

Título: EL COMERCIO ELECTRÓNICO EN ESPAÑA

Autores: Leonel Cerno (1) y Teodosio Pérez Amaral (2)

(1) Centro: Departamento de Economía de la Empresa, Universidad Europea de Madrid

Dirección: Urb. El Bosque, s/n. 28670 Villaviciosa de Odón, Madrid, Spain.

Teléfono: (+34) 91 211 56 44

Fax: (+34) 91 616 82 65

E-Mail: leonel.cerno@uem.es.

(2) Centro: Departamento de Fundamentos del Análisis Económico II (Economía Cuantitativa), Universidad Complutense de Madrid

Dirección: Campus Universitario de Somosaguas, 28223, Pozuelo de Alarcón, Madrid, Spain.

Teléfono: (+34) 91 394 23 80

Fax: (+34) 91 394 23 80

E-Mail: teodosio@ccee.ucm.es.

Resumen

Analizamos los factores que influyen el comercio electrónico para el caso español. Empezamos planteando un modelo para los determinantes de la compra en la Web teniendo en cuenta el nexo existente entre la adquisición de bienes y el acceso y uso del servicio de Internet. Los resultados revelan que la accesibilidad a la Red será el principal factor (seguido de la renta y el género sexual) que influye en la probabilidad de la compra electrónica. Luego nos planteamos un modelo condicionado a los usuarios del comercio electrónico para los determinantes del gasto en la Web. Descubrimos que tener en cuenta la disparidad (dispersión) en los datos será importante para este tipo de modelos, y para obtener un modelo final utilizamos el algoritmo RETINA, que permite el tratamiento de la heterogeneidad y mejora sustancialmente los resultados finales del modelo.

Abstract

In this article we analyze the factors that affect the e-commerce for the Spanish case. We begin with a model specification for the determinants of the buy in the Web taking into account the existing link among the acquisition of goods and the access and use of the Internet service. The results reveal that the accessibility to the Net will be the principal factor followed by the income and the genre that impact in the probability of the electronic buy. Then we raising a model determined to the users of the e-commerce for the determinants of the expenditure in the Web. We discover that bring it the disparity (dispersion) in the data will be important for this type of models, and to obtain a final model we use the algorithm RETINA, which allows the treatment of the heterogeneity and improves substantially the final results of the model.

Palabras clave: Gasto por Internet, comercio electrónico, RETINA, pendientes específicas, k-medias.

JEL-Code: C2, C21, C25

EL COMERCIO ELECTRÓNICO EN ESPAÑA¹

Leonel Cerno

Universidad Europea de Madrid

Teodosio Pérez Amaral

Universidad Complutense de Madrid

En este artículo analizamos los factores que influyen el comercio electrónico para el caso español. Empezamos planteando un modelo para los determinantes de la compra en la Web teniendo en cuenta el vínculo existente entre la adquisición de bienes y el acceso y uso del servicio de Internet. Los resultados revelan que la accesibilidad a la Red será el principal factor (seguido de la renta y el género sexual) que influye en la probabilidad de la compra electrónica. Luego nos planteamos un modelo condicionado a los usuarios del comercio electrónico para los determinantes del gasto en la Web. Descubrimos que tener en cuenta la disparidad (dispersión) en los datos será importante para este tipo de modelos, y para obtener un modelo final utilizamos el algoritmo RETINA, que permite el tratamiento de la heterogeneidad y mejora sustancialmente los resultados finales del modelo.

Palabras clave: Gasto por Internet, comercio electrónico, RETINA, Pendientes específicas, k-medias.

1. INTRODUCCIÓN

Dentro de las principales características que diferencian a los mercados virtuales de los mercados convencionales destaca el hecho de que habría menos fricciones. A su vez, una de las formas de crear poder de mercado en el caso del comercio virtual es reducir lo que se denomina *costes de búsqueda*, pues Internet ha incrementado sustancialmente la disponibilidad de la información de precios y productos, permitiendo a los consumidores identificar la mejor opción y mejorar su posición frente a los vendedores on-line. Es decir que los consumidores actuales han ido mejorando y fortaleciendo su posición a medida que el Comercio electrónico se ha desarrollado. En la literatura

¹ Los autores agradecen el apoyo financiero obtenido a través del Proyecto Cicyt SEJ2004-06948 del Plan Nacional de I+D del Ministerio de Educación, Ciencia y Tecnología.

existente se puede destacar a Clemons, Hann y Hitt (1999), Combers y Patel (1997) y Brynjolfsson y Smith (1999) que analizan la conducta del consumidor a partir de la aparición del comercio electrónico.

Los mercados online parecen más eficientes en términos de precios y de costes. Algunos estudios han encontrado una dispersión de precios apreciable (Bakos, 2001). Esta dispersión puede ser explicada, en parte, por la heterogeneidad de factores específicos tales como la confianza o el conocimiento del sitio Web o de la marca.

El análisis del nivel de fricción de los mercados de Internet se hace desde dos puntos de vista: o bien comparando las características de los mercados electrónicos y los convencionales, o bien analizando la conducta dentro de los mercados electrónicos. El presente capítulo se centra en la segunda opción.

En los mercados electrónicos, la eficiencia se mide en cuatro dimensiones: el nivel de precios, su dispersión, las elasticidades de los bienes a los precios y los costes (Smith, Bailey y Brynjolfsson, 1999). En cuanto a la creación de los precios, habrá factores que no han influido demasiado en la formación de los precios de los mercados convencionales, como por ejemplo el efecto de los costes de búsqueda que presiona los precios a la baja e intensifica el intento de diferenciación para poder mantener los precios por encima del coste marginal (Peddibhotla, 2000). A su vez, el servicio post-venta para ciertos tipos de bienes o incluso el efecto del tamaño de la Red en el equilibrio podrán en principio ser tratados con elementos de la teoría microeconómica tradicional sobre todo utilizando los supuestos tradicionales (por ejemplo la información perfecta), pero yendo más allá de la creencia común de que los precios en Internet son bajos porque los consumidores tienen la facilidad de encontrarlos rápidamente. En la investigación reciente de Shapiro y Varian (1999) se analiza bajo qué condiciones ocurre todo esto.

En el presente estudio se propone una aproximación empírica utilizando una muestra de datos de compradores por la Red en España, se mide el impacto sociodemográficos en el comercio electrónico y se contrastan sus implicaciones en la demanda de bienes y servicios en la Red.

2. COMERCIO ELECTRÓNICO Y USO DE INTERNET

Para entender el perfil de los demandantes de bienes a través de la Red, es importante considerar que los determinantes de la probabilidad de comprar a través de Internet incluyen información acerca de los determinantes de la demanda de uso de Internet (Cerno y Pérez-Amaral, 2005). Un regresor en un modelo de respuesta binaria explica el impacto en la probabilidad de que el agente elija una opción particular (Train, 2002). Al explicar aquí el comportamiento y los determinantes de la demanda de bienes a través de Internet, hay que considerar al acceso y uso de Internet como una variable explicativa.

