Facultad de Economía, Empresa y Turismo
Universidad de La Laguna

MEMORIA DEL TRABAJO FIN DE GRADO

Estimación de la probabilidad de elección entre alternativas binarias: una aplicación al consumo de tabaco en España

(Estimation of the probability of choice between binary alternatives: an application to tobacco consumption in Spain)

Nieto González, Imanol Lorenzo

Tutores:
D. Ginés Guirao Pérez
D. Domingo Jesús Lorenzo Díaz

Grado en Administración y Dirección de Empresas
Facultad de Economía, Empresa y Turismo
Curso Académico 2018/2019
Convocatoria de julio

San Cristóbal de La Laguna, a 9 de julio de 2019

Este trabajo aborda un análisis de los factores demográficos, económicos y sociales que explican la decisión que toma un individuo ante el hecho de consumir tabaco o no. Los datos de la Encuesta Nacional de Salud (2017) serán los que sean utilizados como principal fuente de información. La especificación y estimación de los modelos de elección discreta, en particular, de un modelo logit, será lo que permita explicar la decisión, en un sentido u otro, todo ello basado en el fundamento teórico que otorga la Teoría de la Utilidad Aleatoria. Una vez obtenidos los principales resultados, se observa como los hombres, entre 40 y 60 años, con estudios secundarios, nacidos en España y con hábitos como la no realización de actividad física y el consumo de alcohol de forma habitual, son los que muestran una mayor probabilidad de fumar.

Palabras clave: *fumar, modelos de elección discreta, modelo logit, utilidad aleatoria.*

This paper discusses an analysis of the demographic, economic and social factors that explain the decision considered for the individual about smoke or not smoke. The data from the National Health Survey (2017) becomes the main source of information. The specification and estimation of the discrete choice models, in particular, a logic model, pretend to explain the decision can be applied, in one sense or another, everything is based on the theoretical background that the Theory of Random Utility guarantees. Once the main results are obtained, it is observed as men, between 40 and 60 years old, with secondary studies, born in Spain and with habits such as the non-realization of physical activity and the habitual consumption of alcohol, are those that show a more likely to smoke.

Keywords: *smoking, discrete choice models, logit model, random utility.*

Contenido

En el siguiente enlace se pueden consultar los anexos de este trabajo: https://cutt.ly/ET7TZR

Índice de tablas:

Índice de gráficos:

1. INTRODUCCIÓN

La creciente preocupación por llevar un estilo de vida saludable, materializado en una mejor y más cuidada alimentación y en la adopción de hábitos saludables, como el ejercicio físico diario, entre otros, constituye el principal elemento motivacional de este trabajo. Pues, a pesar de esta clara tendencia, hábitos como el consumo de alcohol y de tabaco se siguen dando en una proporción importante de la población española (22,08% de fumadores en 2017). Incluso, resulta aún más preocupante observar esta conducta en individuos cada vez más jóvenes, a pesar de las demostradas consecuencias que esto tiene sobre la salud. Por ello, en la actualidad, el consumo de tabaco es, sin duda, uno de los principales problemas de salud a los que se enfrenta la sociedad, junto con el consumo de alcohol y la obesidad.

Por supuesto, este no pretende ser un documento que contribuya a la prevención de estas características en aras de eliminar el consumo, pues son, en sí mismos, factores inherentes al propio individuo. Sin embargo, sí que se propone hacer una radiografía que permita definir, al menos en gran parte, al consumidor tipo y, a partir de ahí, estudiar posibles medidas de prevención, asunto que se escapa del alcance de este estudio, pero que podría resultar crucial a la hora de elaborar programas específicos de prevención enfocados a los colectivos de riesgo.

Este trabajo pretende, entonces, analizar los factores sociales que llevan a que el individuo consuma tabaco. Por tanto, el objetivo que persigue es el de estudiar los factores demográficos y económicos que explican la propensión del consumidor español de tabaco, así como la probabilidad asociada a esta. Para la consecución del objetivo se articulan algunas propuestas de modelos de elección discreta, particularmente la especificación de modelos logit, que permita evaluar, de acuerdo con la base teórica en la que se fundamenta este trabajo, la Teoría de la Utilidad Aleatoria, la mayor o menor relación que tiene con el hecho de fumar las características propias del individuo.

La fuente de datos utilizada será, principalmente, la Encuesta Nacional de Salud 2017, que elabora el Instituto Nacional de Estadística (INE). Particularmente se trabajará con las encuestas realizadas a sujetos con 15 años o más. Se trata, en consecuencia, de unos datos de corte transversal que permitirán definir, para ese momento concreto, al fumador español.

El contenido se articula en cinco capítulos. Se abordan, en el capítulo uno y dos, la introducción y los antecedentes, realizando una revisión de la literatura, un análisis temporal del consumo de tabaco y la propuesta de algunos conceptos operativos. En el capítulo tres se explica y detalla la base de datos empleada, así como las distintas recodificaciones que se han realizado para adaptarla a las necesidades de este estudio. También en este apartado se aborda la base teórica y metodológica que se emplea, haciendo un recorrido por los principales modelos de variable dependiente limitada. El capítulo cuarto, uno de los más extensos, desarrolla un amplio análisis descriptivo de la muestra y de los pesos relativos que, en cada colectivo, presentan los individuos que fuman. De esta forma, se pretende obtener posibles resultados esperados que puedan ser, en el capítulo cinco donde se aborda la estimación, contrastados con los resultados extraídos de los modelos. Finalmente, tras el capítulo quinto, se relacionan las conclusiones de este trabajo, además de las limitaciones detectadas y las propuestas de continuación que podrían derivarse a partir del mismo.

2. ANTECEDENTES

En 2006, con la entrada en vigor en España de la primera ley que regula[1], entre otras, la venta y el consumo de tabaco, más conocida como "ley antitabaco", comienza un proceso que pretende reducir el consumo de este por los más que probados riesgos que tiene para la salud. La propia exposición de motivos de la ley lo describe como tal en su exposición de motivos[2]. Además, el riesgo que supone este producto va más allá del propio consumidor, pues el humo de tabaco en el ambiente, aquel que respira el fumador pasivo, constituye también una importante causa de mortalidad. En total, en España, el tabaco provoca 52.000 muertes al año[3].

En este sentido, y aludiendo también al texto constitucional, en su artículo 43.1.2.[4], donde se encomienda a los poderes públicos el apoyar la prevención y el cuidado de la salud, el Estado da por iniciada la lucha antitabaco.

Ahora bien, a pesar de las restricciones que esta ley impuso, no es hasta 2010, con la promulgación de la denominada "nueva ley antitabaco"[5], que venía a modificar la preexistente, que se batalla de forma definitiva contra el consumo de tabaco. Entre otras medidas, la nueva ley ampliaba la prohibición de fumar a cualquier espacio público, salvo que este estuviera al aire libre. Con este cambio normativo, no solo se avanzaba en el camino hacia la eliminación del consumo, sino que se protegía a aquellos que involuntariamente, fumadores pasivos, veían perjudicada su salud por el mero hecho de frecuentar un lugar público donde se permitía fumar.

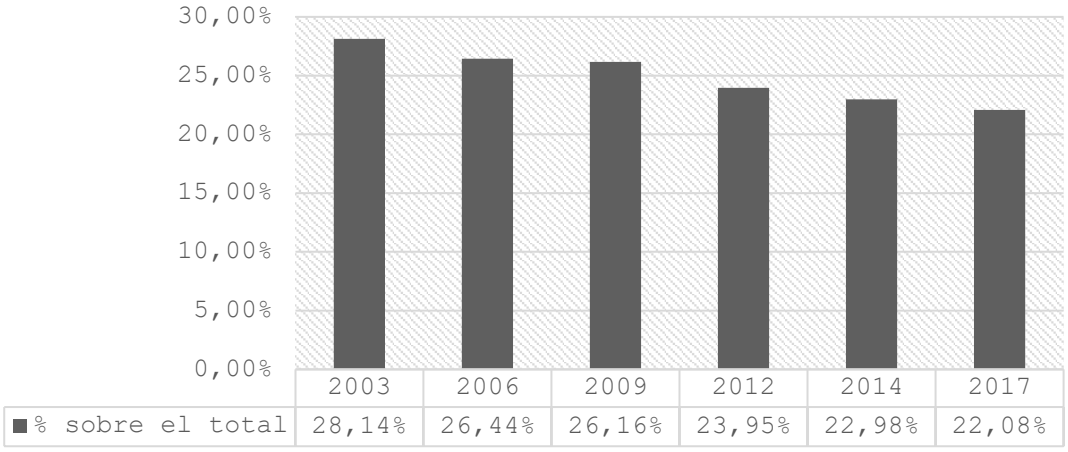
Por supuesto, la modificación no estuvo a salvo de detractores que aludían a ciertos asuntos que habría que sopesar si deben prevalecer frente a la protección de la salud. Estos hablaban de quebrantamiento de libertades individuales[6], como la libertad de empresa, que se veía – según sus detractores – claramente recortada al no depender de la voluntad del dueño del establecimiento la posibilidad de fumar en sus dependencias. Pérdidas económicas, que pudieran acabar en pérdidas de empleos, aumento del contrabando, etc., son algunas de las razones que argumentaban su malestar.

Es probable que otros motivos, como los mencionados en la introducción de este trabajo, hayan contribuido a la reducción del número de personas que, en España, son fumadoras a diario, según datos del Instituto Nacional de Estadística, pero, indudablemente, la entrada en vigor de estas leyes ha provocado una clara desincentivación del consumo de tabaco. El siguiente gráfico (gráfico 1) muestra la evolución temporal del consumo a diario de tabaco, haciendo un recorrido que se inicia en 2003, previo a la implantación de la primera ley, y que finaliza en 2017, año que servirá

[1] Ley 28/2005, de 26 de diciembre, de medidas sanitarias frente al tabaquismo y reguladora de la venta, el suministro, el consumo y la publicidad de los productos del tabaco.
[2] "En España, al igual que en otros países desarrollados, el tabaquismo es la primera causa aislada de mortalidad y morbilidad evitable. La evidencia científica sobre los riesgos que conlleva el consumo de tabaco para la salud de la población es concluyente" (Exposición de motivos de la Ley 28/2005).
[3] https://cutt.ly/wlnb7C
[4] Artículo 43.1. de la Constitución Española (1978):
"2. Compete a los poderes públicos organizar y tutelar la salud pública a través de medidas preventivas y de las prestaciones y servicios necesarios. La ley establecerá los derechos y deberes de todos al respecto".
[5] Ley 42/2010, de 30 de diciembre de 2010, por la que se modifica la Ley 28/2005, de 26 de diciembre, de medidas sanitarias frente al tabaquismo y reguladora de la venta, el suministro, el consumo y la publicidad de los productos del tabaco.
[6] https://cutt.ly/ilbvDw

de referencia en este trabajo. En 2003 el porcentaje de fumadores a diario ascendía hasta el 28% de la población, a partir de ahí, ha ido experimentando ligeras caídas hasta el 22% que marca el año 2017, por tanto, 6 puntos porcentuales menos en 14 años. El descenso experimentado ha sido relativamente lento, sobre todo antes de la entrada en vigor de las leyes antitabaco, pues la caída entre 2006 y 2012 es la más pronunciada, casi tres puntos, y desde entonces ha ido consolidándose esa tendencia a la baja. No obstante, en 2014, según datos de la Encuesta Europea de Salud, la media de la Unión Europea de los 28 era del 19% y, por tanto, España se situaba casi cuatro puntos porcentuales por encima de la media en ese año.

Gráfico 1 Evolución del número de fumadores a diario en España (2003 – 2017)



	2003	2006	2009	2012	2014	2017
% sobre el total	28,14%	26,44%	26,16%	23,95%	22,98%	22,08%

Fuente: Elaboración propia a partir de los datos de la Encuesta Nacional de Salud de España (INE: 2003, 2006, 2012 y 2017) y la Encuesta Europea de Salud en España (INE:2009 y 2014) Los porcentajes son poblacionales (factor de elevación)

Algo que resulta aún más preocupante y que cuyas cifras serán tratadas más adelante (capítulo 4: análisis descriptivo de los datos) es la edad a la que empiezan a consumir tabaco los jóvenes. Por ejemplo, según los datos de la Encuesta Nacional de Salud de España 2017, casi el 9% de los encuestados con 15 y 16 años, fuman a diario actualmente, a pesar de estar prohibida su venta a menores de edad. Sin embargo, la cifra más alarmante es que el 74% de los fumadores a diario, lo son desde los 18 años o antes.

