15. Dic. 97

Muchas gracias.

D/ Enrique GARCÍA PÉREZ con D.N.I 50.275.521
profesor de la Facultad de CC Económicas y Empresariales, adscrito al Dpto de Estadística e Investigación Operativa, se compromete a la consulta de la Tesis Doctoral , Análisis de la demanda de vivienda secundaria con modelos de elección discreta, de Mª Cruz MOLES MACHI con fines exclusivamente de investigación y a respetar los derechos de autor.

En Madrid a 22 de octubre de 1997.

UNIVERSITAT DE VALÈNCIA

FACULTAT DE CIÈNCIES ECONÒMIQUES I EMPRESARIALS
Departament d' Economía Aplicada

ANÁLISIS DE LA DEMANDA DE VIVIENDA SECUNDARIA CON MODELOS DE ELECCIÓN DISCRETA

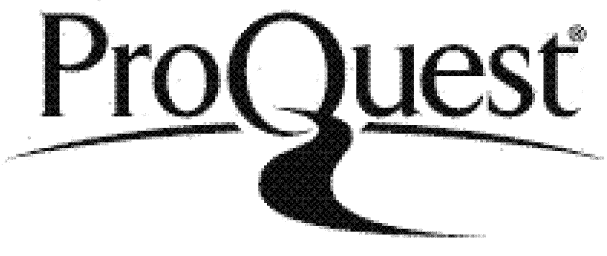
Tesis doctoral presentada por:

Mª Cruz Molés Machí

Directores de la tesis:

J. Santiago Murgui Izquierdo
Mª Luisa Moltó Carbonell

Valencia, 1996



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ÍNDICE

1. INTRODUCCIÓN

El objetivo que se pretende en este trabajo es analizar la demanda de vivienda secundaria en España. Se ha considerado que este estudio podría tener un notable interés por partida doble. En primer lugar por entrar en el marco de estudios del mercado de la vivienda, que es un sector muy importante en la Economía de cualquier país (la producción y el mantenimiento de la vivienda constituyen un importante segmento del sector productivo de la economía) y en segundo lugar por el hecho de que en la literatura econométrica el análisis de la vivienda secundaria, ha recibido todavía muy poca atención.

Debido a la importancia de este bien, desde el punto de vista de la política económica, se está muy interesado en los estudios sobre la vivienda. Para realizar análisis de la vivienda es necesario conocer las características que presenta la vivienda como mercancía y los mercados de la misma.

La vivienda es un bien de consumo duradero que tiene unas características particulares que lo distinguen de los usuales mercados de bienes y servicios.

En primer lugar, la vivienda tiene la particularidad de satisfacer la "necesidad" humana de resguardarse. También destaca el hecho de que representa una "inversión financiera", para algunos hogares es la mayor de las inversiones que realizan a lo largo de toda su vida. Este aspecto de la vivienda ha recibido recientemente considerable atención. Se han realizado debates sobre si ha habido cambios en el tipo de inversión, sobre si se ha sustituido la inversión en capital mobiliario por la inversión en vivienda, contribuyendo a la crisis de productividad.

Otra característica es la "fijación espacial" de las viviendas, éstas no pueden ser transportadas de un lugar a otro. La cuarta y la quinta características son la "indivisibilidad" de las viviendas, no siendo posible establecer fracciones de unidades de vivienda y el "alto coste" de producción. Otra característica del mercado de este bien es la "limitación" del mismo, resulta difícil que una familia encuentre una vivienda con las características ideales para ella.

La información en el mercado de la vivienda es imperfecta. El precio de una vivienda, en alquiler o en propiedad, depende del ofertante y del demandante. Si un ofertante pone su vivienda en el mercado a un cierto precio y no llega a un

acuerdo con un demandante, el riesgo de perder los ingresos asociados puede llevarle a reducir el precio de la vivienda.

Las familias no modifican instantáneamente la elección de su vivienda cuando las condiciones cambian. Habitualmente permanecen en una vivienda varios años y pocas familias se moverán inmediatamente en respuesta a la introducción de una nueva política de vivienda.

En la mayoría de los países, el sector público en todos los niveles se halla fuertemente involucrado en el mercado de la vivienda: interviene en la designación de la zona urbanizable (introduce controles sobre el uso y densidad del suelo), en el control de los alquileres, en la regulación del tipo de interés en los programas de renovación urbana, en los impuestos sobre la propiedad inmobiliaria, en los tratamientos fiscales favorables a la vivienda, en las subvenciones dirigidas a familias de renta baja, etc...

Los subsidios de vivienda se suelen encontrar implícitos en el impuesto sobre la renta. En muchos paises se otorga un tratamiento fiscal favorable a la vivienda ocupada por su propietario. Estos subsidios son de naturaleza política, y este tratamiento fiscal tiende a favorecer no sólo al beneficiario sino también a los productores del bien en cuestión. Si se incrementa la demanda se tenderá a beneficiar a la industria de la construcción.

En España los propietarios ocupantes de la vivienda se benefician de desgravaciones en el impuesto sobre la renta, ya que pueden deducirse de la base imponible todos los intereses de capitales ajenos invertidos en la compra de la vivienda, y en la cuota íntegra la desgravación es el 15% de los capitales invertidos. Además se les imputa una renta del 2% del valor catastral que está por debajo del valor del mercado.

Estos incentivos fiscales suponen el 74% del total de ayudas para la compra de viviendas en España. Comparando la distribución de las ayudas en algunos países desarrollados (Alemania, Dinamarca, España, Francia, Países Bajos, Reino Unido y Estados Unidos) el país cuyo porcentaje de ayudas fiscales es mayor es Estados Unidos con un 81%, seguido de España con el 74% y Dinamarca con un 70%, (Lasheras et al. ,1991).

Todas estas características comentadas hacen de la vivienda un producto de mercado muy importante en la economía de un país y por supuesto en la economía doméstica.

Un análisis sobre la demanda de vivienda puede plantearse desde distintos puntos de vista, según cual sea el objetivo perseguido. Una posible línea de trabajo es el estudio de las características propias del mercado de la vivienda, como la relación precio-calidad de las viviendas, cambios en la oferta o demanda de vivienda en respuesta a una nueva política económica (como una modificación en el tratamiento fiscal a los propietarios ocupantes de su vivienda), tipos de vivienda que están en el mercado, etc.

Otra posibilidad es plantear un análisis a nivel del propio hogar o familia que toma decisiones acerca de la vivienda.

Desde el punto de vista del hogar, la primera decisión que puede considerarse es la elección de la forma o régimen de tenencia de la vivienda que va a ocupar. En España se presentan dos opciones, alquiler y propiedad, aunque en diversos paises es común la propiedad cooperativa.

Una aproximación de los condicionantes de la elección de tenencia de vivienda, así como los parámetros que determinan el gasto realizado en ella, pueden proporcionar una base para determinadas actuaciones de política económica.