2.1. ANÁLISIS DESCRIPTIVO Y DEFINICIÓN DE VARIABLES

En una primera descripción del conjunto de datos se observa que sólo el 3,7% de la población española compra habitualmente por Internet (últimos tres meses). A continuación se presenta el Gráfico 1 donde se pueden ver los porcentajes del número de transacciones de bienes y servicios adquiridos a través de Internet:

Gráfico1: Productos comprados a través de Internet en los últimos 12 meses



Se observa que para la muestra estudiada los productos más demandados son las actividades de ocio tales como los viajes (41,1%) y las entradas a espectáculos (26,5%),

mientras que en contrapartida están la adquisición de acciones y las apuestas que serán las menos demandadas, con un 4,9% y un 0,9% respectivamente.

2.1.1. CARACTERÍSTICAS DE LOS DATOS

Los datos utilizados en este artículo son parte de la encuesta TIC-H (2003) del Instituto Nacional de Estadística (INE). Se trata de los 5.273 individuos que se declaran usuarios de Internet. A continuación se presentan las Tablas 1 y 2 donde se definen las variables y se presentan estadísticos descriptivos.

Tabla 1. Definición de Variables

<i>Variable</i>	<i>Definición</i>
<i>COMPRA</i>	Ficticia =1 si el individuo compra por Internet
<i>IS_Q</i>	Quintil de Renta (variable secuencial 1-5) (=1 si se encuentra en el quintil más bajo... =5 si se halla en el quintil más alto)
<i>VIVINT</i>	Ficticia =1 si el individuo accede a Internet desde el hogar
<i>USO</i>	Cantidad de sitios donde usa Internet (de 1 a 4)
<i>USOIN</i>	Intensidad de uso de Internet (horas semanales)
<i>USOR</i>	Intensidad de uso del ordenador (horas semanales)
<i>SEXH</i>	Ficticia =1 si el individuo es hombre
<i>EDAD</i>	Edad del individuo
<i>POBLAC</i>	Tamaño poblacional relativo (nivel provincial) ²
<i>TOTMIEM</i>	Cantidad de miembros en el hogar del individuo

Tabla 2. Estadísticos Descriptivos

<i>Variable</i>	<i>Desviación</i>			
	<i>Media</i>	<i>Típica</i>	<i>Mínimo</i>	<i>Máximo</i>
<i>COMPRA</i>	0,126	0,336	0	1
<i>IS_Q</i>	4,030	0,850	1	5
<i>VIVINT</i>	0,615	0,487	0	1

² Se refiere al tamaño poblacional de cada provincia a la que pertenece el individuo dividido por la población total de España.

<i>USO</i>	1,490	0,653	1	4
<i>USOIN</i>	38,001	29,805	0	70
<i>USOR</i>	46,642	29,257	0	70
<i>SEXH</i>	0,519	0,499	0	1
<i>EDAD</i>	33,862	12,465	15	88
<i>POBLAC</i>	4,237	4,429	0,163	13,277
<i>TOTMIEM</i>	3,333	1,192	1	6

A partir de este primer análisis a continuación se propone estimar un modelo para los determinantes de la compra de bienes a través de Internet, considerando determinantes sociodemográficos tales como la renta, el sexo, la edad y el hábitat poblacional. También se consideran las características del individuo respecto al servicio de Internet tales como el acceso desde el hogar, el uso desde otros sitios además del hogar y la intensidad de utilización del servicio medida en horas semanales. La especificación, realizada con un modelo logit binario, quedará entonces:

$$\ln\left(\frac{P_i}{1-P_i}\right) = b_0 + b_1 IS_Q_i + b_2 VIVINT_i + b_3 USO_i + b_4 USOINT_i + b_5 USOR_i + b_6 SEXO_H_i + b_8 EDAD_i + b_9 EDAD_i^2 + b_{10} POBLAC_i + b_{11} TOTMIEM_i + u_i$$

(4.4)

Donde $P_i = \Pr(COMPRA_i)$ es la probabilidad de que el individuo i -ésimo compre a través de Internet. A continuación, en la siguiente tabla se presentan los resultados de las estimaciones de los parámetros y de los ratios de ventaja:

Tabla 3. Resultados del Modelo Logit para la Compra por Internet

Variable			
Dependiente:	Coeficiente	Ratio de	z
Compra por Internet		Ventaja	
Constante	-6,19	---	13,67
<i>IS_Q</i>	0,23	1,26	3,75
<i>VIVINT</i>	0,48	1,62	4,39

<i>USO</i>	0,39	1,48	5,84
<i>USOIN</i>	0,01	1,01	6,31
<i>USOR</i>	0,01	1,01	2,61
<i>SEXH</i>	0,38	1,46	4,24
<i>EDAD</i>	0,09	1,09	4,01
<i>EDAD</i> ²	-0,001	0,99	4,22
<i>POBLAC</i>	0,04	1,04	4,77
<i>TOTMIEM</i>	-0,08	0,92	2,16

Observaciones: 5.223

Log-Verosimilitud: -1773,45

$\chi^2_{(10)}$: 397,09 (Prob. = 0,000)

Pseudo R²: 0,1007

La significatividad conjunta del modelo es alta, la razón de verosimilitudes que se obtiene es de 397,09, valor elevado para una chi-cuadrado con 10 grados de libertad. La bondad de ajuste del modelo se la considera también como buena luego de ver el grado de acierto de las predicciones dentro de la muestra:

Tabla 4. Predicciones del Modelo Estimado

		Valor Real		
		<i>COMPRA</i> = 0	<i>COMPRA</i> = 1	Total
Valor Predicho	<i>COMPRA</i> = 0	3.063	208	3.271
	<i>COMPRA</i> = 1	1.521	451	1.972
	Total	4.584	659	5.243

El modelo predice acertadamente 3.514 de las 5.243 observaciones (67,02%). Debido a la gran cantidad de ceros dentro de la muestra, se utiliza como umbral la proporción de unos en la variable endógena (0,126). Con esto, la especificidad (porcentaje de aciertos iguales a 0) es de 66,81% y la sensibilidad (porcentajes de aciertos iguales a 1) de 68,43%.