Sin duda, constituye un objetivo a corto plazo para las autoridades sanitarias, detectar los factores sociales y particulares que llevan, a este colectivo, a consumir tabaco a una edad tan prematura. Algunos autores como Sánchez, Moreno, Muñoz y Pérez (2007) apuntan a que los jóvenes que consumen sustancias de uso frecuente en nuestra sociedad, como el tabaco y alcohol, tienen amigos que realizan este tipo de consumos, convirtiendo, por tanto, al entorno social en un claro fundamento del inicio en el consumo. Añaden que “el consumo de sustancias habituales en nuestra sociedad ha terminado formando parte de un sistema más amplio de ritos de transición a la adultez” (Sánchez, 2007, p. 305). Cabe añadir, de acuerdo con la necesidad descrita al inicio del párrafo, que han fracasado los programas preventivos que se han aplicado en esta línea hasta el momento. Algunos apuntan a la generalidad de estos y a la no adaptación a las realidades individuales de los colectivos y grupos como causas de la frustración de los objetivos de los programas (Inglés et al, 2007).

En esta línea, sobre aquello que explica la propensión a fumar de un individuo, en este trabajo se emplean variables de tipo sociodemográfico, esto es, variables explicativas del propio individuo. A pesar de que la base teórica empleada sea la Teoría de la Utilidad Aleatoria (Domencich y McFadden, 1975) y esta permita la utilización de variables propias del individuo y atributos del producto[7], en esta ocasión, solo se trabajará con las primeras, dado que no se disponen características propias del producto consumido por cada uno de los sujetos encuestados. Particularmente, dado que también esta teoría es aplicable a cualquier alternativa planteada (independientemente de las variables, pero dentro del marco descrito), las variables de tipo sociodemográficos son las más usuales en estudios de este tipo, cuando se pretende modelizar el comportamiento de un individuo frente a la toma de una decisión. Algunos trabajos que las han empleado, para consumos de otros bienes o servicios, son, por ejemplo, el de Guirao, Rodríguez, Cano y Romero (2016) donde estudian el consumo del vino, a partir de modelos de probabilidad, y el de González, Guirao y Pérez (1995) que constituye una aplicación similar en el campo de la Economía del Transporte. Adicionalmente, se incluyen variables propias del consumo de tabaco (frecuencia de consumo, intentos de dejarlo…) y determinantes del estado de salud del individuo (consumo de alcohol, práctica habitual de ejercicio físico, estado de salud percibido, etc.), con la intención de enriquecer el análisis y verificar, si así es finalmente, relaciones esperadas que se traduzcan en resultados empíricos del trabajo.

Por último, a continuación, se detallan algunos conceptos operativos que serán empleados en el trabajo y que conviene, a efectos esclarecedores, definir:

- Tabaco: se incluirán, no solo el tabaco convencional, sino también otros métodos como vapeadores, puros, tabaco en pipa, tabaco de liar, cigarrillo electrónico y otros similares.

- Fumador: "Se incluyen tanto los individuos que consumen tabaco de forma habitual como los que lo hacen de forma esporádica", definición extraída de la web del Consejo Estatal de Estudiantes de Medicina de España. Esta definición se asemeja a otras encontradas en fuentes tanto nacionales como internacionales. En este sentido, todas tienen en común la dificultad de definir qué se debe considerar o no, fumar de "forma esporádica". A los efectos de este trabajo, se considera fumador solo a aquel que consume tabaco diariamente, dado que se pretende explicar los factores que definen la propensión a fumar, por lo que parece más apropiado centrar el estudio en los fumadores habituales dado que el hábito de compra de estos sí que permitiría evaluar, por ejemplo, el efecto de medidas de tipo económico (carga impositiva, precio de la cajetilla, etc.) o las consecuencias que tiene sobre su salud. Por tanto, solo se tendrá en cuenta a los catalogados por la encuesta como "fumador a diario".

- Consumidor español: residente en territorio nacional, así se consigue adaptar el estudio a la principal fuente de datos (Encuesta Nacional de Salud 2017).

[7] "La probabilidad de que los individuos elijan una determinada alternativa es función de sus características socioeconómicas y de la relativa atractividad de la alternativa" (Ortúzar y Willumsem, 2008, p. 334).

3. METODOLOGÍA

En este apartado se aborda, en primer lugar, la fuente de datos utilizada, así como la parte de esta que ha sido empleada para el estudio, y, en segundo lugar, se detalla el fundamento teórico, para posteriormente definir el método y la justificación del modelo empleado.

3.1. FUENTE DE INFORMACIÓN: ENCUESTA NACIONAL DE SALUD DE ESPAÑA 2017

"La Encuesta Nacional de Salud de España (ENSE) es un conjunto seriado de encuestas que constituye la principal fuente de información sobre la salud percibida por la población residente en España" (Metodología ENSE, 2017, p. 5).

Los microdatos de la citada encuesta para el año 2017 constituyen la principal fuente de información a partir de la cual se elabora el modelo propuesto. Por tanto, se trata de información secundaria y los datos son de corte transversal.

Esta encuesta se elabora en España con periodicidad quinquenal, alternándose cada dos años y medio con la Encuesta Europea de Salud en España. Dado que ambas comparten un importante número de variables y han sido armonizadas entre sí, se dispone de datos ciertamente periódicos que permiten observar el comportamiento temporal del nivel de salud percibido. Esta periodicidad es la que ha permitido que, en el capítulo anterior, se haya abordado un breve análisis temporal sobre el consumo de tabaco.

La encuesta cuenta con tres tipos de cuestionarios: hogares, adultos y menores. Este trabajo se ha centrado en los microdatos del cuestionario para adultos (15 y más años) pues, dado el tema que se pretende abordar, este colectivo constituye un conjunto de información mayor y que, *a priori*, podría enriquecer el resultado. Dentro de este colectivo, se cuenta con 23.089 encuestas y 455 variables.

Del total de variables disponibles, para la elaboración del modelo, se utilizan doce de ellas, aunque para los análisis descriptivos abordados en el siguiente capítulo se puede utilizar alguna adicional. Las variables seleccionadas se enmarcan en los apartados de la ENSE:

- Datos de identificación.
- Características demográficas de la persona adulta seleccionada.
- Estado de salud.
- Actividad física.
- Consumo de tabaco y exposición al humo de tabaco.
- Consumo de alcohol.

Dado que el modelo propuesto solo recoge dos decisiones alternativas (fumar o no fumar), se ha procedido a realizar una recodificación que haga que la respuesta a la pregunta ¿fuma actualmente?, recoja exclusivamente dos respuestas: sí o no, en lugar de las seis alternativas que contempla el cuestionario. En la siguiente tabla (tabla 1) se ofrece un esquema de la recodificación propuesta:

Tabla 1 Recodificación de la variable V121

Respuesta a variable V121: ¿Fuma actualmente?			
Valores iniciales		Valores recodificados	
1	Sí, fumo a diario	1	Sí fumo
2	Sí fumo, pero no a diario	0	No fumo
3	No fumo actualmente, pero he fumado antes	0	No fumo
4	No fumo ni he fumado antes de manera habitual	0	No fumo
8	No sabe	0	No sabe
9	No contesta	0	No contesta

Con la recodificación propuesta, los individuos objeto de estudio quedan agrupados en dos colectivos: fumadores y no fumadores. Por otra parte, se procede a excluir a los individuos cuya respuesta no ha ido en ninguno de los sentidos propuestos, esto es, aquellos que han dado como respuesta no sabe o no contesta. Se trata de 22 individuos, un 0,09% del total de la muestra. En esta línea, también se procede a eliminar a aquellos individuos que han respondido de igual forma (no sabe o no contesta) a algunas de las variables que incluirá el modelo, dado que no aportarán información en ninguno de los sentidos propuestos. Se trata de las variables V126, T112, G22 y W127[8]. El total de eliminaciones realizadas es de 56 individuos, un 0,24%.

Como resultado de las modificaciones y eliminaciones practicadas, el trabajo se realiza con las respuestas de 23.033 individuos.

Además, con la intención de simplificar el amplio abanico de posibilidades que presentan otras variables, como por ejemplo la actividad económica actual desempeñada por el individuo o su nivel de estudios, se procede a realizar también las siguientes recodificaciones (tablas 2 a 5):

Tabla 2 Recodificación de la variable ACTIVa

Respuesta a variable ACTIVa: Actividad Económica Actual			
Valores iniciales		Valores recodificados	
1	Trabajando	1	Activo/a
2	En desempleo	1	Activo/a
3	Jubilado/a, prejubilado/a	0	Inactivo/a
4	Estudiando	0	Inactivo/a
5	Incapacitado/a para trabajar	0	Inactivo/a
6	Las labores del hogar	0	Inactivo/a
7	Otros	0	Inactivo/a

[8] Denominación de la ENSE. Se omite el nombre y contenido de las variables en aras de expresarlas, en esta ocasión, de una forma más breve. Estos datos están recogidos en el ANEXO I de este trabajo.

Tabla 3 Recodificación de la variable NIVEST

Respuesta a variable NIVEST: Nivel de estudios			
Valores iniciales		Valores recodificados	
02	No sabe leer o escribir	1	Sin estudios
03	Educación Primaria incompleta (ha asistido menos de 5 años a la escuela)	1	Sin estudios
04	Educación Primaria completa	2	Estudios Primarios
05	Primera etapa de Enseñanza Secundaria, con o sin título (2º ESO aprobado, EGB, Bachillerato Elemental)	3	Estudios Secundarios, Medios y Bachillerato
06	Estudios de Bachillerato	3	Estudios Secundarios, Medios y Bachillerato
07	Enseñanzas profesionales de grado medio o equivalentes	3	Estudios Secundarios, Medios y Bachillerato
08	Enseñanzas profesionales de grado superior o equivalentes	4	Estudios superiores
09	Estudios universitarios o equivalentes	4	Estudios superiores

Tabla 4 Recodificación de la variable T112

Respuesta a variable T112: Frecuencia con la que realiza alguna actividad física en su tiempo libre			
Valores iniciales		Valores recodificados	
1	No hago ejercicio. El tiempo libre lo ocupo de forma casi completamente sedentaria	0	No, no realizo, de forma regular, actividad física en mi tiempo libre
2	Hago alguna actividad física o deportiva ocasional	0	No, no realizo, de forma regular, actividad física en mi tiempo libre
3	Hago actividad física varias veces al mes	1	Sí realizo, de forma regular, actividad física en mi tiempo libre
4	Hago entrenamiento deportivo o físico varias veces a la semana	1	Sí realizo, de forma regular, actividad física en mi tiempo libre
8	No sabe	8	No sabe
9	No contesta	9	No contesta

Tabla 5 Recodificación de la variable W127

Respuesta a variable W127: Frecuencia de consumo de alcohol en los últimos 12 meses			
Valores iniciales		Valores recodificados	
01	A diario o casi a diario	1	Consumo habitual
02	5-6 días por semana	1	Consumo habitual
03	3-4 días por semana	1	Consumo habitual
04	1-2 días por semana	1	Consumo habitual
05	2-3 días en un mes	0	Consumo esporádico o no consumo
06	Una vez al mes	0	Consumo esporádico o no consumo
07	Menos de una vez al mes	0	Consumo esporádico o no consumo
08	No en los últimos 12 meses, he dejado de tomar alcohol	0	Consumo esporádico o no consumo
09	Nunca o solamente unos sorbos para probarlo a lo largo de toda la vida	0	Consumo esporádico o no consumo
98	No Sabe	98	No Sabe
99	No contesta	99	No contesta

Estas recodificaciones se realizan atendiendo a la similitud en la consideración de las categorías agrupadas, a pesar de que tengan ciertas diferencias entre sí, y al número de individuos al que afectan. En consecuencia, a excepción de la tercera, todas las recodificaciones tienen como resultado una única variable cualitativa o *dummy* (Stock y Watson, 2012). En el caso de la tercera, así como con otras que no han sido recodificadas, se articulan variables cualitativas por categorías. Este asunto queda reflejado en el anexo I, donde se relacionan las variables finales que contiene el modelo, su correspondencia con las variables de la ENSE y los valores que estas toman.

Asimismo, en relación con la variable V126 que mide la frecuencia de exposición a ambientes de humo en lugares cerrados, se procede a transformar dicha variable en una variable numérica y continua, utilizando, para asignar dicho número, el promedio de los límites por categoría propuestos por la encuesta. Este cambio responde a la intención de aportar variables de tipo no cualitativo al modelo, dado que, en su mayoría, son estas las que se han definido. Esta transformación resulta de la siguiente manera (tabla 6):

Tabla 6 Transformación de la variable V126

Respuesta a variable V126: Frecuencia de exposición a ambientes de humo en lugares cerrados		
Valores iniciales		Valores transformados (en minutos)
1	Nunca o casi nunca	0
2	Menos de una hora al día	30
3	Entre 1 y 5 horas al día	180
4	Más de 5 horas al día	300

Con este cambio se adapta la encuesta a los datos que serán utilizados y esta variable queda definida de tal forma que contempla con mayor precisión el tiempo que un individuo nuevo a incorporar en la muestra, o a estimar, está expuesto a humo en lugares cerrados.