Cuando se realiza la elección de una vivienda, bien comprándola o bien alquilándola, la familia debe considerar que no sólo obtiene la unidad física sino también, a causa de la fijación espacial, un vecindario, un conjunto de bienes y servicios públicos y ciertas obligaciones impositivas. El tipo de vivienda que cada familia adquiere está, presumiblemente, muy relacionado con su renta y sus características demográficas, así como con los precios relativos (precios de mercado del resto de viviendas). Además, la composición de la familia (formación, disolución, número de hijos, etc...), el tipo de empleo y lugar de trabajo de sus miembros, la ubicación de la vivienda principal (municipio interior o costero), la distancia entre la vivienda principal y la secundaria, así como el uso que se piensa realizar de una vivienda secundaria podrían resultar también factores determinantes en el análisis.

En este trabajo se pretende realizar un análisis empírico sobre el mercado de la vivienda secundaria en España. El objetivo perseguido es modelizar el comportamiento de los hogares españoles frente a decisiones relativas a este tema, tales como disponer o no de una vivienda secundaria, el régimen de tenencia o el tamaño de la misma.

Los problemas que se pretende resolver no responden al esquema de una regresión clásica. Son situaciones en las que la variable o fenómeno de interés

(disponer o no de una vivienda secundaria, el régimen de tenencia, etc.) está determinado a partir de una serie de variables o características observables, pero la relación funcional entre ellas no puede plantearse utilizando el modelo lineal básico debido a que la variable o fenómeno estudiado no es de tipo continuo, sino que es una variable cualitativa.

Se pueden encontrar muchos ejemplos de este fenómeno en diversos campos de investigación, y no sólo en el análisis de la vivienda. La elección del régimen de tenencia de la vivienda principal entre comprar y alquilar, la reacción de un insecto que vive o muere, ante la administración de una droga, la elección del modo de transporte para ir al trabajo entre coche, autobús y metro, son ejemplos relacionados con este tipo de modelos.

Para analizar ese tipo de variable dependiente será necesario cuantificarla mediante la introducción de valores numéricos que representen las diferentes categorías. De esta forma se tendrá que analizar el comportamiento de una variable discreta en función de unas características o variables observadas y otras no observables o aleatorias.

Los modelos que plantean una relación funcional entre la variable respuesta y las variables observadas que determinan el fenómeno estudiado, buscando una explicación a su comportamiento, reciben el nombre de modelos de respuesta discreta o cualitativa, o modelos de elección discreta.

El objetivo de los modelos de respuesta cualitativa es la determinación de la probabilidad asociada a cada posible alternativa o valor de la variable discreta dependiente. La probabilidad buscada estará condicionada a las observaciones de los factores o características que determinan estos valores.

En ocasiones el valor observado de la variable dependiente es el resultado de un proceso de decisión en el que un individuo debe elegir una alternativa, de entre un conjunto de alternativas mutuamente excluyentes, basándose en una serie de características del propio individuo y de las alternativas. En estas situaciones los modelos de respuesta cualitativa pueden plantearse como un problema de maximización de la utilidad: un individuo, que se supone racional y con capacidad decisora, elegirá aquella alternativa que le produzca mayor utilidad.

También en este caso el objetivo de estos modelos es analizar la probabilidad de que el individuo elija cada una de las alternativas posibles en función de todas las características observadas sobre el individuo y sobre las alternativas.

La analogía entre los dos planteamientos anteriores ha condicionado la aparición de modelos de respuesta cualitativa o de elección discreta, genéricamente, englobando las dos situaciones y considerándolas como equivalentes.

Los modelos de respuesta cualitativa se clasifican según el número de alternativas que hay en el conjunto de elección (conjunto de las posibles alternativas) en modelos binomiales y modelos multinomiales. En el primer tipo el número de alternativas es dos y en el segundo este número debe ser superior a dos.

La elección entre comprar o alquilar una vivienda es una situación en la que se utilizará un modelo binomial. Sin embargo, si el problema ofrece como posibilidades alquilar una vivienda, comprarla o ser copropietario, será necesario buscar un modelo multinomial.

La literatura estadística proporciona además de los modelos de respuesta discreta otros modelos denominados genéricamente modelos de variable dependiente limitada y que tienen muchos puntos en común con los modelos de respuesta discreta.

Supóngase que se considera el problema de analizar el gasto que realizan las familias en vivienda. La literatura econométrica distingue entre ser propietario o inquilino de la vivienda, ya que el gasto que realiza una familia propietaria de su vivienda no es supuestamente el mismo que realiza una familia que disfrute de su vivienda en régimen de alquiler. Ahora se plantearía un modelo de respuesta discreta para la elección del régimen de tenencia de la vivienda y un modelo con dos ecuaciones para el gasto, una para propietarios y otra para inquilinos que deberán estimarse utilizando la información del modelo de elección discreta anterior.

En este ejemplo la variable dependiente es continua, pero su observación está limitada. Los individuos de la muestra están clasificados por alguna condición sobre las características observadas, y las variables dependientes continuas se observan según la condición especificada.

En esta tesis se realizará una revisión metodológica sobre todos estos modelos comentados. En este primer capítulo se ha realizado una breve introducción a los problemas planteados en el trabajo y a continuación se presenta la organización que se ha seguido.

En el capítulo 2 se ofrece una revisión de los artículos y trabajos que se han encontrado en la literatura econométrica acerca de la demanda de vivienda. El tipo de estudio que se pretende realizar sobre la vivienda secundaria en España tiene sus antecedentes en estudios similares realizados para la vivienda principal en diferentes países. Siguiendo la metodología de estos estudios, se desarrollarán los modelos correspondientes a la vivienda secundaria.

La revisión metodológica sobre los modelos de respuesta cualitativa y los modelos de variable dependiente limitada se recoge en el capítulo 3. Además de la presentación de los modelos y sus particularidades se incluye todo el proceso inferencial necesario para su completa determinación.

En el capítulo 4 se comenta la fuente de datos que se ha utilizado en el análisis de la vivienda secundaria, así como las muestras y las variables (dependientes y explicativas) de cada uno de los análisis particulares realizados.

El análisis de los resultados se realiza en el capítulo 5. En el primero de los análisis se estudian las características sociodemográficas y los factores económicos que determinan si los hogares disfrutan o no de vivienda secundaria; es decir, se pretende encontrar qué características son las que discriminan a las familias que tienen a su disposición una vivienda secundaria de aquellas que no la tienen, independientemente del régimen de tenencia en el que disfrutan de la misma.

Un segundo análisis que se ha considerado interesante es plantear un modelo que analice conjuntamente la elección del régimen de tenencia de la vivienda principal de las familias y la disponibilidad de la vivienda secundaria.

Posteriormente se estudia la determinación de las características que presentan aquellos hogares que disfrutan de más de una vivienda secundaria. En este caso se plantea la discriminación de los hogares que solo tienen una vivienda secundaria a su disposición frente a aquellos que tienen más de una.

Se comentan después los resultados obtenidos al modelizar el régimen de tenencia de la vivienda secundaria, considerando como variables determinantes del modelo el mismo conjunto de variables explicativas utilizado en el modelo que analiza si el hogar disfruta o no de vivienda secundaria.