Hay que mantener siempre en mente que se está utilizando una muestra condicionada a los usuarios de Internet, de manera que los resultados obtenidos y las interpretaciones realizadas serán sólo aplicables a esta parte de la muestra. Analizando los valores de los ratios de ventaja, se puede decir:

- Exceptuando la cantidad de miembros en el hogar, todas las demás variables tendrán efectos positivos en la probabilidad de comprar por Internet.
- El acceso a Internet en la vivienda (*VIVINT*) y la cantidad de lugares de uso de Internet (*USO*), ambas variables referidas a los hábitos de Internet del individuo, tienen los coeficientes de mayor valor. El acceso a Internet en la vivienda tendrá el valor mayor en el ratio (1,62). Le sigue la cantidad de sitios de uso de Internet con un ratio de 1,48. Es decir que el acceso a Internet en el hogar contribuye en la probabilidad de adquirir productos a través de la Red en un 62%, y el utilizar el servicio en varios sitios además del hogar contribuirá a incrementar la probabilidad de comprar en 48%. Estos resultados tienen lógica al considerar que antes de comprar por Internet, primero hay que conectarse a la Red, o bien desde el hogar, o bien desde otros sitios.
- Se observa que pertenecer al sexo masculino tiene influencia en la probabilidad de comprar por Internet. Su ratio de ventaja es de 1,46. Esto sugiere mayor afinidad del género masculino con las compras a través de la Red.
- La influencia de los niveles de renta también es positiva. La variable que contiene sus efectos tiene un ratio de ventaja de 1,26. Tendrá también su lógica debido a que al haber una mayor renta habrá un mayor gasto. Es decir que también tendrá una influencia positiva en la probabilidad de comprar por Internet.
- El hábitat poblacional (*POBLAC*), las intensidades de uso de Internet (*USOIN*) y del ordenador (*USOR*) también impactarán de forma positiva en la probabilidad de compra por Internet, pero de manera menor que las detalladas en los puntos anteriores. Los ratios de ventaja serán más cercanos a 1 (1,04 para la variable *POBLAC*, y 1,01 tanto para la variable *USOIN* y *USOR*).
- El caso de la variable *EDAD* es específico debido a que está incluida dentro del modelo tanto en niveles como al cuadrado. Su derivada es una función lineal en *EDAD*: $\partial \hat{COMPR} / \partial EDAD = 0,09 - (2) \cdot (0,001) EDAD$. Es decir que a medida

que aumenta la edad disminuyen los aumentos de la probabilidad de comprar por Internet. El punto de inflexión donde el efecto total de la edad empieza a hacer decrecer la probabilidad sería sobre los 45 años.

- La única variable de las consideradas que disminuye la probabilidad de comprar bienes y servicios en Internet será la variable referida al total de miembros en el hogar (*TOTMIEM*). Su ratio de ventaja está por debajo de 1 (0,92). Es decir que a mayor cantidad de miembros en el hogar, menor probabilidad de adquirir bienes en la Red.

Por último se evalúa la variación en las probabilidades pronosticadas observando el cambio que se produce en estas cuando el individuo adquiere una conexión a Internet desde el hogar, pasar de utilizar Internet cada vez desde más lugares o pasar de un quintil de ingresos inferior a uno superior. A continuación se presentan las siguientes tablas:

Tabla 5. Probabilidades de Comprar en Internet en los Distintos Niveles de Renta

Quintil de Renta	1°	2°	3°	4°	5°
Probabilidad	0,0815	0,1004	0,1232	0,1503	0,1821
Diferencia	0,0189	0,0228	0,0271	0,0318	

Tabla 6. Probabilidades de Comprar en Internet para la Cantidad de Sitios de Uso

Sitios de Uso de Internet	1	2	3	4
Probabilidad	0,1283	0,1786	0,2431	0,3217
Diferencia	0,0503	0,0645	0,0786	

Tabla 7. Probabilidades de Comprar en Internet para el Acceso desde el Hogar

Internet en el Hogar	No	Sí
Probabilidad	0,1171	0,1765
Diferencia	0,0479	

De estas tablas se puede decir:

- Las diferencias muestran que el mayor incremento en la probabilidad de comprar por Internet se debe a que los individuos utilicen el servicio en muchos sitios, siendo este incremento de un 7,86% cuando se pasa de utilizar Internet en tres sitios a utilizarlo en cuatro.
- Pasar a un quintil de ingresos superior no siempre provoca el mismo incremento en la probabilidad de comprar en Internet, aunque vemos que esta diferencia de probabilidades va incrementándose a medida que se va subiendo de categoría de renta
- Como es lógico, el contar con el servicio de Internet desde el hogar incrementa la probabilidad de comprar productos en la Red, con un incremento superior a los experimentados al pasar a quintiles superiores de ingresos (4,79% contra el mayor incremento al pasar del quintil 4º al 5º con una diferencia del 3,18%).

Como conclusión de esta primera parte se puede decir que, dentro de los determinantes considerados para la compra de bienes y servicios en la Red, los más importantes serán las características de los individuos como tener acceso al servicio desde la vivienda, y la cantidad de lugares de uso del mismo además del hogar. Curiosamente la intensidad de uso de Internet (horas semanales conectado a la Red) impactará positivamente, pero no con la magnitud que lo harán estas dos variables mencionadas (acceso desde la vivienda y cantidad de lugares de uso). Las otras dos variables que provocarán impactos positivos en la probabilidad serán la pertenencia al género masculino y el nivel de renta, con impactos importantes tanto una como otra variable, pero no con el mismo grado que las dos mencionadas anteriormente. Esto tiene sentido debido a que antes de comprar bienes y servicios en la Red hay que tener acceso a la misma, y además hacer uso del servicio. Una vez cubierta esta parte, los factores como el nivel de renta o el género también influirán positivamente en la probabilidad, pero en menor medida.

Una vez estudiados el perfil de quién compra en la Red dentro de los que tienen acceso a la misma, a continuación se elabora un modelo de gasto individual en comercio electrónico para los usuarios de Internet. A continuación se especificará y estimará una función de demanda de bienes y servicios y una función de gasto a través de Internet

que sea lineal, contando con la información de los usuarios utilizada en los apartados anteriores y con inclusión de nuevas variables que a continuación se describen.

3. VARIABLES PARA EL MODELO DE GASTO INDIVIDUAL

Modelizar funciones de demanda o de gasto en las telecomunicaciones generalmente no suele ser una tarea fácil, debido a que con frecuencia la mayoría de los datos disponibles son binarios y a veces no se tiene información de los precios pagados por los consumidores. Asimismo, no se tendrá gran confianza acerca de una forma funcional ni sobre la homogeneidad de los datos.

Antes de estimar las funciones de demanda y de gasto a través de Internet, en este apartado se va a describir en detalle las variables que se utilizarán para tal fin. Luego se recogerá la heterogeneidad en los datos utilizando una metodología exploratoria habitual: el algoritmo k-Medias (Mac Queen, 1967; Hartigan y Wong, 1979) para identificar los distintos grupos con características similares en los datos.