3.2. FUNDAMENTO TEÓRICO

Vale la pena comenzar este apartado indicando la distinción entre los modelos agregados de demanda (de primera de generación) y los modelos desagregados (de segunda generación). Esto permitirá enlazar, de una forma más sencilla, el propósito de este trabajo con lo que será el marco de referencia que sustente, en materia de análisis económico y del comportamiento del consumidor, el desarrollo que se propondrá a continuación.

Los modelos agregados pretenden explicar las decisiones de los consumidores a partir de las observaciones efectuadas a grupos o mediante la generalización del comportamiento promedio. En cambio, los modelos desagregados se basan en las elecciones que cada individuo efectúa, consiguiendo entonces un mayor acercamiento a la realidad del consumidor. Esta razón es la que hace que, estos últimos, sean considerados más adecuados para proponer modelos más realistas. Sin embargo, a pesar de que los trabajos de Warner (1962) y Oi y Shuldimer (1962) abordaban las carencias de los modelos convencionales, esto es, los de primera generación, no provocaron que estos cayeran en desuso hasta principios de los ochenta. Los modelos de segunda generación son, por tanto, una herramienta adecuada para modelizar el comportamiento del individuo y, más particularmente, para explicar la postura adoptada frente a la decisión de fumar o no fumar. La

utilización de los microdatos que realiza este trabajo va en esta línea, en la de adecuar el modelo a lo que individualmente ha decidido cada sujeto y, lo que resulta más importante, cuáles han sido las circunstancias que lo han llevado a tomar esa decisión y no otra.

La respuesta a la pregunta anterior está en la utilidad que reporta la alternativa seleccionada al individuo. Ahora bien, la decisión adoptada no posee utilidad para el individuo en sí misma, sino que se deriva de las características propias de este, así como de los atributos que presenta cada una de las posibles alternativas. Por tanto, la utilidad es un concepto subjetivo e indudablemente particular para cada decisor. Esto, como es normal, no se diferencia del concepto económico de utilidad que está comúnmente aceptado y que permite asignar útiles a cada alternativa, consiguiendo entonces tomar una decisión, tras comparar la magnitud de cada una de ellas.

En este sentido, la Teoría de la Utilidad Aleatoria (Domencich y McFadden, 1975) es la que, habitualmente, sirve como base teórica para la propuesta de modelos de elección discreta que estudien las decisiones del consumidor, además de que se fundamenta en aquello que se considera crucial a la hora de decidir: la utilidad percibida por el individuo. Esta teoría defiende lo siguiente:

I. Los individuos actúan racionalmente y cuentan con toda la información necesaria para la toma de decisiones. Por tanto, pretenderán maximizar su utilidad, de acuerdo con sus restricciones.

II. Se cuenta con un conjunto $A = \{A_1, A_2, \dots, A_j, \dots, A_N\}$ de alternativas susceptibles de ser elegidas y un conjunto X de atributos de las alternativas y de características de los individuos.

III. Cada alternativa A_j genera en el individuo (i) cierta utilidad U_{ij}. Como ya se señaló, la utilidad es un concepto sumamente subjetivo y propio de cada individuo, por tanto, es muy complejo poseer información completa sobre lo que determina su valor. En consecuencia, el investigador reconoce que la utilidad no dependerá solo de los atributos y de las características observadas, sino que existirá un componente aleatorio que recoja la omisión de aspectos no medibles como los gustos o las preferencias, además de poner de relieve los posibles errores que, durante el proceso de observación, cometa el investigador. De esta manera, la utilidad se compone de una parte sistemática (V_{ij}) y de otra aleatoria (ε_{ij}), tal que:

$$U_{ij} = V_{ij}\,(X) + \varepsilon_{ij} \qquad (3.1)$$

Cabe destacar que, de no reconocer la existencia de un componente aleatorio, el modelo podría llevar a resultados no consistentes en términos de racionalidad, pues dadas las características observadas, la decisión adoptada por el individuo debería ser distinta a la efectivamente adoptada. Esa diferencia radica en el término estocástico.

IV. Una vez resuelto el problema de maximización, el individuo i elegirá la alternativa $A_{j^*} \in A$ si y solo si

$$U_{ij^*} \geq U_{ij}, \forall j \neq j^* \tag{3.2}$$

es decir

$$V_{ij^*} + \varepsilon_{ij^*} \geq V_{ij} + \varepsilon_{ij} \rightarrow V_{ij^*} - V_{ij} \geq \varepsilon_{ij} - \varepsilon_{ij^*} \tag{3.3}$$

Dado que el investigador desconoce el valor de $\varepsilon_{ij} - \varepsilon_{ij^*}$, no puede comprobar si la expresión anterior se verifica, por lo que

$$\begin{aligned} P_{ij} &= P\left(U_{ij^*} \geq U_{ij}\right) \\ &= P\left(V_{ij^*} + \varepsilon_{ij^*} \geq V_{ij} + \varepsilon_{ij}\right) \\ &= P\left(\varepsilon_{ij^*} - \varepsilon_{ij} \geq V_{ij} - V_{ij^*}\right) \\ &= P(\varepsilon_{ij} - \varepsilon_{ij^*} \leq V_{ij^*} - V_{ij}) \\ &\boxed{= P(\varepsilon_{ij} \leq \varepsilon_{ij^*} + V_{ij^*} - V_{ij})} \end{aligned} \tag{3.4}$$

A partir de lo expuesto en este epígrafe, se debe formular el modelo de elección discreta que permita, de acuerdo con los principios relacionados anteriormente, estimar la decisión del individuo. En el siguiente epígrafe se aborda la descripción de estos modelos, su justificación y los diferentes modelos que pueden ser aplicados.

3.3. MODELOS DE VARIABLE DEPENDIENTE DISCRETA

El objetivo de este trabajo es definir, a partir de ciertas variables explicativas, la decisión que toma un individuo ante el hecho de fumar o no fumar. Más particularmente, se trata de conocer qué probabilidad o qué nivel de propensión muestra este a partir de sus características socioeconómicas.

En este sentido, se trata, por tanto, de estudiar la menor o mayor relación que, en términos estadísticos, existe entre una decisión y ciertos valores cualitativos y cuantitativos que toman las variables exógenas del modelo que se pretende proponer.

Como ya se señaló, se plantea un escenario en el que el individuo objeto de estudio tenga dos alternativas de respuesta: fumar o no fumar. A este respecto, resulta necesario que el modelo econométrico a proponer contemple esta circunstancia, esto es, tenga carácter discreto. Particularmente, que tome valores 0 y 1 (1 si la decisión es fumar y 0 si la decisión adoptada es no fumar), y así ha sido recodificada, a partir de los datos de la encuesta, la variable dependiente, tal y como consta en los epígrafes anteriores.

Ahora bien, este análisis no puede plantearse en los términos habituales de una regresión con variable dependiente continua, cuando, por ejemplo, se estudia el coste de producción de cierto producto a partir de los niveles de input que precisa, sino que se debe articular alrededor de un modelo que contemple esta peculiaridad, es decir, que se enfrente a una variable dependiente limitada, y en este caso binaria. Algunos ejemplos que gozan de la misma naturaleza son la participación o no en el mercado laboral por parte de un individuo, la decisión de votar a un partido político o los determinantes que explican que un país reciba ayuda exterior.

Los modelos de elección discreta o de respuesta cualitativa son los que posibilitan este estudio con las características propias que han sido mencionadas. En este sentido, se presentan los principales modelos de probabilidad que contemplan estas circunstancias: modelo de probabilidad lineal, modelo probit y modelo logit. Además, aunque no se abordarán en este trabajo, existen varias extensiones de estos modelos como el probit ordenado y los modelos Tobit (Tobin, 1958), también modelos de datos de enumeración y aplicaciones a conjuntos de datos cuya variable dependiente ya no es binaria, sino policotómica.

3.3.1. Modelo de Probabilidad Lineal

Se pretende transformar $\mathrm{X}\beta$ en una probabilidad. Es decir, se precisa de una función F tal que:

$$prob\ (y_i = 1) = F(X_i\beta) \tag{3.5}$$

Una elección natural para la función F, que transforma $\mathrm{X}\beta$ en un número comprendido entre 0 y 1, es una función de distribución, o la densidad acumulada. De hecho, podemos definir así los modelos de respuesta binaria.

Diciendo que F es una función de identidad tal que

$$prob\ (y_i = 1) = (X_i\beta) \tag{3.6}$$

obtenemos el modelo de probabilidad lineal. (Johnston y Dinardo, 2001, p. 478)

Sin embargo, el modelo de probabilidad lineal está casi en desuso en la actualidad. Su uso se reduce a contadas excepciones, siendo su sencillez la ventaja que explica su utilización. Esto está explicado por una de sus mayores debilidades: el modelo no es capaz de delimitar el valor de la probabilidad predicha entre cero y uno. Además, este es heterocedástico, esto es, "la varianza de la perturbación no es constante a lo largo de las observaciones" (Greene, 1999, p. 469), sino que depende de $\mathrm{X}\beta$, aspecto que se deriva de la estructura que sigue su residuo, que solo toma uno de estos valores: $1 - \mathrm{X}\beta$ o $-\mathrm{X}\beta$.

3.3.2. Modelo Probit[9]

El modelo probit parte de la utilización de la distribución normal estándar para modelizar la parte estocástica del modelo. Esta se suele representar como $\emptyset\ (.)$. En este sentido

$$P(Y = 1) = \int_{-\infty}^{X\beta} \Phi\ (t)dt = \emptyset(X\beta) \tag{3.7}$$

Si se expresa de forma desarrollada, suponiendo que depende de k regresores, resulta

$$P\left(Y = 1/_{X_1, X_2, \dots, X_k}\right) = \emptyset(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \tag{3.8}$$

[9] Elaborado a partir de Stock y Watson (2012) y Greene (1999).
También es importante que se señale que, tanto para el modelo probit como para el logit, las probabilidades que se calculan son condicionadas a los valores que toman, para cada individuo, las distintas variables.

3.3.3. Modelo Logit[10]

El modelo de regresión logit, por el contrario, utiliza la distribución logística. Esta se representa, habitualmente, así: $\Lambda(.)$. De esta forma

$$P(Y=1)=\frac{e^{X\beta}}{1+e^{X\beta}}=\Lambda(X\beta) \tag{3.9}$$

Si se expresa de forma desarrollada, suponiendo que depende de k regresores, resulta

$$P\left(Y=1/_{X_1,X_2,\dots,X_k}\right)=\Lambda(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)$$
$$=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}}\,^{11} \tag{3.10}$$

Para cerrar este capítulo metodológico, a continuación, se abordan algunas similitudes y diferencias entre los dos últimos modelos probabilísticos mencionados (probit y logit) a partir de Greene (1999).

- Las dos distribuciones tienden a dar probabilidades muy similares a los valores intermedios de $\beta'X$. En cambio, cuando $\beta'X$ es muy pequeño, la logística tiende a dar probabilidades mayores para $y=0$. Ocurre al revés para $\beta'X$ muy grande.
- Se espera, por el contrario, que ambos modelos no originen predicciones iguales cuando la muestra contiene pocas respuestas afirmativas o pocas respuestas negativas, además de que tenga gran variación en una variable independiente de importancia.
- Desde un punto de vista teórico, resulta difícil justificar la no elección de un modelo en favor del otro. En la mayoría de las ocasiones, coinciden los resultados obtenidos con ambas distribuciones.

[10] Elaborado a partir de Stock y Watson (2012) y Greene (1999).

[11] Esta expresión (3.10) resulta de dividir la descrita inicialmente (3.9) entre $e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}$, de tal manera que:

$$P\left(Y=1/_{X_1,X_2,\dots,X_k}\right)=\frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}/_{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}}}{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}/_{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}}+1/_{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}}}$$
$$=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}}$$

4. ANÁLISIS DESCRIPTIVO DE LOS DATOS[12]

El capítulo que se presenta a continuación pretende, por un lado, contextualizar en términos descriptivos la muestra obtenida por la ENSE 2017 (distribuciones por sexo, Comunidades Autónomas, etc.), y, por otra parte, analizar combinaciones de variables que, a su vez, permitan – en cierto modo – establecer resultados esperados y correlaciones.