A continuación se ha estimado un modelo para analizar las preferencias de los hogares por el tipo de vivienda. Se pretende analizar si los hogares prefieren viviendas secundarias unifamiliares o por el contrario prefieren viviendas situadas en edificios colectivos.

Otro aspecto de la vivienda que se ha considerado de notable interés para este trabajo es explicar el tamaño de la vivienda secundaria que está disfrutando el hogar a partir de características sociodemográficas del mismo.

Para finalizar, en el capítulo 6 se presentan los resultados y las conclusiones más destacadas del trabajo realizado.

2. LA DEMANDA DE VIVIENDA

Aunque es difícil encontrar estudios empíricos sobre la demanda de vivienda secundaria, sí que hay una amplia literatura sobre el análisis general de la vivienda. A continuación se realizará un breve repaso a los estudios más relevantes de la literatura econométrica sobre la vivienda.

Los análisis de demanda de vivienda empiezan a desarrollarse a partir de los años 50. Muth (1960) publica un artículo sobre la demanda de este bien. Este mismo autor en el año 1969 publicó un libro sobre la vivienda en las ciudades. También por esta época aparece un libro publicado por Reid (1962) dedicado a la demanda de vivienda.

En 1968, Lee publica un estudio sobre el gasto en servicios de vivienda. Este mismo año Winger realiza también un trabajo sobre el análisis del gasto en servicios de vivienda.

Los artículos anteriores son comparados en un artículo de De Leeuw (1971), quien realiza una revisión de cinco trabajos sobre la demanda de vivienda en Estados Unidos (dos de Reid, uno de Lee, uno de Muth y uno de Winger). La conclusión obtenida por el autor es que las diferencias en los resultados de los estudios anteriores se deben a las distintas definiciones dadas a la variable gasto en vivienda y a la variable renta del hogar, aparte de las diferencias debidas al uso de muestras diferentes y a la utilización de procedimientos de estimación diferentes. En el artículo, De Leeuw presenta además un análisis del modelo del gasto en vivienda utilizando la información empírica de hogares residentes en 19 áreas metropolitanas diferentes, obtenida del Census of Housing de 1960. Los datos considerados en el análisis están agrupados según el número de miembros del hogar. Este autor estima de forma separada las ecuaciones de gasto para los hogares propietarios e inquilinos de su vivienda. Como variables explicativas utiliza únicamente la renta disponible del hogar, sin ninguna otra variable explicativa que refleje las características sociodemográficas del hogar.

En ese mismo año Maisel, Burnham y Austin reestiman el modelo de De Leeuw con datos no agrupados. Estos autores critican la pérdida de eficiencia del estimador debido al uso de datos agrupados.

En 1972, Kain y Quigley realizan un estudio sobre la discriminación racial en el mercado de la vivienda. Analizan el régimen de tenencia de la vivienda considerando como variables explicativas las características sociodemográficas del hogar, incluyendo la variable raza del sustentador principal en este conjunto con datos referentes a 1.200 familias de St. Louis. Los resultados obtenidos indican que la raza es un factor que influye en el tipo de vivienda ocupada. También proponen la estimación de un segundo modelo en el que introducen además como variable explicativa una variable que indica el tipo de régimen de tenencia en que se disfrutaba la vivienda anterior del hogar.

En estos primeros años otro artículo interesante fue publicado por Carliner (1973). En éste se realiza una revisión del artículo de De Leeuw. Carliner estima el modelo del gasto en servicios de vivienda con datos de panel y con un conjunto de variables explicativas donde se incluyen variables demográficas como la raza, el sexo y la edad del sustentador principal. La estimación separada para las ecuaciones de gasto de los propietarios y los inquilinos se realiza mediante el método de los mínimos cuadrados ordinarios, y la muestra utilizada está sacada del Panel Study of Income Dynamics (PSID) dirigida por el Survey Research Center of the University of Michigan.

Li en 1977 analiza qué tipo de régimen de tenencia prefieren los hogares americanos. El objetivo del estudio es analizar las características sociodemográficas (edad, tamaño de la familia y raza) y los factores económicos del hogar (renta) que determinan la decisión de comprar o alquilar la vivienda en dos ciudades (Boston y Baltimore). El modelo utilizado es un modelo logit binomial y los datos son del Metropolitan Housing Characteristics del año 1970. En el trabajo se estima el modelo de elección del régimen de tenencia con diferentes conjuntos de variables explicativas. Considera como variables explicativas unas variables interacción que reflejan efectos conjuntos de otras dos variables, tal como el efecto conjunto de la edad y la raza del sustentador principal.

En 1978, Lee y Trost publican un trabajo sobre la elección del régimen de tenencia y el gasto en servicios de vivienda. Este estudio se realiza con un planteamiento diferente a los análisis anteriores. Lee y Trost destacan que las diferencias existentes en el gasto de los hogares propietarios de su vivienda y los inquilinos de la misma no queda reflejado adecuadamente si se considera una ecuación para cada tipo de régimen de tenencia y se estiman separadamente. Ellos consideran que los hogares deciden conjuntamente el régimen de tenencia de su vivienda y el gasto que van a realizar en ella. Por ese motivo debe realizarse una estimación conjunta de la elección del régimen de tenencia, propietarios e inquilinos, y el gasto en servicios de vivienda.

En su artículo presentan una metodología general para los modelos de variable dependiente limitada. Para la estimación de los correspondientes parámetros del modelo proponen dos procedimientos. En primer lugar un procedimiento de estimación en dos etapas, corrigiendo el sesgo de selección muestral con la inversa del ratio de Mills. Este procedimiento garantiza la consistencia del estimador. La segunda propuesta es realizar una estimación máximo-verosímil, utilizando como estimaciones iniciales las obtenidas en el procedimiento de estimación en dos etapas. Estos estimadores son consistentes y además eficientes asintóticamente. Lee y Trost sugieren también contrastar la correlación (supuestamente distinta de cero) entre la demanda de vivienda y el régimen de tenencia. El estadístico propuesto para contrastar la existencia de simultaneidad (correlación no nula) de las dos decisiones es el test de la razón de verosimilitudes.

Después de la primera parte metodológica, estos autores en su artículo presentan una aplicación para el análisis del gasto en servicios de vivienda con interdependencia en la elección del régimen de tenencia. El trabajo empírico se realiza con datos obtenidos del Panel Study of Income Dynamics (PSID) de 1972. El modelo se estima por los dos procedimientos de estimación propuestos en la primera parte del trabajo. El contraste de simultaneidad entre la decisión de comprar o alquilar la vivienda y la cantidad de servicios consumidos se realiza con el test de la razón de verosimilitudes. Para su muestra obtienen que existe evidencia de simultaneidad en las decisiones.