La información de cantidades pagadas, bienes demandados por los consumidores o distintos costes asumidos por los vendedores se extrae a partir de variables *proxy* que contengan estos efectos. Las variables adicionales a las ya expuestas (comprar en Internet –*COMPRA*–, quintil de renta –*IS_Q*–, acceso a Internet desde el hogar –*VIVINT*–, cantidad de sitios en donde usar Internet –*USO*–, intensidad de uso del ordenador –*USOR*–, Intensidad de uso de Internet –*USOIN*–, género –*SEXH*–, edad –*EDAD*–, tamaño poblacional –*POBLAC*– y cantidad de miembros en la familia –*TOTMIEM*–) se detallan a continuación en el Cuadro 3, se puede observar que en su mayoría son variables ficticias. Otra cosa que se destaca es que ahora el índice de renta individual tiene sus valores originales, es decir ya no está categorizado:

Tabla 8. Definición de Variables Adicionales

<i>Variable</i>	<i>Definición</i>
G_i	Gasto realizado en productos por la Red (en euros).

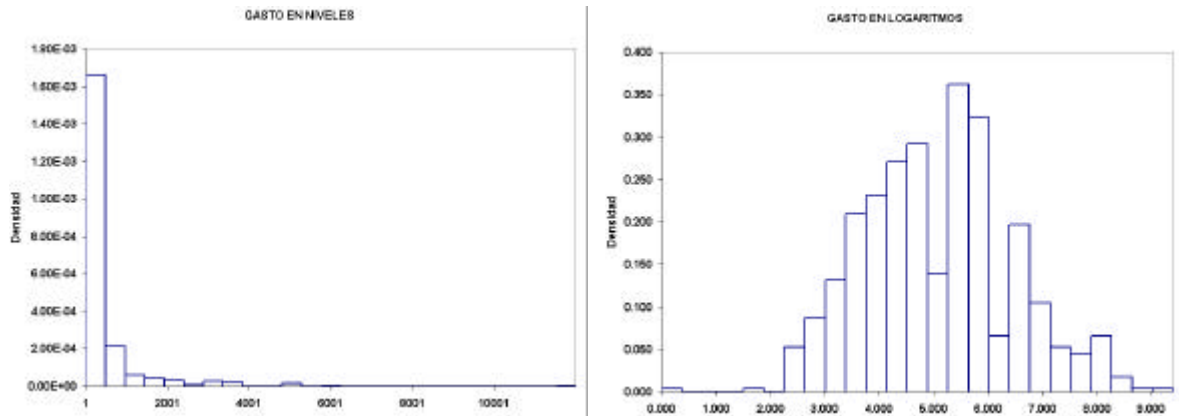
IS_i	Índice de renta individual ³ .
P_1	Ficticia =1 si el individuo adquirió productos del hogar.
P_2	Ficticia =1 si adquirió música.
P_3	Ficticia =1 si adquirió libros.
P_4	Ficticia =1 si adquirió ropa.
P_5	Ficticia =1 si adquirió software.
P_6	Ficticia =1 si adquirió material informático.
P_7	Ficticia =1 si adquirió artículos electrónicos.
P_8	Ficticia =1 si adquirió acciones y títulos financieros.
P_9	Ficticia =1 si adquirió paquetes de viajes.
P_{10}	Ficticia =1 si adquirió entradas a espectáculos.
P_{11}	Ficticia =1 si compró lotería o realizó apuestas online.
P_{12}	Ficticia =1 si adquirió otros productos o servicios.
$CANT_i$	Número de transacciones para adquisiciones en Internet.
$FP1_i$	Ficticia =1 si la forma de pago es con tarjeta de crédito
$FP2_i$	Ficticia =1 si la forma de pago es con transferencia bancaria
$FP3_i$	Ficticia =1 si la forma de pago es contra reembolso
$FP4_i$	Ficticia =1 si la forma de pago es por suscripción

A continuación se hace un análisis gráfico-descriptivo de las variables comparándolas con el gasto en la Red. La variable referida al gasto en euros en la Red (G_i) se refiere al gasto realizado durante los últimos tres meses. La respuesta de los encuestados era de tres tipos: en pesetas, en euros y en categorías predefinidas de gastos (también en pesetas y en euros). A partir de esta información se elaboró la variable G_i , llevando todas las observaciones a valor euro (1 euro = 166,36 pesetas). En las observaciones referidas a intervalos de gasto, se utilizó el punto medio del mismo.

³ Explicada en detalle en el capítulo 2, esta variable está construida a partir de agregar ponderaciones de capital humano y no humano.

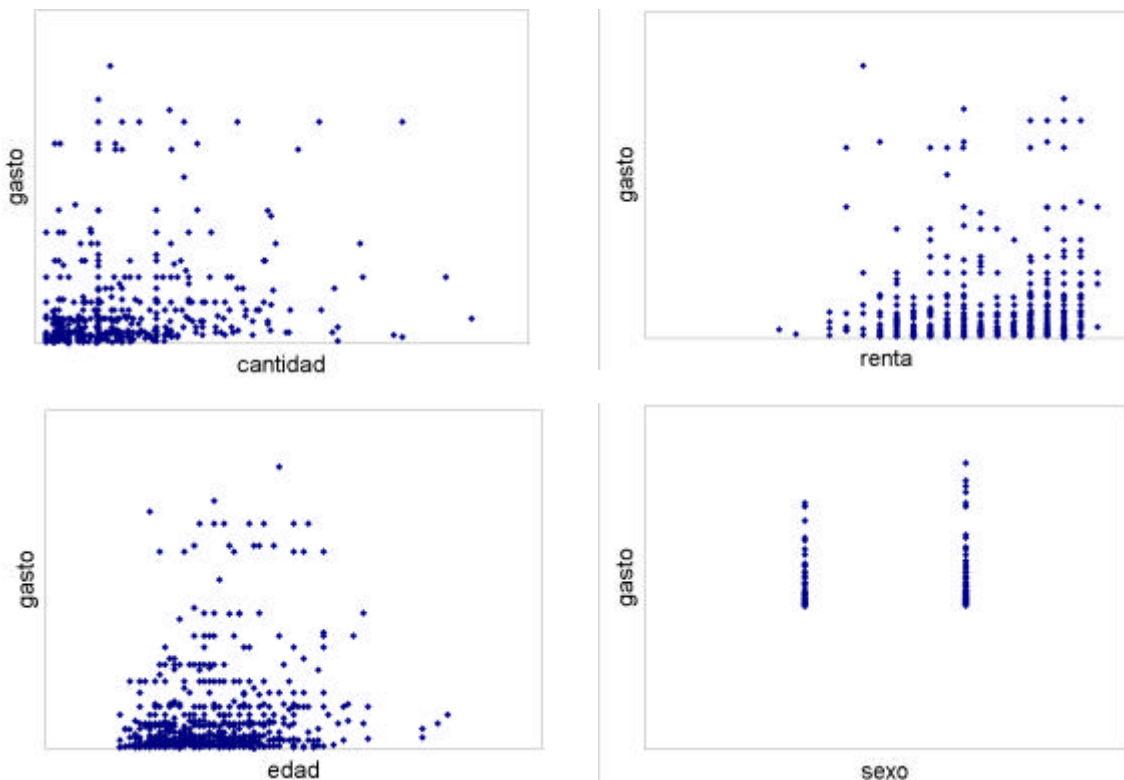
Observando el histograma de la variable gasto G_i en niveles y en logaritmos, se evidencia la existencia de valores extremos:

Gráfico 3. Histograma de Gasto en Internet (en niveles y en logaritmos)



Haciendo uso del algoritmo de detección y exclusión de valores extremos de Peña y Yohai (1999) incorporado en RETINA y aplicándolo al logaritmo del gasto, se detectan y excluyen momentáneamente de la muestra 22 valores extremos. Luego se realizan gráficos bivariantes de la variable Gasto con algunas variables explicativas:

Gráfico 4. Gasto vs. Variables Explicativas Seleccionadas sin Valores Extremos



El método de recogida de la información realizada por el INE se basó en una combinación de procedimientos CATI⁴ (entrevista telefónica) y PAPI⁵ (entrevista personal). Para cada sección de la muestra se dispuso de dos listas de reservas para sustituir las incidencias en las viviendas originalmente seleccionadas. Luego de la recogida de datos se utilizaron distintos tipos de algoritmos de filtro para depurar la muestra. Es decir que es difícil pensar que estos valores extremos equivalgan a valores mal registrados. Además, al ser una muestra de datos de corte transversal, la heterogeneidad puede estar presente. Será necesario recogerla en los posteriores modelos que se vayan a ajustar.

A continuación se procede a detectar grupos de individuos con perfiles homogéneos. Para ello se utiliza la probabilidad individual de comprar en Internet estimada en el apartado anterior, y el k -medias. El algoritmo de k -Medias de Mac Queen (1967) es un clásico método estadístico para detectar subgrupos homogéneos dentro de una muestra. El criterio de homogeneidad que se utiliza es minimizar la suma de cuadrados dentro de los subgrupos para todas las variables dada por:

$$\min SCDG = \sum_{g=1}^G \sum_{j=1}^P \sum_{i=1}^{n_g} (x_{ijg} - \bar{x}_{jg})^2 = \min \sum_{g=1}^G \sum_{j=1}^P n_g s_{jg}^2$$

donde x_{ijg} es el valor para el individuo i de la variable j clasificado dentro del grupo g , y $\bar{x}_{jg}$ es la media de esta variable dentro del grupo. Se puede ver que esto es equivalente a minimizar la suma ponderada de las varianzas de las variables en los grupos. Estas varianzas grupales se las puede considerar una buena medida de heterogeneidad, porque al minimizarlas se obtendrían grupos más homogéneos. El algoritmo k -Medias funciona con las siguientes etapas (Peña, 2002):

- a) Seleccionar los k puntos más alejados entre sí como centros de los grupos iniciales.

⁴ Acrónimo de “Computer Assisted Telephonic Interview”.

⁵ Acrónimo de “Paper and Pencil interview”.

- b) Calcular las distancias euclídeas de cada elemento a los centros de los k grupos y asignando a cada elemento el grupo de cuyo centro esté más próximo.
- c) Definir el criterio de homogeneidad (en nuestro caso será minimizar la $SCDG$) e ir reasignando.
- d) Si luego de la reasignación el criterio no mejora, terminar el proceso.

En este caso se ha agrupado la muestra en 5 grupos a partir de intervalos de probabilidad calculada a partir de los resultados de la estimación del modelo (4.4). Aunque esta cantidad podría reducirse o aumentar dependiendo del grado de agregación y de homogeneidad pretendido dentro de cada grupo, siendo el máximo de grupos igual a la cantidad de observaciones. Se tienen entonces los siguientes grupos:

Tabla 9. Usuarios de Internet Agrupados por Probabilidad de Compra en la Red

GRUPO	Probabilidad de Compra en Internet	Tamaño
Menos Propensos (Grupo 1)	0,010 – 1,101	1.757
Poco Propensos (Grupo 2)	0,101 – 0,191	1.341
Propensión Media (Grupo 3)	0,191 – 0,298	1.037
Propensión Alta (Grupo 4)	0,299 – 0,429	684
Propensión muy Alta (Grupo 5)	0,430 – 0,701	411

Incluyendo estos grupos en forma de constantes específicas dentro del modelo de demanda de bienes en la Red, se recoge la heterogeneidad de los datos observada más arriba.

3.1. MODELO DE DEMANDA DE BIENES ADQUIRIDOS EN LA RED

En este apartado se estimarán los parámetros de una función de demanda de bienes y servicios en Internet. Para especificar dicho modelo se utiliza como variable endógena la variable $CANT_i$ referida a la cantidad de bienes que el individuo ha adquirido en la Red. Esta variable es el número de productos que los individuos han declarado comprar

en Internet. Al ser esta una variable cuyos datos son números enteros no negativos, se utilizará una especificación acorde. Aunque se podría utilizar una regresión mínimo cuadrática, sin embargo por el tipo de datos (gran cantidad de ceros, datos enteros y ninguno negativo) y por la forma que tiene la distribución de la variable endógena, los modelos que más se utilizan son el de Poisson y el binomial negativo, este último utilizado sobre todo cuando los datos tienen sobredispersión. Aquí se especifica una función de demanda de bienes en Internet con un modelo de Poisson para la variable endógena $CANT_i$ en función de la renta y de características tales como cuatro formas de pago consideradas (dentro de las características económicas), el sexo, la edad, el nivel de estudios y la población a la que pertenece (características sociodemográficas). Para recoger la heterogeneidad incluiremos además constantes específicas referidas a cuatro de los cinco grupos detallados en la Tabla 9.

El modelo entonces queda especificado de la siguiente manera:

$$\ln l_i = CANT_i = a_0 + a_1 IS_i + a_2 NIVELEST_i + a_3 SEXO_H_i + a_4 EDAD_i + \\ + a_5 POBLAC_i + \sum_{k=1}^4 f_{ik} FP_k + \sum_{m=1}^5 x_{im} GR_m + u_i$$

en donde $i = 1, \dots, 5.218$. Estos son sólo los que sí compran en Internet, las variables son las que están descritas más arriba en las Tablas 1 y 8 (índice de renta individual, nivel de estudios alcanzado, pertenencia al género masculino, edad y población), FP_k se refiere a las cuatro formas de pago de las compras consideradas, y GR_m se refiere a los grupos por categoría de probabilidad de comprar a través de Internet, y donde l_i es el número de veces que el individuo i ha comprado por Internet en un período dado, y equivale a la esperanza condicional de una variable aleatoria con distribución de Poisson.