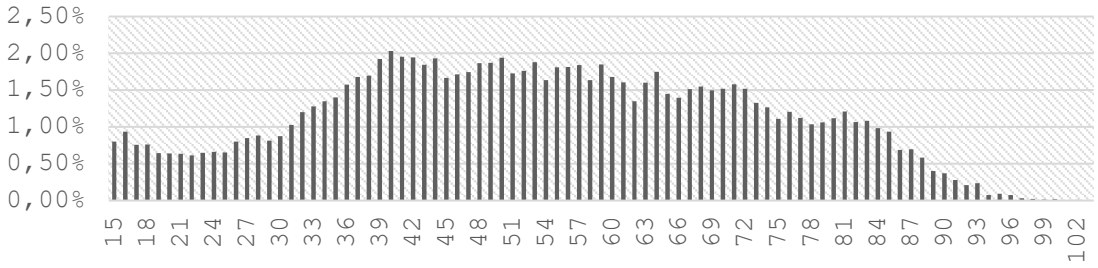
4.1. DESCRIPCIÓN DE LA MUESTRA[13]

La muestra total de la Encuesta Nacional de Salud de España, dentro de las encuestas realizadas a adultos, cuenta con 23.089 individuos, pero, tras las eliminaciones practicadas en el capítulo 3, la muestra objeto de estudio resulta de 23.033 casos.

Por sexos, la muestra tiene 10.560 hombres (45,85%) frente a 12.473 mujeres (54,15%). Por tanto, en estos términos, la distribución está bastante equilibrada y, como es lógico, ajustada a la proporción que, en la sociedad, tienen ambos sexos.

La edad, por su parte, abarca un amplio abanico de datos. La edad mínima disponible es de 15 años (edad mínima de los encuestados de esta muestra) y la edad máxima es una observación de un individuo de 103 años. A continuación, se muestra la distribución relativa, sobre el total de encuestados, de cada una de las edades (gráfico 2).

Gráfico 2 Distribución de la muestra por edades (% sobre el total)



Fuente: Elaboración propia a partir de los datos de la Encuesta Nacional de Salud de España (2017)

A pesar de que, en ciertos intervalos, no se proporcione una cantidad representativa de individuos, se opta por no agrupar las edades para observar gráficamente la forma de la distribución y así detectar más fácilmente, si los hubiera, rangos de edad más representados o, sencillamente, observar donde se concentra mayor número de encuestados. Para las edades centrales, de los 35 a los 60 años, se observa un importante número de observaciones que,

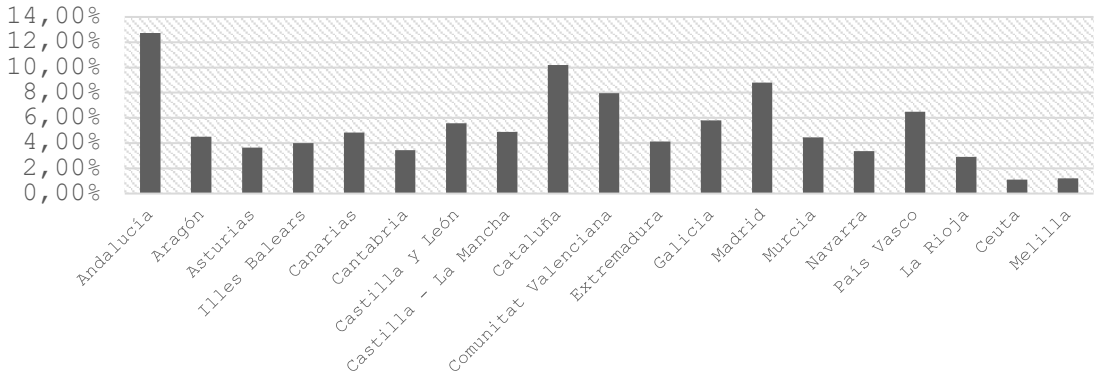
[12] Se hace notar que los cálculos expuestos, dentro de este capítulo, se realizan sin hacer uso de los factores de elevación propuestos por la fuente de datos. En cualquier caso, se ha comprobado, para la proporción de individuos fumadores, que no existen importantes diferencias entre la media muestral y la media muestral ponderada al factor. Además, tras el proceso de eliminación de individuos acometido en el capítulo 3, para ceñir perfectamente los resultados a la realidad, sería necesario recalcular los factores de elevación. Aun así, admitiendo cierto error, dado que la pérdida de individuos es muy pequeña, se proponen, más adelante, estimaciones de los modelos haciendo uso del factor, y, nuevamente, los resultados son muy similares.

[13] Todos los datos contenidos en este epígrafe (4.1.) se muestran agregados en el resto de las variables, salvo por aquella por la que se clasifican. Por ejemplo, en el caso de la distribución por edades, se muestran conjuntamente ambos sexos y no divididos por estos.

posteriormente, va descendiendo cuando se aproxima a ambas colas, donde se ubican los valores extremos. Como ya se había mencionado, los individuos de 90 o más años, son un rango de edad que cuenta con pocos individuos. Este es un aspecto esperado a partir de la radiografía demográfica de España y en consonancia con los métodos de muestreo utilizados por el Instituto Nacional de Estadística.

Por último, en este primer nivel de descripción de la muestra, se abordará la distribución de esta por Comunidades Autónomas. En este caso, la información metodológica de la ENSE señala que "la muestra se distribuye entre comunidades autónomas asignando una parte uniformemente y otra de forma proporcional al tamaño de la comunidad" (Metodología ENSE, 2017, p. 11). De esta forma, la muestra presenta la siguiente distribución de pesos relativos por Comunidades Autónomas:

Gráfico 3 Distribución de la muestra por CC.AA. (% sobre el total)



Fuente: Elaboración propia a partir de los datos de la Encuesta Nacional de Salud de España (2017)

Como ya se señaló, dado que responde a una lógica de reparto en función de la población de la Comunidad Autónoma, aquellas que tienen más población cuentan con mayor número de individuos, es el caso de Andalucía (12,73% del total), y las que cuentan con un menor número de habitantes están menos representadas, como Ceuta y Melilla, con un 1,11% de la muestra en ambos casos. Canarias, por su parte, es la novena comunidad, cuenta con 795 encuestas, un 3,45% del total.

Por otra parte, antes de abordar otras variables en el epígrafe siguiente, se presentan algunos resultados que están relacionados con otras variables de la ENSE que, a pesar de no ser utilizadas en la especificación del modelo, resulta interesante su análisis para conocer mejor el perfil de fumador español.

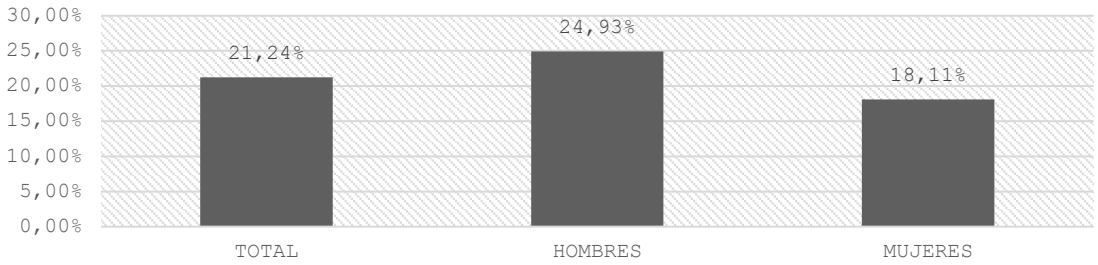
Por ejemplo, si se analiza el número de cigarrillos que el individuo consume al día, resulta un valor medio de 13 unidades por individuo, siendo además el cigarrillo convencional el más usual. Ahora bien, también algunos de los individuos declarados como fumadores, en torno al 24% del total, declaran haber intentado dejar de fumar en los últimos doce meses, al menos, en una ocasión. Se alcanzan valores de hasta 5 intentos, pero, lo más frecuente, es que no lo hayan intentado.

4.2. RELACIONES ENTRE VARIABLES: DECISIÓN ADOPTADA Y COMBINACIONES DE FACTORES EXPLICATIVOS

Se pretende ahora, a través de un análisis descriptivo, detectar posibles relaciones de los colectivos con la elección; en definitiva, se trata de analizar las posibles correlaciones existentes, en términos de una mayor propensión a fumar del individuo, de acuerdo con sus características.

Del conjunto de la muestra, un total de 4.892 individuos se declaran fumadores, esto representa un 21,24% de los encuestados. Si se toma este valor como representativo del conjunto y se compara con los datos que arroja el análisis por sexos (gráfico 4), se observa como las mujeres presentan un menor porcentaje de fumadoras (18,11%), mientras que los hombres superan ese valor representativo del total en más de tres puntos porcentuales, situándose en el 24,93% de los encuestados.

Gráfico 4 Fumadores totales según sexos



Fuente: Elaboración propia a partir de los datos de la Encuesta Nacional de Salud de España (2017)

En consecuencia, los hombres, cabe esperar, que presenten una mayor propensión a fumar que las mujeres, pues no solo superan el porcentaje de fumadores de la muestra, sino que, respecto a ellas, sobrepasan su porcentaje en más de seis puntos. Este asunto es previsible que también se traduzca en una mayor probabilidad de elección de fumar. Si se continúa analizando este asunto por sexos, pero combinándolo con rangos de edad, se sigue observando como los hombres, aunque levemente, siguen siendo más. Es decir, no se trata de un efecto exclusivamente agregado, sino que parece ser un patrón que se repite a lo largo de la muestra.

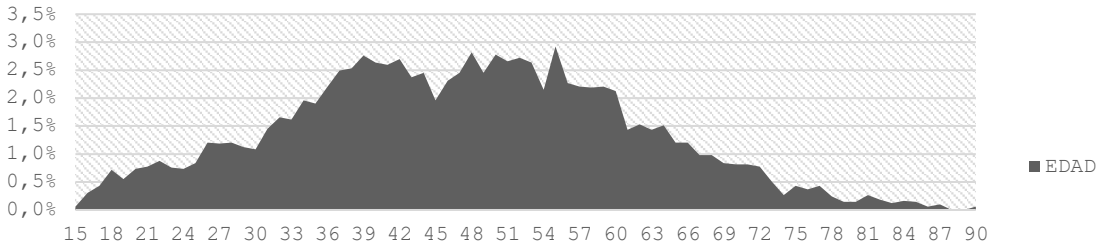
Siguiendo con el factor edad, la edad media del individuo fumador es de 47 años[14]. Quizás, sea la edad, la variable que arroja un resultado más claro (el reparto por sexos y CC.AA. es más homogéneo), en tanto que se observa una evidente concentración de individuos fumadores a partir de los 40 años, donde se alcanzan los primeros picos, y hasta los 60 años, donde ya se define un claro declive. Alcanzar la madurez, con lo que ello implica en términos de una necesidad mayor de cuidar la salud, podría ser lo que explique este comportamiento por edades. Además, eso no varía según sexos, por ello, se muestra en el gráfico 5, una relación entre el porcentaje, sobre el total de fumadores, de los sujetos que fuman y la edad de estos, con independencia de su sexo.

[14] Nótese que la edad media de la muestra se sitúa en los 53 años. En este sentido, cabría pensar que, en la misma línea de lo ya expresado, la concentración de individuos fumadores se da en edades más tempranas, de ahí que la edad media de estos sea menor.

Abordando el último de los factores descriptivos estudiados en el epígrafe anterior, la Comunidad Autónoma de residencia del encuestado, cabe decir que, en general, todas presentan un nivel similar de fumadores respecto al total de individuos de esa zona geográfica.

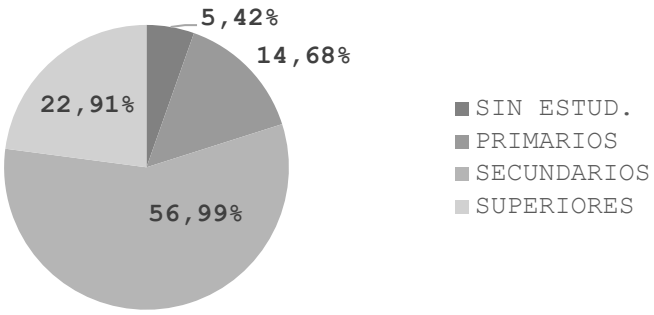
En la línea de lo anterior, ahora se analiza la relación existente entre el nivel educativo y la decisión de fumar. Inicialmente se estudia el peso que los fumadores, según los niveles de estudios considerados, tienen sobre la muestra total (gráfico 6). Aquellos individuos que declararon tener estudios secundarios, esto es, de acuerdo con la recodificación propuesta, estudios secundarios, bachillerato y formación profesional de grado medio, son los que concentran mayor número de fumadores en la muestra (57% del total). Por el contrario, los individuos categorizados como sin estudios representan el 5,42% de los fumadores del conjunto de la muestra. Si bien es cierto que, en ambos casos, estos dos grupos son el más y el menos numeroso de la muestra, respectivamente.