Otros estudios posteriores utilizan los resultados metodológicos obtenidos por Lee y Trost. Un destacado estudio sobre la elección del régimen de tenencia y el gasto en servicios de vivienda es realizado por Rosen en 1979. El objetivo de este autor es analizar los cambios producidos en la demanda de vivienda al introducir cambios en el tratamiento fiscal a los propietarios, para ello analiza el gasto en servicios de vivienda y la elección del régimen de tenencia de varios grupos de hogares. La fuente de información son datos de corte transversal obtenidos de la encuesta Panel Survey of Income Dinamics (PSID). Rosen se interesa únicamente por los propietarios de la vivienda y sólo comenta la ecuación de gasto correspondiente a los mismos. El procedimiento de estimación utilizado es el proceso de estimación en dos etapas y para analizar la simultaneidad realiza un contraste sobre la significatividad del parámetro que acompaña a la inversa del ratio de Mills. Los resultados del contraste indican que se debe rechazar la simultaneidad entre las dos decisiones. En la última parte del trabajo este autor investiga mediante simulaciones el efecto y las implicaciones que los cambios en la política fiscal tienen en la demanda de vivienda.

En 1980, King analiza para los hogares del Reino Unido la decisión conjunta de la elección del régimen de tenencia y la demanda en servicios de vivienda. En este modelo se impone la restricción de que ambas decisiones, discreta y continua, se obtienen a partir de un mismo orden de preferencias. El estudio se realiza para los hogares del Reino Unido y en este país el mercado de la vivienda presenta una importante intervención gubernamental en forma de subsidios financieros a la vivienda pública, concesiones en los impuestos a los propietarios de su vivienda y la legislación que controla los alquileres. Estas características, junto con limitaciones en la elección del régimen de tenencia, ya que no todos los hogares pueden acceder a las diferentes alternativas del régimen de tenencia, hacen que en el problema de maximización de la utilidad aparezcan restricciones en los parámetros del modelo y en la forma funcional de las ecuaciones. Los datos utilizados en este trabajo son 4.238 hogares extraídos del Family Expenditure Survey (FES) realizado en el período 1973/74. El conjunto de variables explicativas utilizado en este modelo no incluye las características propias del sustentador principal. El procedimiento de estimación utilizado es el de la máxima-verosimilitud conjunta.

Un estudio dedicado a la elección del régimen de tenencia de la vivienda es el de Henderson e Ioannides (1983). Estos autores plantean un modelo teórico de elección del régimen de tenencia teniendo en cuenta factores que son fundamentales en dicha elección, tales como las diferencias entre el coste de oportunidad de alquilar y comprar. Además en el artículo se analizan los cambios debidos a las imperfecciones del mercado de la vivienda y a los diferentes tipos de impuestos que tiene la vivienda en cada país. Otro artículo sobre la elección del régimen de tenencia y el consumo en servicios de vivienda de los mismos autores fue publicado en 1986. En este trabajo plantean el modelo de decisión desde la maximización de la utilidad con restricciones económicas. Previamente al análisis propuesto, estiman un modelo probit binomial para decidir si a una familia se le concede o no el préstamo hipotecario en función de sus características demográficas. Los datos para esta aplicación empírica corresponden a 2.663 familias proporcionadas por The Annual Housing Survey (AHS) de los Estados Unidos del año 1975.

En la literatura econométrica sobre la demanda de vivienda, se encuentran también artículos que analizan otros aspectos de vivienda, tal como el tiempo de permanencia en la vivienda, la localización de la misma, su tamaño y calidad o los diferentes sistemas de financiación para adquirir la vivienda. Algunos de ellos se comentan a continuación. Una característica del mercado de la vivienda es que los hogares pueden romper el equilibrio entre oferta y demanda. Los altos costes económicos y físicos de un cambio de domicilio son la causa de que algunas

familias aguanten tiempo en la vivienda que ocupan aunque no tengan la máxima utilidad.

Dynarski publica en 1985 un artículo donde analiza la existencia o no de equilibrio en los hogares que acaban de mudarse de domicilio, utilizando el hecho de ser propietario o inquilino como una variable de estratificación. El modelo estimado en el artículo es el gasto en servicios de vivienda con datos extraídos del Panel Study of Income Dynamics (PSID) de 1970.

En 1987, Henderson e Ioannides analizan la elección del régimen de tenencia de la vivienda y el gasto en servicios de la misma considerando los cambios de domicilio que realiza el hogar en el período de estudio. La muestra utilizada en el análisis está referida al período 1971/79 y consta de 686 familias obtenidas del Panel Study of Income Dynamics (PSID).

Krumm (1987) estima un modelo intertemporal de elección del régimen de tenencia. En su artículo extiende el análisis estático del modelo de elección del régimen de tenencia a una estructura intertemporal del mismo. Este autor estima un modelo logit multinomial con ocho alternativas utilizando los datos del Panel Survey on Income Dynamics (PSID) de Michigan del año 1976 hasta el año 1979 y analiza la variación que producen en las probabilidades de elección de la tenencia los cambios en el tiempo de las variables explicativas.

Rosenthal publicó en 1988 un artículo en el que presenta los resultados de un estudio sobre los cambios de domicilio de las familias. El estudio analiza el tiempo de permanencia de los hogares en la vivienda y la elección del régimen de tenencia en que el hogar disfruta la misma. La decisión del tiempo de permanencia se estima con un modelo semi-Markov y el régimen de tenencia con un modelo logit binomial cuyo conjunto de variables explicativas incluye, además de las características sociodemográficas y económicas del hogar, el régimen de tenencia en que el hogar disfrutaba su vivienda anterior. El análisis se realiza con 626 familias de Estados Unidos del Panel Study of Income Dynamics (PSID). Esta muestra se estratifica según el estado civil del cabeza de familia y el régimen de tenencia de la vivienda anterior del hogar.

Ese mismo año, Blackley y Ondrich (1988) plantean un modelo para la demanda de calidad, tamaño y distancia de la vivienda al centro de la ciudad. El estudio considera dos variables discretas, para la distancia al centro y el tamaño de la vivienda y una variable continua, para la calidad de la vivienda. El análisis conjunto de las variables discretas se realiza, primeramente, con un modelo logit multinomial y después con un modelo GEV (valor extremo generalizado) y la variable continua se modeliza mediante el usual modelo de regresión lineal. La

muestra consta de 470 familias que disponen de la vivienda en régimen de alquiler del área metropolitana de San Francisco.

Börsch-Supan y Pitkin en 1988 analizan varias especificaciones de modelos de elección discreta para estimar el consumo de vivienda. Con datos para el Área Metropolitana Albany-Schenectady-Troy de Nueva York del Annual Housing Survey de 1977 plantean un modelo de elección con nueve alternativas (para propietarios e inquilinos). Los modelos de respuesta cualitativa utilizados son el modelo logit anidado y un modelo de eliminación jerárquica.

Brownstone, Englund y Persson (1988) presentan un artículo en el que se realiza una revisión metodológica del modelo propuesto por Lee y Trost. Éstos obtienen el estimador en dos etapas desarrollado por Lee y Trost para un procedimiento de muestreo endógeno (basado en la elección) para algunos modelos de variable dependiente limitada. Utilizan sus resultados para analizar el gasto en vivienda y la elección del régimen de tenencia. El contraste de simultaneidad se realiza comparando los estimadores obtenidos por el método en dos etapas para los hogares propietarios e inquilinos con los estimadores máximo-verosímiles calculados conjuntamente, mediante el test de Hausman. Sus datos muestrales, obtenidos de la muestra HINK que realiza el Swedish Central Bureau of Statistics, les llevan a la conclusión que existe simultaneidad entre las dos elecciones.