3.2. RESULTADOS Y DISCUSIÓN

La siguiente tabla muestra los resultados de la estimación:

Tabla 10. Resultados del Modelo de Demanda

Variable			
Dependiente	Coeficiente	Ratio de	<i>z</i>
	Estimado	Incidencia	
<i>CANT</i>			
Constante	-3,63	---	17,08
<i>IS</i>	0,42	1,51	4,67
<i>NIVELEST</i>	-2,20	0,11	4,74
<i>SEX _H</i>	0,12	1,13	1,93
<i>EDAD</i>	0,002	1,00	0,77
<i>POBLAC</i>	0,10	1,01	1,69
<i>FP1</i>	2,42	11,23	36,00
<i>FP2</i>	0,82	2,27	10,93
<i>FP3</i>	1,32	3,75	19,79
<i>FP4</i>	0,36	1,42	2,72
<i>GR1</i>	-0,79	0,45	7,69
<i>GR2</i>	-0,22	0,81	2,84
<i>GR3</i>	-1,14	0,32	5,14
<i>GR4</i>	0,11	1,01	0,13
<i>GR5</i>	---	---	---
Observaciones: 5.218			
Log-Verosimilitud: -1.914,34			
$c^2_{(13)}$: 3635,19 (Prob. = 0,000)			
Pseudo R ² : 0,4870			

A partir de los resultados de la Tabla 10 podemos decir que:

- El número esperado de artículos adquiridos en Internet se incrementa notablemente con la posibilidad de pagar con tarjeta de crédito. Su ratio de incidencia es el más grande de todos con diferencia (11,23). Esto quiere decir

que si el individuo posee una tarjeta de crédito (*FP1*), incrementa su número esperado de artículos adquiridos en Internet en mucho más del doble.

- La incidencia del resto de las formas de pago, aunque bastante menor que la que produce el pago con tarjeta de crédito, también es importante. Los pagos con transferencia bancaria (*FP2*) y contra reembolso (*FP3*) son los que inciden en 2,27 y 3,75 respectivamente. Observamos que el impacto del pago contra reembolso es mayor que el pago con transferencia bancaria. El pago por suscripción (*FP4*) es también importante, pero su incidencia es la menor de todas (1,42).
- El nivel de renta individual (representado por la variable *IS*) también tiene una incidencia positiva en la cantidad de bienes y servicios adquiridos por comercio electrónico. Su incidencia es de 1,51. Es decir que el incremento unitario del valor del índice *IS* equivale a un aumento en la cantidad esperada de transacciones para adquirir bienes y servicios en la Red del 51%.
- En cuanto a las constantes específicas de los cinco grupos de individuos, hemos eliminado del modelo al grupo del último intervalo de probabilidad (el de probabilidades más altas) para evitar problemas de multicolinealidad exacta en la función de verosimilitud. Vemos que a medida que aumenta la probabilidad de comprar en Internet, como es lógico, los ratios de incidencia tienden a aumentar. Esto quiere decir que las personas más propensas a comprar en Internet adquieren mayor cantidad de productos mientras que con los menos propensos ocurrirá lo contrario.
- Atributos sociodemográficos como la edad, el nivel de estudio o el tamaño de la población tendrán una escasa influencia en la cantidad esperada de comercio en la Red.

Resumiendo entonces, se ha considerado un modelo para los determinantes de la probabilidad de comprar en Internet. Para ello se ha utilizado la muestra de los que se declaran usuarios de Internet. Una vez obtenida, se ha agrupado a los individuos por su propensión a la compra en Internet y con esto se recogió la heterogeneidad en un modelo de demanda de bienes y servicios a través de Internet utilizando como variable endógena el número de bienes adquiridos a través de la Red. A continuación se

especifica y se estima una función de gasto en comercio electrónico lineal en los parámetros.

3.3. FUNCIÓN LINEAL DE GASTO EN COMERCIO ELECTRÓNICO

La predicción del gasto es importante desde muchos puntos de vista. Permite evaluar la conducta o descubrir perfiles del consumidor en Internet, sirve para determinar nuevas formas de financiación a partir de estos perfiles, etc. En este apartado se buscará obtener un modelo de predicción para el gasto. Para ello, a continuación se incluirán los grupos de la Tabla 9 en forma de constantes específicas dentro de un modelo MCO utilizando como variable endógena el logaritmo del gasto. La especificación es la siguiente:

$$\log(G_i) = g_0 + g_1 \log(IS_i) + g_3 \log(EDAD_i) + g_4 SEXO_H + \sum_{g=1}^4 d_{ig} G_i + \sum_{p=1}^{12} x_{ip} P_i + u_i \quad (4.7)$$

Como se observa, se han incluido las constantes específicas d_{ig} y x_{ip} correspondientes a los cinco grupos de individuos del apartado anterior y a los doce productos considerados, respectivamente.

El fin ahora es conseguir un modelo para previsión que además mejorará la especificación. Para ello se utilizará el algoritmo RETINA de Pérez Amaral, Gallo y White (2003) y con el cual se seleccionará un modelo con buena capacidad predictiva por fuera de la muestra. En la Tabla 11 se observan las variables que incluyó RETINA (los resultados de las estimaciones se encuentran en el Apéndice):

Tabla 11: Detalle de las Variables Utilizadas

Variable Endógena	$\ln(G)$
<i>Variables Continuas Originales</i>	$\ln(IS), \ln(EDAD)$
<i>Constantes Específicas</i>	$SEXO_H, G_g$ donde $g = 1, \dots, 5$
<i>Pendientes Específicas</i>	$SEXO_H, G_g, P_p$ donde $p = 1, \dots, 12$

Recordar que la variable G_i hace referencia a los cinco grupos de individuos detectados en el apartado 3.1, y P_p hace referencia a los doce grupos de bienes y servicios considerados que se adquieren en Internet.

RETINA también es útil cuando no se tiene una hipótesis de especificación clara en cuanto a qué variables hay que considerar en el modelo. En este caso, la inclusión de constantes y pendientes específicas como regresores ha mejorado la explicación de las variaciones del gasto en Internet, y eso se nota a partir de la mejora de la previsión fuera de la muestra, el AIC y la bondad de ajuste dentro de la muestra ($\bar{R}^2$). En la siguiente Tabla 12 se presentan los principales resultados comparativos, en donde se destaca la mejora en la previsión fuera de la muestra con una importante reducción del Error Cuadrático Medio de Predicción Robusto (RCMSPE) de 1,195 a 0,689.