Gráfico 5 Fumadores totales según edades



Fuente: Elaboración propia a partir de los datos de la Encuesta Nacional de Salud de España (2017)

Gráfico 6 Distribución de la muestra por niveles de estudios



Fuente: Elaboración propia a partir de los datos de la Encuesta Nacional de Salud de España (2017)

Ahora bien, si se realiza este mismo análisis desde una perspectiva intragrupo, es decir, el porcentaje de individuos fumadores que incluye cada uno de los niveles de estudios (tabla 7), estos valores se reducen considerablemente, hasta el punto de que los fumadores con estudios secundarios son los únicos que superan el nivel de fumadores total de la muestra. Como ya se mencionó, el peso que este colectivo tiene en la muestra (43%) es lo que provoca, en cierta medida, el ascenso del nivel relativo de fumadores en la muestra.

Tabla 7 Porcentaje de fumadores según niveles de estudios

Porcentaje de fumadores según nivel de estudios (sobre el total de sujetos de cada nivel de estudios)	Sin estudios	9,68%
	Estudios primarios	16,15%
	Estudios secundarios, Medios y Bachillerato	28,11%
	Estudios superiores	18,90%

Fuente: Elaboración propia a partir de los datos de la ENSE (2017)

A continuación, se pretende analizar la relación entre la decisión de fumar y otras características propias del individuo como es el caso de variables que recogen el estado de salud de este, así como sus hábitos de vida, por ejemplo, el consumo periódico de alcohol, la realización habitual de actividad física y la periodicidad con la que se expone a espacios cerrados con humo, entre otras.

La variable ACTIVa[15] recoge si, en el momento que realizó la encuesta, el individuo era considerado como activo (trabajaba o estaba desempleado) o inactivo (jubilado, estudiante, etc.). Al analizar la postura de un sujeto, de acuerdo con esta característica (tabla 8), ante la decisión de fumar, el 28,93% de los activos encuestados declaran ser fumadores. Por el contrario, solo el 12,31% de los encuestados inactivos consumen tabaco de forma habitual. A este respecto, parece entonces que hay cierta mayor propensión a fumar por parte de los activos.

Como se señaló al inicio de este trabajo, los individuos objeto de estudio han sido sujetos residentes en España con 15 años o más. Ahora bien, estos sujetos podrían haber nacido en España o no y, a nivel de país, resulta interesante reconocer si se da un posible comportamiento característico de una condición o la otra, porque, aún sin precisar la nacionalidad, resultaría llamativo observar, por ejemplo, mayor propensión en los nacidos en España. Esto iría, además, en consonancia con lo expuesto más arriba[16], donde se señaló que en comparación con la UE 28, los españoles fuman más que el resto. Sin embargo, a la luz de los datos, este comportamiento diferencial podría no observarse, pues en el caso de los nacidos en España, el porcentaje de fumadores es del 21,46% (tabla 8), mientras que para los nacidos en el extranjero es del 19,22%; no obstante, a pesar de que la diferencia es escasa, la diferencia entre ambos colectivos es estadísticamente relevante, tal y como se verá en el epígrafe siguiente.

Dentro de este apartado de variables propias, se procede a analizar aquella que mide el tiempo que pasa un individuo en ambientes cerrados expuestos al humo del tabaco. En el conjunto de la muestra, una importante mayoría no frecuenta este tipo de lugares, es decir, pasa cero minutos en entornos de este tipo. Por supuesto, se debe recordar que la muestra es de 2017 y, por tanto, tras la aprobación de la "nueva ley antitabaco" (2010) que restringía el consumo de tabaco solo a exteriores. Por ello, de producirse esta situación sería, en principio, en zonas privadas, como los hogares. De hecho, como es de esperar, esta cifra disminuye considerablemente en el caso de los fumadores, pues ya solo el 69,56% de estos no frecuenta estos espacios (tabla 8) y, en consecuencia, la exposición a estos ambientes, además de que pueda que sea el individuo en

[15] Denominación de la ENSE 2017. Ver anexo I.
[16] Dado que el 90,29% de la muestra es de individuos nacidos en España, de detectarse un comportamiento diferenciado según procedencia, este resultado podría establecerse como distintivo del país, de ahí que se relacione con la comparación con la UE 28.

cuestión quien genera el humo, podría ser un aliciente a la hora de comenzar a fumar, si se diera el caso de que la exposición no tuviera como responsable al sujeto en cuestión. Para el resto de los tiempos contemplados[17], el peso de individuos es relativamente similar y decreciente con el aumento del tiempo de exposición.

Tabla 8 Frecuencias relativas conjuntas, según características propias, entre regresores y variable dependiente (fumar – no fumar)

		FUMA	
		NO	SÍ
Actividad económica	Inactivo/a	87,69%	12,31%
	Activo/a	71,07%	28,93%
País de nacimiento	Extranjero	80,78%	19,22%
	España	78,54%	21,46%
Los porcentajes anteriores se realizan sobre el total de cada colectivo			
Tiempo de exposición al humo en lugares cerrados *(porcentaje respecto del total de fumadores y no fumadores)*	0 minutos	93,57%	69,56%
	(0, 300]	6,43%	30,44%

Fuente: Elaboración propia a partir de los datos de la ENSE (2017)

Dentro del segundo bloque propuesto, relacionado con el individuo, pero más particularmente con aquellas variables que inciden directamente sobre su estado de salud, así como sobre sus hábitos, se proponen los siguientes análisis.

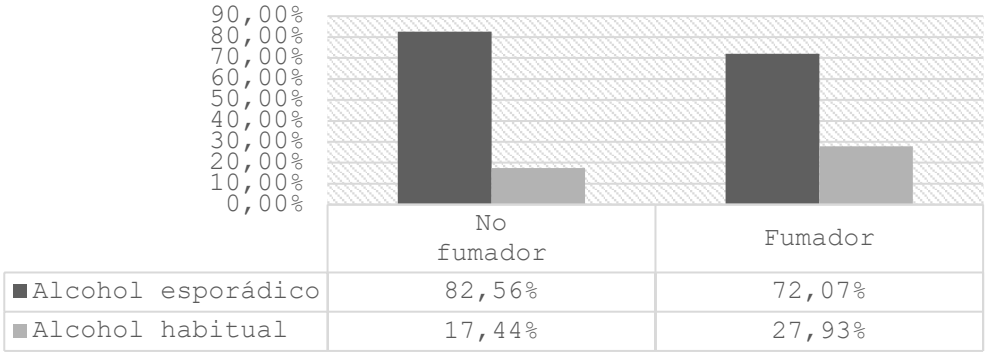
En primer lugar, se aborda la realización de actividad física por parte de cada uno de los individuos de la muestra. Inicialmente, se quisiera hacer hincapié en el dato muestral que revela que el 22,38% de los individuos realiza, de forma regular, actividad física en su tiempo libre. Por su parte, de entre los que realizan actividad física con cierta frecuencia, casi el 18% fuma a diario, frente a una clara mayoría, superior al 80%, que no lo hace. En cambio, la situación contraria (no realizar actividad física con regularidad) arroja unos resultados inesperados y que merece la pena analizar con detenimiento, pues la diferencia entre la tasa de fumadores que realizan actividad física y la de fumadores que no la realizan es de algo más de cinco puntos porcentuales, diferencia que se preveía mayor, ya no tanto por una mayor proporción de fumadores no deportistas[18], sino por un número menor de deportistas fumadores.

La frecuencia con la que se consume alcohol es la siguiente variable que va a ser analizada y que goza de gran relevancia. Incuestionablemente, el alto consumo de alcohol es, al igual que fumar, un hábito que comporta un gran peligro para la salud, en consecuencia, la unión de ambos hábitos constituye una combinación que, cuanto menos, podría relacionarse con el empeoramiento del estado de salud en el que se encuentra el individuo. Un dato que resaltar es que ocho de cada diez encuestados de los que consumen alcohol de forma esporádica, no fuma a diario. Otro dato que destacar es la importante diferencia de casi diez puntos entre los fumadores de un grupo y

[17] Para presentar los resultados en la tabla 8 de una forma más sencilla, el resto de los tiempos considerados se muestra en un único intervalo que permita distinguir a aquellos sujetos que no pasan tiempo en estos espacios y los que sí lo pasan, con independencia del tiempo exacto que pasen.
[18] Se emplea el término deportista para hacer alusión a los individuos que realizan actividad física con regularidad, aunque, estrictamente, no sea posible determinar tal consideración.

otro. Dentro del colectivo de individuos que no consumen alcohol habitualmente, el 17,44% fuman, mientras que, con una diferencia de más de un diez por ciento, el 27,93% de los que sí lo hacen, fuman. Por tanto, parece que, al menos de momento, se podría esperar una clara relación entre ambos hábitos, también en términos probabilísticos. Todos los datos relativos a este asunto se presentan en el gráfico 7.

Gráfico 7 Fumar y consumo de alcohol



	No fumador	Fumador
Alcohol esporádico	82,56%	72,07%
Alcohol habitual	17,44%	27,93%

Fuente: Elaboración propia a partir de los datos de la Encuesta Nacional de Salud de España (2017)

En último lugar, dentro de esta línea relacionada con la salud, se encuentra que el 69% de la muestra padece un problema de salud crónico o de larga duración[19]. Dentro de cada uno de los grupos (fumadores y no fumadores) se observa una importante cantidad de individuos con enfermedades de esta naturaleza.

Sin embargo, un asunto que podría tener una doble lectura es la proporción dentro de cada colectivo, dado que, en el caso de los fumadores, se trata de un 61,84%, mientras que, para los no fumadores, este porcentaje se eleva hasta el 71,27%. A este respecto, se podría esperar que en el colectivo que padece estas dolencias, podría haber un mayor número de personas fumadoras, incluso que sea precisamente el consumo de tabaco el que contribuya a la aparición de esta. Ahora, también podría suponerse que, justo por el padecimiento de esta enfermedad, el individuo deja de fumar y, por esta razón, la proporción es menor.

En el sentido de esta última hipótesis planteada, aunque para el análisis de este trabajo se contemple la situación de exfumador como una situación de no fumador, si analizamos esta opción sobre los microdatos originales, se observa como casi el 29% de la muestra que padece enfermedades crónicas o de larga duración son exfumadores, yendo esto en consonancia con lo recogido en el final del párrafo anterior.

[19] Se entiende por problema de salud de larga duración, según lo descrito en la ENSE 2017, a aquella enfermedad o problema de salud que dure, al menos, 6 meses. La pregunta precisa a la que responden los encuestados es: "*¿Tiene alguna enfermedad o problema de salud crónicos o de larga duración? (Entendemos por larga duración si el problema de salud o enfermedad ha durado o se espera que dure 6 meses o más)*".

5. MODELO DE ELECCIÓN DISCRETA: ¿FUMAR O NO FUMAR?

La estimación y la interpretación de los distintos modelos de elección discreta constituye el asunto que se abordará en este capítulo. Inicialmente se realizará la estimación del Modelo Lineal de Probabilidad, a pesar de los defectos ya expuestos, pues servirá para analizar la consistencia de los resultados (al menos en signo y significación), con independencia del modelo empleado. Posteriormente se estimará el Modelo Logit, que será en el que se centre la mayor parte del capítulo y, tras abordarlo y realizar las primeras estimaciones, se propondrán una serie de modificaciones para mejorar la calidad del modelo propuesto.

Antes de entrar en las estimaciones propiamente dichas, se exponen ciertos aspectos que son importantes y, además, comunes a todo lo expuesto más adelante:

- Los modelos se estiman, al menos inicialmente, con las 37 variables que se relacionan en el anexo I, siendo una de ellas (FUMA) la que se articula como variable dependiente. Los 36 restantes se proponen como regresores.

- Del conjunto de regresores, hay un total de 28 que, estando relacionadas entre sí, se especifican como variables cualitativas o *dummies*. Es el caso de las 19 variables de las CC.AA., de las 4 variables que recogen el nivel de estudios y de las 5 que engloban el estado de salud percibido por el individuo.

- A consecuencia de lo anterior, dado que cuando una de las variables de, por ejemplo, las Comunidades Autónomas, tome valor uno, el resto tomarán el valor cero, existe cierto riesgo. Este se materializa en la existencia de relaciones lineales exactas entre las variables independientes, desembocando esto en un problema de multicolinealidad[20]. Para evitar este problema, los modelos propuestos se estimarán con categorías de referencia[21] o término general, tal y como sigue:

    - Para las CC.AA.: Andalucía.
    - Para los niveles de estudios: Estudios Primarios.
    - Para los estados de salud percibidos: Muy bueno.