También en el año 1988, Goodman realiza un estudio donde analiza diferentes especificaciones del modelo de elección del régimen de tenencia y del gasto en servicios de vivienda. En el trabajo se dedica una atención especial a la renta del hogar, la cual se desglosa en dos componentes, renta permanente y renta disponible. La aplicación empírica se realiza con datos de la muestra AHS National de 1978. En la estimación del modelo aparece un problema informático (no se pueden calcular las estimaciones máximo-verosimiles conjuntas) que no le permite contrastar la simultaneidad de las decisiones.

Hay diversos artículos que tratan el análisis de la elección del régimen de tenencia y el gasto en servicios de vivienda en otros países. Grootaert y Dubois (1988) plantean el estudio para Costa de Marfil con datos de 1979. Éstos estiman únicamente la ecuación de gasto para los hogares propietarios de su vivienda, pero corregida para considerar la influencia de la simultaneidad. El contraste sobre la simultaneidad, como hacía Rosen en su artículo, lo efectúan contrastando la significatividad del coeficiente que acompaña al ratio de Mills. Con sus datos llegan a la conclusión que dicho coeficiente no es significativo y por lo tanto el modelo no presenta simultaneidad.

Horioka en 1988 realiza un artículo sobre la decisión conjunta del régimen de tenencia y del gasto en servicios de vivienda en Japón. En el trabajo se estima la renta del hogar y los precios de las viviendas. Los resultados obtenidos se comparan con los de otros países. La muestra es del año 1981 y consta de 1.075 hogares obtenidos de la SBC Survey que realizó la Universidad de Tokio (Sociology Departamente of Faculty of Letters) y el Japan Research Center. La estimación del gasto en servicios de vivienda se realiza sólo para el subgrupo de los propietarios de la vivienda, ya que no dispone de datos para los inquilinos. Dicha estimación de los parámetros del modelo se realiza por el procedimiento en dos etapas desarrollado por Lee y Trost y por el procedimiento de máxima-verosimilitud conjunta. Con los datos disponibles, el contraste de significatividad sobre el parámetro asociado a la inversa del ratio de Mills indica que no hay simultaneidad en las dos decisiones realizadas.

Zorn (1988) realiza un estudio sobre el régimen de tenencia y la movilidad de los hogares en Corea. Con un modelo logit binomial estudia la decisión del hogar sobre si comprar o alquilar su vivienda y con un modelo logit multinomial los cambios de domicilio que realizan los hogares. La información muestral es del año 1982 y consta de 1.600 hogares de la ciudad de Seul (Corea). La muestra fue realizada por The Korean Research Institute for Human Settlements under Bureau of Statistics.

En 1989, Henderson e Ioannides analizan conjuntamente la elección del régimen de tenencia, el tiempo de permanencia en el domicilio y los niveles de consumo de la vivienda en Estados Unidos. Los datos utilizados son datos de panel que permiten analizar el tiempo que permanece una familia en una vivienda antes de cambiar de domicilio. La fuente de información es Panel Study of Income Dynamics (PSID) durante el período 1971/81. Las conclusiones obtenidas son que las familias cuyo sustentador principal tiene mediana edad, un nivel de renta elevado y estudios superiores, tiene una predisposición mayor a cambiar de domicilio.

Gabriel y Rosenthal (1989) publican un artículo en el que analizan la elección de la localización de la vivienda mediante un modelo logit multinomial considerando como alternativas cinco distritos del área metropolitana de Washington. Los datos del estudio son del año 1981 y provienen de Washington D. C. metropolitan area file of the American Housing Survey. El análisis se repite con otras dos muestras, la primera formada por los hogares cuyo cabeza de familia es blanco y la segunda formada por aquellos en los que el cabeza de familia es negro.

Otro trabajo relacionado con la vivienda es el de Börsch-Supan y Stahl (1991) quienes analizan el tipo de ahorro elegido por los hogares en Alemania cuando van a comprar una vivienda mediante un modelo logit multinomial.

Edin y Englund publican un artículo en 1991 sobre el gasto en servicios en vivienda y la elección del régimen de tenencia en Suecia, utilizando la base de datos HUS de 1984. El esquema utilizado por estos autores es el propuesto por Lee y Trost, aunque ellos incorporan una variable que refleja si el hogar ha realizado un cambio de domicilio recientemente o no.

También Brownstone y Englund (1991) se plantean analizar la elección del régimen de tenencia en Suecia con tres posibles alternativas de elección: compra, alquiler o propiedad compartida, el modelo utilizado es el logit multinomial. Además, analizaron la demanda de vivienda de forma separada para los hogares que disfrutan de la vivienda en propiedad y para los que poseen un apartamento en multipropiedad; en este análisis se toma en consideración el ahorro o riqueza del hogar. Utilizan datos de la muestra HUS de 1986.

Finalmente, en la literatura econométrica se pueden encontrar otros muchos artículos sobre la demanda de vivienda. Para completar esta revisión metodológica se comenta a continuación un estudio realizado en España sobre la demanda de vivienda.

El análisis del gasto en servicios de vivienda y la elección del régimen de tenencia se ha realizado para la Comunidad Autónoma de Andalucia por Jaen y Molina (1994). Los datos utilizados para este trabajo son los que proporciona la Encuesta de Presupuestos Familiares de 1980/81. Las ecuaciones de gasto estimadas, para los propietarios y los inquilinos, tienen corregido el sesgo de selección con la inversa del ratio de Mills. Con los datos utilizados se obtiene que tanto para los propietarios como para los inquilinos, el coeficiente de correlación de la ecuación de gasto y la de la elección del régimen de tenencia resulta significativamente diferente de cero, lo cual lleva presumiblemente a la existencia de simultaneidad entre el gasto en vivienda y la elección del régimen de tenencia. Estos mismos autores realizan un análisis metodológico del mercado de la vivienda en España, analizando sus características tanto a nivel individual de consumo como a nivel de economía del país. En su monografía publicada en el año 1995 se encuentra una revisión completa de sus estudios y de estudios de otros autores.

3. REVISIÓN METODOLÓGICA

En este capítulo metodológico se realizará la presentación y análisis de los modelos de respuesta discreta y los modelos de variable dependiente limitada.

En el epígrafe 3.1 se recogen los modelos de respuesta discreta. Se realiza la presentación de los mismos clasificándolos según las hipótesis distribucionales introducidas en cada uno de ellos. Asímismo se tienen en cuenta las características de la situación que se pretende modelizar.

Así, si las alternativas del conjunto de elección presentan una cierta similitud, se utilizarán modelos que puedan reflejar esta estructura de relaciones. Los modelos que consideran grupos de alternativas similares se denominan modelos anidados. Si las alternativas presentan una relación de orden, se hablará de modelos ordenados.

En determinadas ocasiones los modelos reflejan el proceso de decisión seguido por el individuo. Cuando el decisor realice la elección por eliminar alternativas que no le resulten interesantes se tendrán los modelos de eliminación.

