $(\hat{\alpha}, \hat{\beta})$ are calculated linearly at (26), while $\sum$ and $\prod$ are obtained through the search algorithm by iterating from point 1 to 4.

[11] This work follow Nevo (2001) suggestion of using a simplex search, starting from guessed $\sum$ and $\prod$ values, then use the more precise (but more sensibe to the starting point) Quasi-Newton optimization algorithm to estimate $\sum$ and $\prod$, using as a starting point the $\sum$ and $\prod$ coefficients obtained with the simplex search method