- También sobre multicolinealidad, se han analizado las posibles correlaciones que hubiera entre el conjunto de regresores. En el anexo II se muestra la tabla de correlaciones, a excepción de la edad, el sexo y las CC. AA.

  Las variables presentan relaciones entre sí, pero no lo suficientemente altas como para generar problemas de colinealidad. Este análisis también es útil para descartar que algunas de las variables midan, en realidad, el mismo atributo del individuo, pues en ese caso la correlación entre ellas hubiera sido más alta.

[20] Véase, por ejemplo, Greene (1998, p. 363).
[21] Otras variables como el sexo y la actividad económica actual también tienen término de referencia, que será aquel cuya categoría haya sido designada con el valor cero. La relación que se presenta obvia estas variables por no haber sido divididas en varias variables, sino en una única que permite recoger ambas categorías.

- Los modelos podrán ser estimados con o sin hacer uso del factor de elevación. En cada caso se especificará si se trata de una estimación que incluya este factor o no.
- Para la estimación de los modelos serán empleados el *software* SPSS y Gretl. A pesar de que los resultados sean esencialmente iguales, se hace uso de ambos porque, en cada caso, facilitan ciertos contrastes o estimaciones específicas que serán de utilidad.

5.1. ESTIMACIÓN DEL MODELO DE PROBABILIDAD LINEAL

Una vez estimado el modelo, haciendo uso del factor de elevación, este arroja los resultados que se recogen en el anexo III de este trabajo.

En términos de la significatividad de las variables especificadas, se puede señalar que, salvo algunas Comunidades Autónomas, todas presentan niveles de significación aceptables, al 1%, al 5% o al 10%. En el caso de las CC.AA., Aragón, Comunidad de Madrid, Comunidad Foral de Navarra, Ceuta y Melilla son las que presentan una significación superior al 10%, no encontrándose diferencias significativas con Andalucía (comunidad de referencia) y, por tanto, no juegan el papel explicativo que deben tener dentro del modelo. En los siguientes epígrafes se planteará una solución a este respecto.

Uno de los aspectos relevantes que presenta el modelo lineal es el signo que toman las estimaciones, pues se podrá contrastar este con el que resulte de la estimación del resto de modelos. Si se analiza de forma pormenorizada, se puede observar como la mayoría de los parámetros toman valores estimados superiores a cero. Solo cinco de estas presentan valores negativos: edad, sin estudios, estudios superiores, enfermedad de larga duración o crónica y actividad física. Este hecho implica que los incrementos que experimenten las variables se traducirán en disminuciones de la probabilidad de fumar.

Los parámetros de este tipo de modelos deben ser interpretados como efectos marginales, es decir, como cambios unitarios ante una variación de una unidad en la variable dependiente en cuestión. Ahora bien, estos cambios tienen dos lecturas dentro de este modelo, una para aquellos parámetros que no se articulan como variables con categorías, por ejemplo, la edad, y otra interpretación es la que se deriva de los efectos de las variables categóricas que sí que cuentan con término general.

Por ejemplo, el coeficiente que acompaña a la variable edad es -0,002, lo que indica que la probabilidad de fumar se reduce aproximadamente en esa cantidad, cuando la edad del individuo se incrementa en una unidad (manteniendo el resto de las variables constantes). En cambio, en el caso del estado de salud, como su variable de referencia es el estado muy bueno, si el individuo pasa de tener salud muy buena a regular, la probabilidad de que sea fumador se incrementa en 0,42. Análogamente, una tercera interpretación (aunque muy similar a la anterior) es la de, por ejemplo, la variable sexo. El término general en ese caso sería mujer, por tanto, la probabilidad aumentaría en 0,27 al pasar de la categoría mujer a hombre, pues la variable tomaría valor 1. En conclusión, en estos últimos casos, el efecto se interpreta como la variación que experimenta la probabilidad al pasar de la categoría de referencia a otra del mismo factor.

Terminando con este primer modelo, solo queda recordar que su principal utilidad es la de comprobar la consistencia de los resultados al ser comparado con los modelos que se estiman

más adelante. Principalmente, el hecho de no garantizar que las probabilidades estén limitadas entre cero y uno, además de que es heterocedástico, lo hacen no ser el idóneo.

5.2. ESTIMACIÓN DEL MODELO LOGIT

Antes de estimar los modelos, se pretende introducir algunos aspectos que resultan fundamentales a la hora de explicar los resultados obtenidos, además se argumentan algunas decisiones adoptadas durante la estimación de los modelos, así como el método que se emplea para la misma.

La no linealidad de los parámetros de modelos como el logit (también es el caso del probit), derivada de la no linealidad de la propia función, imposibilita que estos puedan ser estimados por el método de mínimos cuadrados ordinarios, como sí que se hace con los modelos de regresión convencionales. Por ello, la forma de estimación estándar que es utilizada para la estimación de estos es la máxima verosimilitud. Así, el estimador de máximo verosímil será aquel que esté compuesto por los valores de los coeficientes que hacen máxima la función de verosimilitud. Este mecanismo es el que está implantando en la mayoría de *software* que permiten su estimación.

Por otra parte, en la muestra empleada para este trabajo se da una circunstancia que es importante tener presente antes de estimar. Del conjunto de individuos encuestados, solo algo más del 20% ha respondido afirmativamente a la pregunta que da sentido a la variable dependiente del modelo: ¿fuma actualmente? Por tanto, se da un importante desequilibrio entre la cuantía de unos y ceros que toma esta variable a lo largo de la muestra. Esto lleva a que, de no modificarse el umbral de estimación del modelo (ubicado por defecto en el 0,5), esto es, de no variar el punto de corte que, en caso de sobrepasarlo, clasifica al individuo como fumador, provocaría que el modelo a nivel global prediga adecuadamente un porcentaje alto, pero que a nivel particular no ocurra así con los fumadores, es decir, llevaría a errar al modelo al determinar como no fumador a un fumador. Se opta, a partir de la proporción de ceros y unos ya expuesta, por bajar este umbral a 0,3. Esto lleva a que el porcentaje correcto de predicción para los fumadores pase del 17,6% al 46,5%, más del doble de aciertos en puntos porcentuales, a pesar de que se pierda algo de porcentaje global al realizar esta modificación.

5.2.1. Resultados del modelo logit con todas las variables explicativas (LOG1)

Se pasa a estimar el primer modelo logit, de acuerdo con la estructura de variables planteada inicialmente: 33 variables explicativas más la constante del modelo. Además, se toman como variables de referencia las descritas en el inicio de este capítulo y ajustando el punto de corte para la clasificación en 0,3. En esta ocasión, la estimación se realiza sin hacer uso del factor de elevación, pero habiendo sido comprobada la similitud de los resultados en ambos casos, para que pueda contrastarse con los de esta misma estimación, pero con otro software. Esta comparación resultará de suma importancia en el siguiente subepígrafe.

Una vez estimado, el modelo logit inicial (en adelante LOG1) arroja, en términos de calidad de este, los resultados que se recogen en la tabla siguiente (tabla 9):

Tabla 9 Resultados LOG1

Pruebas ómnibus de coeficientes de modelo				
		Chi-cuadrado	**gl**	**Sig.**
Paso 1	**Paso**	3354,738	33	,000
	Bloque	3354,738	33	,000
	Modelo	3354,738	33	,000

Resumen del modelo			
Paso	**Logaritmo de la verosimilitud -2**	**R cuadrado de Cox y Snell**	**R cuadrado de Nagelkerke**
1	20466,325[a]	,136	,210

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001.

Prueba de Hosmer y Lemeshow			
Paso	**Chi-cuadrado**	**gl**	**Sig.**
1	12,608	8	,126

Tabla de clasificación[a]

			Pronosticado		
			FUMA		**Porcentaje correcto**
Paso 1	**Observado**		**NO**	**SÍ**	
	FUMA	**NO**	15450	2691	**85,2**
		SÍ	2616	2276	**46,5**
	Porcentaje global				**77,0**

a. El valor de corte es ,300

A nivel conjunto, se desea saber si los parámetros son significativos. Para ello, hay que enfrentar las estimaciones de estos a un contraste, cuya hipótesis nula (aquella que es objeto de rechazo o no rechazo) es que simultáneamente todos tomen el valor cero. Por tanto, se compara un modelo con variables explicativas y término independiente, con otro en el que solo esté el término independiente, en el caso de no rechazarse la hipótesis nula. A través de la prueba ómnibus de los coeficientes del modelo, se observa que la chi-cuadrado toma un valor de 3.354,738, que al ser comparado con el estadístico correspondiente ($\chi^2_{33;0,05} = 47,3999$) y concluirse que es mayor, la hipótesis nula es rechazada y, por tanto, los coeficientes son, en conjunto, significativos.

Si se aborda la significación individual de cada uno de ellos, cuya hipótesis nula vuelve a ser que el parámetro – individualmente considerado – tome valor cero, y se observa la tabla 10, donde se recoge el valor de la significación de cada uno de los coeficientes, se concluye que, a excepción de las CC.AA., todos los coeficientes son significativos al 99% de confianza (sig. < 0,01), así pues, las variables consideradas son estadísticamente significativas a la hora de explicar el comportamiento del individuo ante la decisión de fumar. Mención aparte merece la significación de

las Comunidades Autónomas. Estas cuentan con Andalucía como región de referencia. A este respecto, comunidades como Aragón, Madrid o Navarra, entre otras, presentan coeficientes no significativos a ninguno de los niveles aceptables, hecho que ya se puso de manifiesto en la estimación del modelo lineal. Esto indicaría que estas comunidades no presentan diferencias significativas con Andalucía, lo que haría plantear una posible similitud de comportamientos, de acuerdo con el factor geográfico, por parte de los individuos de estas zonas. En el subepígrafe 5.2.2. se aborda una propuesta de solución a este asunto, pues no sería coherente mantener en el modelo variables no significativas, es decir, que no tienen capacidad de explicar el comportamiento del sujeto objeto de estudio.

Tabla 10 Variables en la ecuación LOG1[22]

	B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B)	
							Inferior	Superior
SEXO	,200	,037	29,251	1	,000	1,221	1,136	1,313
EDAD	-,017	,001	161,398	1	,000	,983	,980	,986
ACTIVIDAD ECONÓMICA ACTUAL	,842	,044	368,926	1	,000	2,321	2,130	2,529
PAÍS DE NACIMIENTO	,484	,062	60,161	1	,000	1,622	1,435	1,833
SIN ESTUDIOS	-,302	,084	12,891	1	,000	,739	,627	,872
ESTUDIOS SECUNDARIOS, MEDIOS Y BACHILLERATO	,222	,055	16,406	1	,000	1,249	1,122	1,391
ESTUDIOS SUPERIORES	-,330	,063	27,795	1	,000	,719	,636	,813
BUENO_ESTADO DE SALUD	,217	,051	18,445	1	,000	1,242	1,125	1,372
REGULAR_ESTADO DE SALUD	,286	,064	19,922	1	,000	1,331	1,174	1,510
MALO_ESTADO DE SALUD	,380	,090	17,838	1	,000	1,462	1,226	1,744
MUY MALO_ESTADO DE SALUD	,660	,137	23,241	1	,000	1,934	1,479	2,529
ENFERMEDAD DE LARGA DURACIÓN O CRÓNICAS	-,145	,043	11,325	1	,001	,865	,795	,941
ACTIVIDAD FÍSICA	-,649	,046	195,870	1	,000	,522	,477	,572
CONSUMO DE ALCOHOL	,588	,038	233,675	1	,000	1,800	1,669	1,941
EXPOSICIÓN AL HUMO EN LUGARES CERRADOS	,010	,000	1020,534	1	,000	1,010	1,010	1,011
Constante	-2,118	,121	305,098	1	,000	,120		

También de la tabla 9 es posible extraer si el modelo es un buen ajuste, entendiendo que lo será en el caso de que sea capaz de predecir el valor observado. Se cuenta con dos métodos de comprobación: la prueba de Hosmer y Lemeshow (HL) y la tabla de clasificación, de la que además se desprende el análisis de la especificidad y la sensibilidad del modelo. En este caso, observado el valor que toma la prueba HL (12,608), partiendo de que la hipósteis nula es que el modelo es un buen ajuste y dado el valor de la chi-cuadrado para este caso ($\chi^2_{8;0,05} = 15{,}5073$), no se puede rechazar la hipótesis nula y, en consecuencia, el modelo estimado se ajusta a los datos de la muestra a un nivel aceptable.

Hasta aquí se ha trabajado con la primera estimación del logit. Ahora, en el siguiente subepígrafe, se aborda una nueva estimación que pretende solucionar el asunto de las Comunidades Autónomas. Será con este nuevo modelo donde se entre más en profundidad en la interpretación de este, así como en otros aspectos que no se han comentado hasta ahora.