El epígrafe 3.2 desarrolla los modelos de variable dependiente limitada. Dichos modelos son adecuados a situaciones en las que la variable dependiente es una variable continua pero existe un punto de truncamiento.

Tras la exposición de los diferentes modelos se proporciona el proceso inferencial sobre los mismos necesario para su completo análisis. Los epígrafes 3.3 y 3.4 están dedicados a la estimación de los parámetros desconocidos de los modelos.

También se ofrecen diferentes posibilidades de contrastación de hipótesis acerca de estos modelos en el epígrafe 3.5.

La naturaleza del problema a estudiar indicará el tipo de modelo más conveniente para describir la situación de interés, así como el proceso inferencial más adecuado al mismo.

Al final del capítulo 3 se recoge en un apéndice la demostración de las propiedades de que gozan los estimadores propuestos en este trabajo.

3.1. Modelos con variable dependiente discreta

En este epígrafe se presentarán los modelos, genéricamente denominados modelos de respuesta cualitativa, que tratan de explicar el comportamiento de una variable cualitativa en función de unas variables observadas.

Para facilitar la exposición se comenzará con la situación más simple en la que la variable dependiente puede tomar únicamente dos valores o modalidades. Se denominan modelos binomiales.

Tras este planteamiento se realizará la generalización a la situación en la que el número de categorías o modalidades de la variable respuesta es superior a dos, pero en cualquier caso será un número finito. En este caso nos referiremos a modelos multinomiales.

Los modelos de respuesta discreta suponen una relación funcional entre las características observadas y la probabilidad de que la variable respuesta tome cada uno de sus valores posibles.

En el caso de los modelos binomiales, los posibles valores de la variable respuesta y son dos, que sin pérdida de generalidad denotamos como 0 y 1.

Denotando genéricamente como individuo i a la unidad sobre la cual se analiza la variable respuesta y considerando su vector de características observadas x_i, los modelos de respuesta cualitativa permitirán localizar la probabilidad de que la respuesta del individuo i sea el valor 0 o el valor 1 condicionada a la información que aporta para el problema el vector de características x_i que sobre el individuo i se ha observado. Se pretende encontrar la probabilidad de $y_i = j$ en función del vector x_i; es decir, $\mathrm{P}(y_i = j / x_i)$, $j = 0, 1$.

Este planteamiento del problema permite obtener dos posibles soluciones. Una primera es la que proporcionará estimaciones de estas probabilidades directamente. Considerando un modelo de Bernouilli ya que la población en estudio es dicotómica, el interés está en estimar la proporción poblacional de éxitos, es decir la probabilidad de que $y_i = 1$. Para ello se extrae una muestra de observaciones y_i, $i = 1,.....,N$, y si es una muestra aleatoria simple puede utilizarse la proporción muestral como estimación, ya que es la estimación

máximo-verosímil de la proporción poblacional. Esta solución no permite encontrar la relación que liga la respuesta del individuo, y_i, con sus características x_i.

La segunda solución trata de obtener la probabilidad de respuesta considerando la dependencia entre la variable dependiente y_i, y las variables explicativas x_i. Se trata de determinar la respuesta de un individuo en función de sus características. De esta forma, ante un nuevo individuo se podrá efectuar una predicción sobre el valor que la variable dependiente tomará para dicho individuo.

3.1.1. Modelo de variable latente

El primer paso para realizar el análisis de la variable dependiente sería, si esta variable fuera de carácter continuo, plantear un modelo de regresión lineal. Sin embargo, el carácter cualitativo de la variable respuesta dicotómica no permite utilizar estos métodos de análisis.

La solución a este problema viene por introducir una variable intermedia y ficticia, no observable, que sea continua, definida a partir de las características observables como un modelo de regresión lineal: $y_i^* = x_i'\beta - \varepsilon_i$, y de la cual únicamente será necesario observar el signo.

A esta variable, denominada variable latente, se le exige que, aunque no sea observable su valor, permita definir la variable respuesta dicotómica a partir de ella como:

$$y_i = \begin{cases} 1 & sii \quad y_i^* \geq 0 \\ 0 & sii \quad y_i^* < 0 \end{cases}$$

En el modelo de variable latente, β es un vector de parámetros desconocidos y ε_i es la perturbación aleatoria que recogerá la influencia del individuo o de las propias modalidades de la respuesta, que no son observables pero que pueden determinar de alguna forma la respuesta del individuo.

Elegir el valor 0 como umbral de decisión para los valores de la variable respuesta y_i no presenta ninguna restricción en el modelo, ya que siempre puede redefinirse la variable latente y_i^* hasta conseguir este valor como límite.

Por ejemplo, si un investigador considera más adecuado limitar la variable latente por una constante a, se define una nueva variable latente que recoja dicha constante, $z_i^* = y_i^* - a$. Y la variable respuesta definida a partir de ella tiene el valor cero como umbral de decisión:

$$y_i = \begin{cases} 1 & sii \quad z_i^* = y_i^* - a \geq 0 \\ 0 & sii \quad z_i^* = y_i^* - a < 0 \end{cases}$$

Para cualquier otro caso se procedería del mismo modo.

La probabilidad de respuesta $P(y_i = 1 / x_i)$ se calcula, considerando la relación entre la variable respuesta y_i y la variable latente y_i^* como:

$$P(y_i = 1 / x_i) = P(y_i^* \geq 0 / x_i) = P(x_i'\beta - \varepsilon_i \geq 0 / x_i) = P(x_i'\beta \geq \varepsilon_i / x_i) = F_{\varepsilon_i}(x_i'\beta) \tag{3-1}$$

siendo $F_{\varepsilon_i}(\cdot)$ la función de distribución de la perturbación aleatoria ε_i del modelo de variable latente condicionada a la observación del vector de características x_i.

Análogamente se podría calcular la probabilidad para el segundo valor de la variable respuesta

$$P(y_i = 0 / x_i) = P(y_i^* < 0 / x_i) = P(x_i'\beta - \varepsilon_i < 0 / x_i) = P(x_i'\beta < \varepsilon_i / x_i) = 1 - F_{\varepsilon_i}(x_i'\beta) \tag{3-2}$$

O bien, como

$$P(y_i = 0 / x_i) = 1 - P(y_i = 1 / x_i) = 1 - F_{\varepsilon_i}(x_i'\beta)$$

De esta forma, la probabilidad de que la variable respuesta tome cada uno de sus valores se calculará a partir de la distribución de probabilidad de la variable aleatoria ε_i que define el modelo de variable latente.

Diferentes especificaciones de la función de distribución $F_{\varepsilon_i}(\cdot)$ darán lugar a diferentes expresiones para las probabilidades $P(y_i = 1 / x_i)$ y $P(y_i = 0 / x_i)$.

Estas expresiones funcionales de las probabilidades de respuesta son las que darán nombre a los diferentes modelos de respuesta cualitativa.

A continuación se va a generalizar esta solución para el caso en el que la variable respuesta tome más de dos valores, es decir, para el caso multinomial.