Tabla 12: Comparación entre el Modelo Básico (BLM) y el Modelo RETINA (URM) para la Función de Gasto

	MCO (BLM)	RETINA (URM)
Parámetros	8	33
AIC	0,356	-0,813
$\bar{R}^2$	0,260	0,779
RCMSPE	1,195	0,689

Se puede observar como la inclusión de constantes específicas, pendientes específicas y productos cruzados de los regresores mejoran los resultados globales. Además del RCMSPE, también se observa que el $\bar{R}^2$ pasa de un valor de 0,260 a 0,779, que se puede considerar bueno, sobre todo teniendo en cuenta que esta es una medida que penaliza la inclusión de nuevos parámetros, observando que el número de parámetros aumenta de 8 a 33 (se ve en la Tabla 11 que las 12 variables ficticias referidas a los productos adquiridos sólo se incluirán como pendientes específicas). También el AIC mejora mucho pasando del valor 0,356 a -0,813.

De los resultados de las regresiones podemos decir:

- Las variables continuas originales consideradas en los BLM referidas al logaritmo de la edad y al logaritmo de la renta, $\ln(EDAD)$ y $\ln(IS)$, serán ambas significativas en el modelo MCO sin constantes específicas. Sin embargo, al incluir las constantes específicas de los grupos vemos que sólo será significativa el $\ln(EDAD)$.
- Se observa que en el modelo RETINA los regresores originales sólo entrarán en forma de interacciones (productos cruzados entre variables explicativas continuas) o pendientes específicas (productos cruzados en donde una de las dos variables es binaria). Esto quiere decir que las correlaciones individuales más altas entre la variable endógena y los regresores se darán cuando estos últimos se refieran a interacciones o a pendientes específicas, y no como se presentan originalmente en el modelo MCO.
- Con respecto a los cuatro grupos de consumidores considerados, se observa que si bien en el modelo MCO son todos significativos (el G_V no lo incluimos evitando así problemas de multicolinealidad, pero su efecto está en el término constante), en el modelo RETINA el único grupo que aparece es el G_{II} en forma de pendiente específica.
- Se observa además que incluir variables binarias referidas a los productos es importante para recoger la heterogeneidad. Se puede observar que aparecerán 11 de las 12 en el modelo MCO en forma de pendientes específicas. Esto hace que el $\bar{R}^2$ aumente (a pesar de ser una medida que penaliza la inclusión de regresores) hasta valores aceptables que indican un ajuste (0,779). A su vez, el RCMSPE disminuirá casi hasta la mitad pasando de 1,195 a 0,689.
- Como conclusión, se puede agregar que recoger la heterogeneidad e incluir en el modelo no linealidades en los regresores, puede mejorar sustancialmente tanto el ajuste como la habilidad predictiva, cuando se usan modelos sugeridos por RETINA. A su vez observamos que los productos adquiridos a través de Internet y considerados en la muestra serán buenos predictores de la demanda de comercio electrónico y que los regresores continuos originales explicarán mejor dicha demanda a partir de ser no lineales.

4. CONCLUSIONES

En este artículo se ha abordado la modelización de la demanda de comercio electrónico en España usando tres modelos que responden a tres preguntas:

- a. ¿Quiénes usan el comercio electrónico?
- b. ¿Su uso es casual o frecuente?
- c. ¿Cuánto dinero gastan?

El primer modelo está referido al perfil del demandante de la Red teniendo en cuenta principalmente el vínculo existente entre el comercio electrónico como tal y el acceso y uso de Internet. El segundo está referido a la cantidad de transacciones (demanda de bienes) en la Red. El tercero corresponde a una función lineal de gasto en Internet.

Los resultados de la primera especificación confirman la existencia de este vínculo y de que los determinantes del demandante serán su nivel de renta, la edad y el nivel de estudios alcanzado del individuo. El cálculo de las probabilidades de la compra a través de Internet revela que esta será más sensible a los hábitos de acceso y uso de Internet. En cambio el uso del ordenador, el género sexual y el hábitat, aunque también influyen, lo harán en menor medida.

El análisis gráfico bivalente del gasto en euros a través de la Red muestra la existencia de datos anómalos y heterogeneidad en los datos que deberán ser detectados y tratados para alcanzar una representación aceptable. Para ello se utiliza el algoritmo k-medias en donde se detectan 5 grupos de consumidores.

Posteriormente se usa un segundo modelo para los determinantes de la demanda de comercio electrónico. Para ello se especifica y estima un modelo de Poisson para la cantidad de bienes demandados a través de la Red. Se concluye que las distintas formas de pago de los productos y servicios de la Red (tarjeta de crédito, transferencia bancaria, contra reembolso) influirán notoriamente en la cantidad de bienes adquiridos, seguidos por el nivel de renta y los distintos atributos sociodemográficos considerados.

El último modelo planteado para tratar la demanda de comercio electrónico será la especificación y estimación de una función lineal de gasto. Para tal tarea se utiliza el algoritmo RETINA para obtener un modelo flexible con buena capacidad predictiva fuera de la muestra. RETINA generará e incluirá en el modelo nuevos regresores a partir de los originales, que mejorará sustancialmente el ajuste y disminuirá el RCMSPE casi hasta la mitad.

En cuanto a lo que se aprende a partir de estos modelos es que el perfil del individuo demandante de bienes y servicios en Internet en España es el de una persona joven, con estudios avanzados, perteneciente a un ámbito poblacional mediano y con un nivel de renta mediano que le da acceso al crédito y le permite realizar pagos con tarjetas de crédito u otras maneras alternativas.

A su vez, con estos resultados se podría decir que los modelos planteados para los determinantes del comercio electrónico para este tipo de datos serían adecuados para realizar un estudio pormenorizado de los determinantes del comercio electrónico en España. Se considera que la metodología que se plantea para el tratamiento de datos de corte transversal con heterogeneidad en la muestra se podría también utilizar en otros contextos.

En futuras investigaciones sería interesante analizar paneles de datos e incluir dinámica dentro de los modelos especificados.