[22] Para favorecer la adecuada lectura de los datos, se excluyen los resultados relativos a las Comunidades Autónomas. Estos datos se recogen de forma íntegra en el anexo IV.

5.2.2. Resultados del modelo logit con omisión de Comunidades Autónomas (LOG2)

Las Comunidades Autónomas, al menos una parte de ellas, presentan un problema a la hora de estimar el modelo, pues, como ya ocurrió en el modelo de probabilidad lineal, algunas de ellas no tienen consigo parámetros significativos. Esto implica que no deberían estar incluidas en la especificación del modelo, pues el valor natural del parámetro que los acompaña sería cero.

Lo anterior, ha exigido un replanteamiento del modelo, poniendo de relieve la posibilidad de agrupar las comunidades en, por ejemplo, dos grupos diferenciados a partir de los coeficientes y de la significatividad individual de estos.

Tras realizar un proceso iterativo de omisión de las variables que miden la pertenencia a cada una de las comunidades[23] y observado el valor que, a raíz de esta omisión, toman distintos criterios de información, se propone la formación del siguiente grupo de CC.AA., como grupo de referencia: Andalucía, Aragón, Baleares, Castilla y León, Cataluña, Galicia, Madrid, Navarra, País Vasco, La Rioja, Ceuta y Melilla. Un total de doce regiones pasan a formar un único grupo (CCAA1) incorporado en el término independiente del modelo. El resto formará también un único grupo (CCAA2) para facilitar la interpretación del modelo en términos de variaciones de pertenencia a un grupo u otro. Como resultado, no solo mejora la significatividad de los coeficientes (ahora todos son altamente significativos) sino que mejoran los tres criterios de información que se toman en consideración. La siguiente tabla (tabla 11) muestra esta evolución en los criterios.

Tabla 11 Criterios de información. Comparación LOG1 y LOG2

	MODELO LOG1	**MODELO LOG2** (donde se omiten las no significativas)
Criterio de Schwarz	*20.807,84*	*20.710,18*
Criterio de Akaike	*20.534,33*	*20.525,15*
Criterio de Hannan – Quinn	*20.623,20*	*20.585,28*

A partir de lo expuesto hasta ahora, se estima el modelo logit definitivo (LOG3). Este incorpora dos nuevas variables (CCAA1 y CCAA2) y deja de tener las 19 iniciales que conformaban las Comunidades Autónomas. Estas dos nuevas variables toman los siguientes valores:

Tabla 12 Valores de las variables CCAA1 y CCAA2 (modelo LOG3)

		CCAA1 (categ. de referencia)	**CCAA2**
Valores	**1**	Andalucía, Aragón, Baleares, Castilla y León, Cataluña, Galicia, Madrid, Navarra, País Vasco, La Rioja, Ceuta y Melilla	*Asturias, Canarias, Cantabria, Castilla – La Mancha, Com. Valencia, Extremadura y Murcia*
	0	*Para el resto*	*Para el resto*

[23] En el anexo V se presentan los resultados de la estimación, esta vez con el software Gretl, donde figura el modelo inicial y el modelo tras la omisión de las mencionadas Comunidades Autónomas. Nótese que estas estimaciones se realizaron con el punto de corte para la clasificación en 0,5 (por defecto en Gretl), dado que solo se pretende que sirvan de ilustración para observar la mejoría del modelo. Posteriormente se realiza la estimación con el umbral escogido: 0,3.

5.2.3. Resultados del modelo logit con agrupamientos de las Comunidades Autónomas (LOG3)

Se estima el modelo LOG3, originando los siguientes resultados (sin hacer uso del factor de elevación y con el umbral en 0,3). Como en el caso anterior, se muestra, en primer lugar, los resultados correspondientes a la calidad y bondad del modelo (tabla 13).

Tabla 13 Resultados LOG3

Pruebas ómnibus de coeficientes de modelo				
		Chi-cuadrado	**gl**	**Sig.**
Paso 1	**Paso**	3335,825	16	,000
	Bloque	3335,825	16	,000
	Modelo	3335,825	16	,000

Resumen del modelo			
Paso	**Logaritmo de la verosimilitud -2**	**R cuadrado de Cox y Snell**	**R cuadrado de Nagelkerke**
1	20485,239[a]	,135	,209

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001.

Prueba de Hosmer y Lemeshow			
Paso	**Chi-cuadrado**	**gl**	**Sig.**
1	10,426	8	,236

Tabla de clasificación[a]

			Pronosticado		
			FUMA		**Porcentaje correcto**
Paso 1	**Observado**		**NO**	**SÍ**	
	FUMA	**NO**	15458	2683	**85,2**
		SÍ	2642	2250	**46,0**
	Porcentaje global				**76,9**

a. El valor de corte es ,300

Si se aborda la significación individual, de acuerdo con lo que recoge la tabla 14, se observa que todos los coeficientes son significativos al 1% (sig. < 0,01). En consecuencia, ahora sí que todas las variables explicativas son relevantes para el modelo, lo que se traduce en que todas ellas contribuyen a explicar la decisión del individuo, dentro del caso que se está estudiando.

En términos de significación conjunta, se descarta nuevamente que los coeficientes tomen el valor cero, dado que el estadístico de contraste (prueba ómnibus de la tabla 13) resulta ser mayor que el valor crítico asociado a la chi-cuadrado con esos mismos grados de libertad ($3335{,}825 > 26{,}2962 = \chi^2_{16;005}$), resultando entonces rechazada la hipótesis nula del modelo. Por tanto, los parámetros son, en conjunto, significativos.

Si se analiza la bondad del ajuste tomando la prueba de Hosmer y Lemeshow, como ya se hizo con el LOG1, se observa que la chi-cuadrado asociada a esta prueba toma el valor 10,426. Al contrastar ese valor con el valor crítico correspondiente ($\chi^2_{8;0,05} = 15,5073$), se concluye que es menor y, por tanto, no se rechaza la hipótesis nula, que es la que conduce a afirmar que el modelo estimado se ajusta a los datos a un nivel aceptable. Esto significa que las diferencias entre los valores observados y los predichos son pequeñas, siendo esta la principal medida en la que se basa el contraste.

Por su parte, del análisis de la tabla de clasificación, se desprenden dos aspectos de sumo interés. Por una parte, se obtiene el dato de acierto global del modelo que, para el LOG3, es del 76,9%. Este porcentaje podría ser clasificado como aceptable, al ir acompañado además por el análisis favorable del HL, además de que, para modelos de este tipo, un ajuste correcto superior al 70% suele ser calificado como adecuado. Por otra parte, se puede analizar un segundo aspecto: la sensibilidad y la especificidad, que están asociadas a los niveles de acierto por categorías. En el caso de la especificidad, el modelo presenta un porcentaje del 85,2%, es decir, goza de una alta especificidad. Este porcentaje coincide con el de acierto que tiene el modelo en individuos no fumadores. Sobre la sensibilidad, el valor es inferior (46%), siendo este el porcentaje de aciertos que consigue la regresión cuando el individuo es fumador.

Adicionalmente, cabe destacar que este último modelo donde, además de eliminar las variables cuyos parámetros no eran significativos, se han agrupado las CC.AA., los criterios de información contemplados han vuelto a mejorar, teniendo un valor menor[24] a los que habían resultado hasta ahora.

Tabla 14 Variables en la ecuación LOG3

	B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B)	
							Inferior	Superior
SEXO	,201	,037	29,580	1	,000	1,222	1,137	1,314
EDAD	-,017	,001	156,905	1	,000	,983	,981	,986
CCAA2	,264	,037	50,598	1	,000	1,303	1,211	1,401
ACTIVIDAD ECONÓMICA ACTUAL	,841	,044	370,202	1	,000	2,319	2,128	2,526
PAÍS DE NACIMIENTO	,477	,062	59,550	1	,000	1,610	1,427	1,818
SIN ESTUDIOS	-,347	,083	17,437	1	,000	,707	,601	,832
ESTUDIOS SECUNDARIOS, MEDIOS Y BACHILLERATO	,216	,054	15,915	1	,000	1,241	1,116	1,380
ESTUDIOS SUPERIORES	-,329	,062	28,137	1	,000	,720	,637	,813
BUENO_ESTADO DE SALUD	,203	,049	16,893	1	,000	1,226	1,112	1,350
REGULAR_ESTADO DE SALUD	,276	,063	19,048	1	,000	1,318	1,164	1,492
MALO_ESTADO DE SALUD	,371	,089	17,241	1	,000	1,449	1,216	1,726
MUY MALO_ESTADO DE SALUD	,639	,136	21,977	1	,000	1,894	1,450	2,474
ENFERMEDAD DE LARGA DURACIÓN O CRÓNICAS	-,145	,043	11,447	1	,001	,865	,795	,941
ACTIVIDAD FÍSICA	-,656	,046	203,317	1	,000	,519	,474	,568
CONSUMO DE ALCOHOL	,576	,038	231,272	1	,000	1,780	1,652	1,917
EXPOSICIÓN AL HUMO EN LUGARES CERRADOS	,010	,000	1013,641	1	,000	1,010	1,010	1,011
Constante	-1,976	,110	321,068	1	,000	,139		

a. Variables especificadas en el paso 1: SEX, ED, CCAA2, ACT, NAC, NOEST, SEEST, SUEST, SBU, SRE, SMA, SMM, ELD, AFI, ALCOH, EXHUM.

[24] Ver anexo VI.

Entrando en las interpretaciones del modelo, es importante señalar que, en el modelo logit, las estimaciones de los parámetros no pueden interpretarse directamente como cambios en la probabilidad. Ahora bien, el signo[25] de esos parámetros sí que puede tomarse como relevante, pues indica la dirección en la que variará la probabilidad, ante variaciones en los regresores. Lo que sí es interpretable directamente es la columna Exp(B). Esta representa, para las variables cualitativas, los *Odds Ratios*[26] (OR) respecto a las categorías que han sido fijadas como de referencia. Por ejemplo, para la variable SEX, el valor del OR es 1,222, lo que indica que la probabilidad de fumar para un hombre es 1,222 veces mayor que para una mujer (que es la categoría de referencia). Esto coincide con lo ya expresado en el apartado de análisis descriptivo, donde se señaló que los hombres podrían presentar una mayor propensión a fumar basada en los datos de la muestra. En el caso de la variable CCAA2, su valor señala que el valor es 1,303 veces mayor para las comunidades del grupo 2 que para las del grupo 1. Otro ejemplo sería el de la variable SUEST, que toma valor uno si el individuo tiene estudios superiores. En este caso, la probabilidad es casi tres veces menor para un sujeto con estudios superiores frente a primarios. De forma análoga se podría proceder con todas las variables categóricas. En el caso de la variable EDAD, que no es categórica sino continua, su OR indica la ratio cuando se incrementa en una unidad la edad. Particularmente, es 0,2 veces menor cada vez que se incrementa una unidad.

Si se analizan el resto de los resultados expuestos en la tabla, se observa como el país de nacimiento del individuo incrementa la probabilidad de fumar si ha nacido en España frente a haber nacido en el extranjero. Igualmente ocurre con la actividad económica desempeñada actualmente por el individuo. Si este es activo, su probabilidad de elegir fumar es mayor.

Dos variables que precisan mención son AFI (realización de actividad física) y ALCOH (consumo de alcohol). Como se habría presupuesto inicialmente y en consonancia con los datos descriptivos, un individuo que practica deporte tiene una probabilidad menor de ser fumador frente a aquel que no realiza actividad física. En términos de OR se materializa en que una es más de cuatro veces menor en individuos que la realizan que los que no. En cambio, en el caso del alcohol, ocurre lo contrario, si el individuo lo consume de forma habitual, es más propenso al consumo también de tabaco, más concretamente, su probabilidad es 1,78 veces mayor que para el que no bebe habitualmente.

Aunque descriptivamente suponía una diferencia menor, el hecho de pasar más tiempo expuesto a humo en espacios cerrados, hace que la probabilidad sea más de una vez superior para aquellos que sí están expuestos.

Dentro del grupo de variables que recogen el estado de salud percibido por el individuo (SBU, SER, SMA y SMM) es importante señalar que todos sus coeficientes toman valores positivos y crecientes con el empeoramiento de la salud del individuo. Por ejemplo, la probabilidad es 1,449

[25] Si el signo es positivo, indicará un aumento, y si fuera negativo, ante un aumento, indicará una disminución (salvo en los casos de variables categóricas que directamente implicará una caída). Este signo irá también en consonancia con los OR, en tanto que, si el signo es negativo, el OR será menor que cero e indicará las veces menor que es una frente a la de referencia.