Ahora la respuesta del individuo será una de las J posibles modalidades de la variable dependiente.

Para resolver esta situación multinomial, la variable respuesta se transforma en una variable J-dimensional, donde cada componente y_{ij} toma el valor 1 si la respuesta del individuo es la categoría j y 0 en otro caso. Es decir, para un individuo i, su respuesta será un vector J-dimensional $(y_{i1},.....,y_{iJ})$ cuyas componentes serán todo ceros salvo la componente correspondiente a la alternativa elegida por dicho individuo, que tomará el valor 1.

El modelo de variable latente univariante $y_i^* = x_i'\beta - \varepsilon_i$ considerado para la situación binomial, se generaliza ahora a un vector de variables latentes, que se construye a partir de una transformación lineal de las características observadas sobre el individuo i y las J alternativas de elección.

La variable latente se define como un vector h-dimensional $y_i^* = x_i'\beta - \varepsilon_i$ donde β es un vector de k parámetros desconocidos, ε_i es un vector aleatorio con h componentes, y x_i la matriz de características observadas de orden $h \times k$.

A partir del modelo de variable latente, la variable respuesta puede definirse como el vector $y_i = (y_{i1},.....,y_{iJ})'$ cuyas componentes verifican que $y_{ij} = 1,\ y_{ir} = 0,\quad \forall r \neq j \quad sii \quad y_i^* \in A_j,\quad j \in \{1,.....,J\}$ siendo A_j el elemento j-ésimo de una partición de R^h.

Considerando la definición del modelo de variable latente se obtiene que la probabilidad de respuesta se calculará como:

$$P_{ij} = P(y_i = j / x_i) = P(y_{ij} = 1, y_{ir} = 0, \forall r \neq j / x_i) = P(y_i^* \in A_j / x_i) =$$

$$P(x_i'\beta - \varepsilon_i \in A_j / x_i) \qquad (3\text{-}3)$$

cuya expresión se obtendrá a partir de la función de distribución F_{ε_i} del vector aleatorio ε_i.

El número de componentes del vector de variables latentes, h, la estructura de los conjuntos A_j de la partición de R^h y la forma de la función F_{ε_i} serán los que determinarán el modelo concreto que se está planteando para analizar la situación.

3.1.2. Maximización de la utilidad

Como se ha comentado anteriormente, los modelos de variable dependiente discreta pueden interpretarse como el resultado de un proceso de maximización de la utilidad. En este apartado se van a plantear los modelos de respuesta cualitativa desde este punto de vista.

Se asume que el individuo i tiene capacidad decisora y el valor que toma la variable dependiente será la elección que él ha realizado entre todas las alternativas que tenía a su disposición.

Los modelos de respuesta cualitativa, desde el punto de vista de la maximización de la utilidad plantean modelizar el comportamiento de un individuo, o decisor, ante el problema de elegir una de las J posibles alternativas que tiene a su disposición. Este comportamiento va a estar influenciado por características propias del individuo y de las alternativas. El objetivo será encontrar de qué forma estas características determinan la decisión final del individuo.

Bajo el planteamiento de considerar que el individuo tiene capacidad decisora se está asumiendo la existencia de una relación de preferencias entre las alternativas. El decisor elegirá la alternativa que ocupa el primer lugar en esta relación.

Para establecer esta relación de preferencias el investigador deberá encontrar una forma de cuantificar la importancia que el decisor da a cada alternativa. En esta situación se define una función que a cada alternativa le asignará un valor, dependiendo de las características del individuo y de las alternativas de elección.

Esta función se conoce con el nombre de función de utilidad y su objetivo es cuantificar la importancia o utilidad que el decisor o individuo i da a una alternativa frente al resto. Denotando como U_{ij} a la utilidad que el individuo i tiene si elige la alternativa j y suponiendo un comportamiento racional del

decisor, se tendrá que el individuo i elige la alternativa $j \in C = \{1,2,....,J\}$ si y solo si $U_{ij} \geq U_{ir}, \quad \forall r \neq j, r \in C.$

Es decir, un individuo racional elegirá aquella alternativa que le proporcione la mayor utilidad.

Aunque la utilidad, U_{ij}, va a depender de las características del individuo y de las alternativas, no todas ellas son observables y en consecuencia será necesario asumir una descomposición de la utilidad en una componente determinista y observada y otra componente no observable y aleatoria.

El vector de características que determinan la utilidad y en consecuencia, la decisión del individuo, estará formado por dos componentes: un subvector x_{ij} que incluirá a todas las características observadas y que estará en la parte determinista de la utilidad, y una variable aleatoria ε_{ij} que recogerá la influencia de todos los factores y características no observables y constituirá la parte aleatoria de la utilidad.

Admitiendo una estructura lineal entre las componentes de la utilidad puede escribirse $U_{ij} = V_{ij} + \varepsilon_{ij}, \quad j = 1,....., J$, siendo V_{ij} la parte determinista que estará identificada a partir del vector de características observables x_{ij} y para la que sin pérdida de generalidad se puede asumir la relación $V_{ij} = x_{ij}' \beta.$

Puesto que el decisor elegirá la alternativa que le proporcione la mayor utilidad, las probabilidades de respuesta, P_{ij}, se calcularán como:

$$P_{ij} = P\left(y_{ij} = 1,\ y_{ir} = 0 \quad \forall r \neq j, r \in C / x_i\right) = P\left(U_{ij} \geq U_{ir},\ \forall r \neq j, r \in C / x_i\right) =$$

$$P\left(V_{ij} + \varepsilon_{ij} \geq V_{ir} + \varepsilon_{ir},\ \forall r \neq j,\ r \in C / x_i\right) =$$

$$P\left(\varepsilon_{ir} \leq V_{ij} - V_{ir} + \varepsilon_{ij},\ \forall r \neq j,\ r \in C / x_i\right)$$

Es decir, la probabilidad de elegir la alternativa j se calculará a partir de la distribución de probabilidad asignada a las variables $\varepsilon_{ir}, r \in C$. Denotando por $f_{\varepsilon_i}(\cdot)$ a la función de densidad conjunta del vector aleatorio $\varepsilon_i = \left(\varepsilon_{i1},....,\varepsilon_{iJ}\right)'$, esta probabilidad se calculará como:

$$P_{ij} = \int_{-\infty}^{+\infty}\int_{-\infty}^{V_{ij}-V_{i1}+\varepsilon_{ij}} \cdots\cdots \int_{-\infty}^{V_{ij}-V_{iJ}+\varepsilon_{ij}} f_{\varepsilon_i}(\vec{t})\, d\vec{t} \qquad (3\text{-}4)$$

resolviendo la primera integral para la componente ε_{ij}.

Con diferentes especificaciones distribucionales para el vector aleatorio ε_i se obtendrán diferentes expresiones de estas probabilidades de respuesta que darán lugar a distintos modelos.