5. REFERENCIAS

Angrist, J. (2004) "Treatment Effect Heterogeneity in Theory and Practice" *Economic Journal*, Royal Economic Society, vol. 114(494), pp. C52-C83

Bakos, Y. (1997) "Reducing Buyer Search Cost: Implications for Electronic Marketplaces" *Management Science*, vol.43, Issue 12

Bakos, Y. (2001) "The Emerging Landscape for Retail E-Commerce" *Journal of Economic Perspectives*, vol. 15, nº 1, pp. 69-80

Brynjolfsson, E. y Smith, M (1999) "Frictionless Commerce? A Comparison of Internet and Conventional Retailers" Working Paper, MIT Sloan School of Management

Cerno, L. y Pérez-Amaral, T. (2005) "Demand of Internet Access and Use in Spain" Documento de Trabajo 0506. Instituto Complutense de Análisis Económico (ICAE)

Cerno, L. y Pérez-Amaral, T. (2006) "Medición y Determinantes de la Brecha Tecnológica en España", mimeo

Fundación AUNA (2003) "Informe España 2003", disponible en http://www.fundacionauna.com/areas/25_publicaciones/

Greene, W. (2000) "Econometric Analysis" 4th. Edition, Prentice Hall

Hartigan, J. y Wong, M. (1979) "A k-means Clustering Algorithm" Applied Statistics, 28, 100-108

Heckman, J. (1978) "Dummy Endogenous Variables in a Simultaneous Equation System" Econometrica 46: 931-959

Hotelling, H. (1929) "Stability in Competition" The Economic Journal, pp. 41-57

INE, "Encuesta Sobre Equipamiento y Uso de Tecnologías de Información y Comunicación de los Hogares del Segundo Trimestre de 2005" disponible en www.ine.es/prensa/np388.pdf

Mac Queen, J. (1967) "Some Methods for Classification and Analysis of Multivariate Observations" Proc. Symposium of Mathematics, Statistics and Probability, 5th, pp. 281-297, Berkeley

Maddala, G. (1983) "Limited Dependent and Qualitative Variables in Econometrics" Cambridge University Press

Marinucci, M. y Pérez-Amaral, T. (2005) “Econometric Modeling of Business Telecommunications Demand Using RETINA and Finite Mixtures” Documento de Trabajo 0501. Instituto Complutense de Análisis Económico (ICAE).

Mas-Colell, A., Whinston, M. y Green, J. (1995) “Microeconomic Theory” Oxford University Press

McKinsey, J. (2000) “Internet Pricing. A Creator of Value Not a Destroyer. McKinsey Marketing Practice” Documento disponible en http://marketing.mckinsey.com/solutions/McK-Internet_Pricing.pdf

Peddibhotla, N. (2000) “Are Economic Theories Good Enough to Understand E-Commerce?” en Wiseman, A. “The Internet Economy: Access, Taxes, and Market Structure”, Brookings Institution Press, Washington D.C.

Peña, D. (2002) “Análisis de Datos Multivariantes”, Mc Graw Hill

Peña, D. y Yohai, V. (1999) “A Fast Procedure for Outlier Diagnostic in Large Regression Problems” Journal of the American Economic Association, 94: 434-445

Pérez-Amaral, T. Gallo, G. y White, H. (2003) “A Flexible Tool for Model Building: the Relevant Transformation of the Inputs Network Approach (RETINA)” Oxford Bulletin of Economics and Statistics, 65:821-838

Red.es (2004) – Ministerio de Industria, Turismo y Comercio “Las TIC en los Hogares Españoles” disponible en <http://observatorio.red.es/estudios/consumo/index.html>

Shapiro, C. y Varian, H. (1998) “Information Rules: A Strategic Guide to the Network Economy”, Harvard Business School Press

Smith, M., Bailey, J. y Brynjolfsson, E., (1999) “Understanding the Digital Markets: Review and Assessment” en Brynjolfsson, E. y Kahin, B. “Understanding the Digital Economy”, MIT Press

Train, K. (2002) "Discrete Choice Methods with Simulation" Cambridge University Press

Varian, H. (1992) "Análisis Microeconómico" 3ra. Edición, Antoni Bosch ed.

Wilde, J. (2000) "Identification of Multiple Equation Probit Models with Endogenous Dummy Regressors". Journal of Econometrics 30: 415-443

Zellner, A. (1962) "An Efficient Method of Estimating Seemingly Unrelated Regressions and Test of Aggregation Bias" Journal of the American Statistical Association, 57, pp. 500-509

APÉNDICE: Función Lineal de Gasto recomendada por RETINA

Observaciones	614	
R ² ajustado	0,779	
Error estándar de la estimación	0,647	
RCMSPE	0,689	
AIC	-0,813	
Variable	Coeficiente	Estadístico t
Constante	7,87	5,65
Interacciones		
$1/(\log(EDAD))^2$	-29,28	3,70
$(\log(EDAD))^2$	-0,19	3,24
$1/\log(IS)$	-0,12	1,59
Pendientes Específicas		
$G_{II}/\log(EDAD)$	-3,73	5,53
$P_1/\log(EDAD)$	0,72	2,71
$P_1/[\log(EDAD)]^2$	-0,12	1,58
$P_2/[\log(EDAD)*\log(IS)]$	4,40	2,77
$P_2/[\log(EDAD)]^2$	-0,20	2,37
$P_2*[\log(EDAD)/\log(IS)]$	-0,43	2,46
$P_2*\log(IS)$	-9,54	4,35
$P_2*[\log(IS)]^2$	-6,94	4,33
$P_3/\log(EDAD)$	3,04	13,63
$P_3*\log(EDAD)*\log(IS)$	-0,45	2,01
$P_4*[\log(EDAD)/\log(IS)]$	-0,11	1,46
$P_4/[\log(EDAD)]^2$	-0,07	0,90
$P_5*\log(EDAD)*\log(IS)$	-0,67	2,91
$P_5/[\log(EDAD)]^2$	7,85	2,66
$P_5*[\log(IS)/\log(EDAD)]$	6,04	1,87
$P_6*\log(EDAD)*\log(IS)$	-0,21	3,36
$P_6/[\log(EDAD)*\log(IS)]$	-0,87	3,98
$P_7/\log(EDAD)$	2,59	2,57
$P_7*\log(EDAD)*\log(IS)$	-0,14	1,84
$P_7/[\log(EDAD)*\log(IS)]$	1,47	1,23
$P_8*[\log(EDAD)]^2$	0,25	1,97
$P_8/[\log(EDAD)]^2$	100,78	1,92

$P_8/\log(EDAD)$	-37,06	1,79
$P_9 * [\log(EDAD)/\log(IS)]$	-0,07	5,46
$P_9 * \log(IS)$	-0,95	6,21
$P_{10}/[\log(EDAD)*\log(IS)]$	-3,15	2,75
$P_{10}/\log(IS)$	0,90	2,49
$P_{10} * [\log(EDAD)]^2$	0,08	4,75
$P_{11} * [\log(EDAD)]^2$	0,05	2,63

Abstract

In this article we analyze the factors that affect the e-commerce for the Spanish case. We begin with a model specification for the determinants of the buy in the Web taking into account the existing link among the acquisition of goods and the access and use of the Internet service. The results reveal that the accessibility to the Net will be the principal factor followed by the income and the genre that impact in the probability of the electronic buy. Then we raising a model determined to the users of the e-commerce for the determinants of the expenditure in the Web. We discover that bring it the disparity (dispersion) in the data will be important for this type of models, and to obtain a final model we use the algorithm RETINA, which allows the treatment of the heterogeneity and improves substantially the final results of the model.