[26] "El cociente de *odds-ratios* cuando la variable explicativa x_{im} cambia en una unidad, es decir, $e^{(\beta_{jm}-\beta_{km})}, j = 1, \dots, J, j \neq k$, indica el efecto de dicho cambio en x_{im} sobre el patrón de sustitución entre las alternativas j y k, mientras que $e^{\beta_{jm}}$ mide el efecto sobre el patrón de sustitución entre la alternativa j la alternativa de referencia" (Guirao et al, 2016).

veces mayor para un individuo que declara tener un estado de salud malo frente al que declara tenerlo muy bueno. A este respecto, cabe sospechar que no se trata de que el individuo, al empeorar su salud comience a fumar, sino que podría ser el hecho de fumar lo que le hace, en ese momento, tener un estado de salud peor.

Quizás el dato más contradictorio es el aportado por la variable ELD (enfermedades de larga duración o crónicas). Descriptivamente se observó que, en términos globales, son más los individuos con enfermedades crónicas que no fuman (81,04%) que los que fuman (18,96%), aunque dentro de cada colectivo (por el efecto de la proporcionalidad intragrupo) el peso de los sujetos que pacen problemas crónicos es similar. Probablemente este sea el motivo que hace que la estimación de este parámetro sea negativa, con las implicaciones que ello tiene. Ahora bien, sería positivo estudiar la correlación que en estos términos existe, pues, a pesar de lo anterior, hay un alto porcentaje de exfumadores que padecen estas dolencias y, quizás, son exfumadores justo por esa razón.

Merece la pena, antes de finalizar este bloque de análisis e interpretación, reconocer que el análisis descriptivo, aunque parezca un instrumento "menos poderoso" que un modelo econométrico, es capaz de describir y dar una clara y acertada intuición de los resultados que, posteriormente, el modelo es capaz cuantificar en magnitudes y probabilidades.

Todos estos resultados son consistentes, como ya se ha especificado más arriba, con los obtenidos inicialmente con el modelo de probabilidad lineal. Además, se ha comprobado – queda a disposición del lector en el anexo VII – que los resultados obtenidos en el modelo LOG3, obtenidos sin hacer uso del factor de elevación, son idénticos en signo y significatividad, y muy próximos en magnitud, a los obtenidos en el mismo modelo, pero incorporando los factores de elevación en el momento de la estimación.

Para finalizar este capítulo, a continuación, se construye la ecuación de la regresión logística resultante del modelo planteado. Esta permite, a partir de la adecuación a las características y los atributos propios del individuo, obtener la probabilidad condicionada de que el individuo decida consumir tabaco, es decir, la probabilidad de que la variable dependiente, FUMA, tome valor uno.

$$P\left(FUMA = 1 / CARACTERÍSTICAS\ DEL\ INDIVIDUO\right) = \frac{1}{1 + e^{-(\beta X)}} \qquad (5.1)$$

Donde el valor de βX se obtendría al evaluar cada una de las variables de la regresión, de acuerdo con las características del individuo. Por ejemplo, para un hombre de Canarias con 25 años, estudiante, nacido en España, con estudios superiores, estado de salud regular, que no padece enfermedades crónicas, que no realiza actividad física ni tampoco bebe habitualmente, y que está expuesto a humo en espacios cerrados unos 30 minutos, tendría una probabilidad:

$$P\left(FUMA = 1 / CARACTERÍSTICAS\ DEL\ INDIVIDUO\right) = \frac{1}{1 + e^{-(-2{,}337)}} = 0{,}0881 \rightarrow 8{,}81\% \qquad (5.2)$$

Con esta probabilidad, por debajo del punto de corte para la clasificación (0,3), el individuo sería calificado como no fumador.

6. CONCLUSIONES Y LIMITACIONES

6.1. CONCLUSIONES

Tras el análisis de los datos y la propuesta de los modelos, se obtienen las siguientes conclusiones:

- La estimación del modelo de probabilidad lineal permite, de forma sencilla, observar las relaciones que en signo y significación de los parámetros puede presentar el resto de las modelos. Además, junto con los análisis descriptivos, estos sirven para establecer resultados esperados. Es importante señalar la más que similares conclusiones que se obtienen con el uso de todas estas técnicas.
- Desde 2006, año en el que entra en vigor la primera "ley antitabaco", el consumo de tabaco en España ha ido cayendo paulatinamente. A pesar de ello, presenta un porcentaje de fumadores superior al de la UE 28. La proporción de fumadores en la población española en el año 2017 era del 22,08%.
- El consumo de tabaco en los más jóvenes constituye un importante riesgo para la salud que debe ser controlado. Esta circunstancia se debe frenar con medidas efectivas que debiliten lo que se ha convertido en un rito hacia la madurez.
- De las variables sociodemográficas seleccionadas, la pertenencia a cierta Comunidad Autónoma parece, inicialmente, no ser una variable con poder de explicación, al menos, para el consumo de tabaco. Estas deben ser agrupadas en colectivos homogéneos, pero heterogéneos entre sí, para que constituya un factor explicativo.
- La franja de edades comprendida entre los 40 y los 60 años es en la que mayor concentración de fumadores se observa, siendo 47 años la edad media de la muestra fumadora (mientras que la de la muestra total es 53 años). Este hecho puede indicar que, con la entrada de las leyes, el consumo del tabaco se ha ido desincentivando, pero, para franjas de edad superiores a los 40 año, este efecto no ha sido tan eficaz, posiblemente, porque el hábito de fumar ya está consolidado en ellos.
- El nivel de estudios juega también un papel importante. Más del 50% de los fumadores cuenta con el nivel de estudios secundarios. En cambio, son los activos, si se habla de actividad económica, quienes más consumen tabaco. La caída del consumo en edades muy avanzadas y la menor propensión en edades muy tempranas, en ambos casos se trata de individuos inactivos, puede explicar este hecho.
- El 21,46% de los individuos nacidos en España fuma, mientras que, para los residentes nacidos en el extranjero, esta proporción baja hasta el 19,22%.
- La diferencia entre la tasa de fumadores que realizan actividad física y la de fumadores que no la realizan es de algo más de cinco puntos porcentuales, diferencia que se preveía mayor, ya no tanto por una mayor proporción de fumadores no deportistas, sino por un número menor de deportistas fumadores.

- Sobre hábitos saludables, se comprueba que los fumadores tienen mayor propensión al consumo de alcohol de forma habitual, que los no fumadores. Por tanto, parece que se podría tender a la acumulación de hábitos dañinos para la salud.
- Aunque en términos comparativos, fumadores y no fumadores presentan una proporción similar en el número de estos que padecen enfermedades de larga duración, sí que parece muy relevante la tasa de personas con estas dolencias que, en la actualidad, ya no fuman: un 29%.
- Se concluye obteniendo como resultado esperado una mayor propensión al consumo de tabaco para aquellos individuos que, aunque no cuenten con todas las características, presenten uno o varios de los atributos y/o hábitos analizados. Sin pretender generalizar en exceso, solo se enuncian características que podría tener un fumador español tipo, basando esta apreciación en, exclusivamente, el análisis planteado a lo largo de este estudio y en la mayor proporción que, sobre el total de fumadores, presenta cada colectivo. En particular, los hombres, entre los 40 y 60 años, con estudios secundarios, trabajando o en desempleo (activo), de nacionalidad española y que consume alcohol de forma habitual, sería aquel perfil individual que obtuviera una mayor probabilidad, *a priori*, de fumar.

6.2. LIMITACIONES

A lo largo del trabajo, el supuesto de racionalidad es el que ha primado en la decisión que toma el individuo. Ahora bien, economistas como, el Premio Nobel, Richard Thaler, apuntan a que los agentes económicos se desvían de forma sistemática del comportamiento racional. Justo esa posible irracionalidad representa una de las limitaciones detectadas en este trabajo. A pesar de que el individuo maximice su utilidad, al tener información completa, quizás debería plantearse decidir no fumar, dada su peligrosidad, en lugar de fumar. Aun así, el individuo decide consumir tabaco. El poder adictivo de este tipo de sustancias podría ser lo que lo lleva a fumar.

Otra limitación que presenta el trabajo es el nivel de sensibilidad que tiene, por ejemplo, el modelo LOG3. Se aconseja que tenga un nivel de especificad y sensibilidad en torno al 70%, porcentaje que sí alcanza la primera, quedándose la segunda en un 46%. Sería pertinente analizar este asunto, pues llevaría al modelo a mejorar aún más sus niveles de acierto.

Por último, para cerrar este trabajo de fin de grado, se proponen dos líneas de continuación que, a su vez. parten como una limitación. Se propone ampliar este trabajo con la introducción de atributos del producto que abarquen un abanico, sin duda, más amplio y exacto de alternativas, y también se propone la incorporación de variables de tipo económico (renta, precio y tipo impositivo) que permitan evaluar efectos concretos de política económica.

7. BIBLIOGRAFÍA

Cáceres Hernández, J. J. (2010). *Conceptos básicos de estadística para Ciencias Sociales*. Madrid: Delta Publicaciones.

Constitución Española, 27 de diciembre de 1978. *BOE, núm. 311.1*, 29 de diciembre, pp. 29315 – 29424. Recuperado de https://cutt.ly/xT0Dfx.

Domencich, T. y McFadden, D. (1975). *Urban Travel Demand: A Behavioural Analysis*. Amsterdam: North-Holland Publishing Company.

González Marrero, R. M., Guirao Pérez, G. y Pérez Marante, Nieves R. (1995). Modelos logit y probit binomiales. La estimación del valor del tiempo en la línea Gran Canaria-Tenerife. En A. Marrero (Presidencia). *Economía de los Servicios*. Conferencia llevada a cabo en el congreso V Congreso Nacional de Economía. Consejo General de Colegios de Economistas de España, Las Palmas de Gran Canaria.

Greene, W. H. (1999). *Análisis Econométrico (3ª ed.)*. Madrid: Prentice Hall Iberia.

Guirao Pérez, G., Rodríguez Donate, M. C., Cano Fernández, V. J. y Romero Rodríguez, M. E. (2016). Modelos de probabilidad y el consumo del vino. En *Investigaciones en Métodos Cuantitativos para la Economía y la Empresa* (pp. 359 – 392). Granada: Editorial Universidad de Granada.

Inglés, C., Delgado, B., Bautista, R., Torregrosa, M., Espada, J., García.Fernández, J., Hidalgo, M. D. y García-López, L. (2007) Factores psicosociales relacionados con el consumo de alcohol y tabaco en adolescentes españoles. *International Journal of Clinical and Health Psychology*, 7 (2), 403 – 420.

Instituto Nacional de Estadística (2018). *Encuesta Nacional de Salud de España (ENSE) 2017, Metodología*. [Versión electrónica]. Madrid. Recuperado el 10 de enero de 2019 de: https://cutt.ly/3T0A3z

Johnston, J. y Dinardo, J. (2001). Modelos de Variable Discreta y Variable Dependiente Limitada. En *Métodos de Econometría* (pp. 471 – 518). Barcelona: Vicens Vives.

Ley 28/2005, de 26 de diciembre, de medidas sanitarias frente al tabaquismo y reguladora de la venta, el suministro, el consumo y la publicidad de los productos del tabaco. *BOE, núm. 309*, 27 de diciembre, pp. 42241 – 42250. Recuperado de https://cutt.ly/ET0SvK.

Ley 42/2010, de 30 de diciembre, por la que se modifica la Ley 28/2005, de 26 de diciembre, de medidas sanitarias frente al tabaquismo y reguladora de la venta, el suministro, el consumo y la publicidad de los productos del tabaco. *BOE, núm. 318*, 31 de diciembre, pp. 109188 – 109194. Recuperado de https://cutt.ly/3T0SJp.

Oi, K. I. Y. y Shuldiner, P. W. (1962). *An Analysis of Urban Travel Demands*. Evanston: Northwerstern University Press.

Ortúzar, J. de D. y Willumsen, L. G. (2008). Modelos de elección discreta. En *Modelos de Transporte* (pp. 333 – 371). Santander: Ediciones de la Universidad de Cantabria.

Sánchez Queija, M., Moreno Rodríguez, M., Muñoz Tinoco, M. y Pérez Moreno, P. (2007) Adolescencia, grupo de iguales y consumo de sustancias. Un estudio descriptivo y relacional. *Apuntes de Psicología*, 25 (3), 305 – 324.

Stock, J. H. y Watson, M. M. (2012). Regresión con variable dependiente binaria. En *Introducción a la Econometría* (pp. 275 – 302). Madrid: Pearson Educación, S.A.

Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, 26, 24 – 36.

Warner, S. L. (1962). *Strategic Choice of Mode in Urban Travel: A Study of Binary Choice*. Evanston: Northwerstern University Press.