Analizando los modelos de respuesta cualitativa desde el punto de vista de la maximización de la utilidad puede reducirse en un grado la integración necesaria:

$$P_{ij} = P\left(V_{ij} + \varepsilon_{ij} \geq V_{ir} + \varepsilon_{ir}, \forall r \neq j, r \in C / x_i\right) =$$

$$P\left(V_{ij} - V_{ir} \geq \varepsilon_{ir} - \varepsilon_{ij}, \forall r \neq j, r \in C / x_i\right) = F_{\varepsilon_{i-j}}\left(V_{i-j}\right)$$

siendo $F_{\varepsilon_{i-j}}(\cdot)$ la función de distribución del vector aleatorio J-1 dimensional $\varepsilon_{i-j} = \left(\varepsilon_{i1} - \varepsilon_{ij}, \ldots\ldots, \varepsilon_{iJ} - \varepsilon_{ij}\right)'$ y V_{i-j} representa el vector de J-1 componentes de la forma $V_{ij} - V_{ir} \quad r \neq j, r = 1,2,\ldots\ldots,J$.

Así:

$$P_{ij} = F_{\varepsilon_{i-j}}\left(V_{i-j}\right) = \int_{-\infty}^{V_{ij}-V_{i1}} \cdots\cdots \int_{-\infty}^{V_{ij}-V_{iJ}} f_{\varepsilon_{i-j}}(\vec{t})\, d\vec{t} \qquad (3\text{-}5)$$

con $f_{\varepsilon_{i-j}}(\cdot)$ representando a la función de densidad conjunta del vector ε_{i-j}.

El planteamiento de la teoría de la maximización de la utilidad es equivalente al del modelo de variable latente. Obsérvese esta equivalencia en el caso más simple de un modelo binomial:

$$P(y_i = 1 / x_i) = P(y_{i1} = 1, y_{i0} = 0 / x_i) = P(U_{i1} \geq U_{i0} / x_i) =$$

$$P(V_{i1} + \varepsilon_{i1} \geq V_{i0} + \varepsilon_{i0} / x_i) = P(V_{i1} - V_{i0} \geq \varepsilon_{i0} - \varepsilon_{i1} / x_i) = F_{\varepsilon_{i0}-\varepsilon_{i1}}(V_{i1} - V_{i0})$$

siendo $F_{\varepsilon_{i0}-\varepsilon_{i1}}(\cdot)$ la función de distribución de la variable aleatoria $\varepsilon_{i0} - \varepsilon_{i1}$ condicionada al vector de características x_i.

Tomando la hipótesis de que $V_{i1} - V_{i0}$ sigue una relación lineal de la forma $x_i'\beta$ y denotando por ε_i a la diferencia $\varepsilon_{i0} - \varepsilon_{i1}$ se tiene que la probabilidad de respuesta $P(y_i = 1 / x_i)$ se puede expresar como $F_{\varepsilon_i}(x_i'\beta)$. Es decir, la misma relación (3-1) obtenida desde el modelo de variable latente $y_i^* = x_i'\beta - \varepsilon_i$.

No obstante, la facilidad de interpretación de las probabilidades de elección a partir de la teoría de la maximización de la utilidad, es la razón por la que se utilizará este planteamiento para deducir las probabilidades de elección en los modelos multinomiales.

Aunque en cualquier caso, tanto desde el modelo de variable latente como desde la maximización de la utilidad, el cálculo de las probabilidades de respuesta está supeditado a la especificación de una distribución de probabilidad para la variable aleatoria que determina el modelo. A continuación se desarrollarán los modelos más usuales clasificados según la distribución de probabilidad asignada.

3.1.3. Modelo Probit

Cuando se utiliza la distribución Normal para la componente aleatoria en un modelo de respuesta cualitativa, se habla de un modelo probit.

Modelo Probit Binomial

Si el modelo es binomial, es decir, únicamente hay dos alternativas, se considera que la distribución de probabilidad asignada a la variable ε_i es la de una variable Normal tipificada. La probabilidad de respuesta vendrá dada por:

$$P(y_i = 1 / x_i) = \Phi(x_i'\beta) = \int_{-\infty}^{x_i'\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt \qquad (3\text{-}6)$$

El hecho de considerar la distribución tipificada no supone ninguna restricción, ya que si la media no es cero, $\mu \neq 0$, bastará incluir un término constante en la relación lineal $x_i'\beta$ que lo tenga en cuenta y si la varianza es distinta de la unidad, $\sigma^2 \neq 1$, la probabilidad anterior adoptará la forma:

$$P(y_i = 1 / x_i) = P(x_i'\beta - \varepsilon_i \geq 0 / x_i) = P(x_i'\beta \geq \varepsilon_i / x_i) =$$

$$P\left(\frac{x_i'\beta}{\sigma} \geq Z / x_i\right) = \Phi\left(\frac{x_i'\beta}{\sigma}\right)$$

siendo

$$Z \xrightarrow{d} N[0,1]$$

En este último caso, el vector de parámetros β estará afectado por esta varianza.

Modelo Probit Multinomial

Cuando la situación que se va a analizar corresponde a un modelo multinomial y se asigna una distribución Normal al vector aleatorio ε_i se hablará del modelo probit multinomial.

No existe una expresión funcional para las probabilidades de respuesta y es necesario realizar una integral múltiple para calcularlas. Con el planteamiento de la maximización de la utilidad, se reduce la dimensión de la integración en un grado.

Supóngase que el vector aleatorio $\varepsilon_i = (\varepsilon_{i1},.....,\varepsilon_{iJ})'$ de la utilidad sigue una distribución conjuntamente Normal con vector de medias $\vec{0}$ y matriz de varianzas-covarianzas Ω_i

$$\varepsilon_i \xrightarrow{d} N[\vec{0}, \Omega_i]$$

La probabilidad de que el individuo elija una alternativa j se obtiene como:

$$P_{ij} = \int_{-\infty}^{+\infty}\int_{-\infty}^{V_{ij}-V_{i1}+\varepsilon_{ij}} \cdots \int_{-\infty}^{V_{ij}-V_{iJ}+\varepsilon_{ij}} \frac{1}{(2\pi)^{J/2}|\Omega_i|^{1/2}} e^{-\frac{1}{2}\vec{t}\,\Omega_i^{-1}\vec{t}'}\, d\vec{t} \qquad (3\text{-}7)$$

El vector J-1 dimensional ε_{i-j} antes definido también seguirá una distribución conjunta Normal, ya que se obtiene como combinación lineal del vector original, $\varepsilon_{i-j} = A\varepsilon_i$, siendo A la matriz $(J-1)\times(J-1)$ de constantes que define la

transformación. El vector de medias del vector ε_{i-j} será $E[\varepsilon_{i-j}] = A\,E[\varepsilon_i] = \vec{0}$ y la matriz de varianzas-covarianzas vendrá dada por $\Omega_{i-j} = A\,\Omega_i\,A'$.

Las probabilidades de elección calculadas ahora sobre la distribución de probabilidad del vector aleatorio ε_{i-j}, $P_{ij} = F_{\varepsilon_{i-j}}(V_{i-j})$, necesitan una integral con un grado menor que si se calculan sobre la distribución del vector aleatorio ε_i.

Aunque la integración se ha reducido en un grado, es necesario realizar J-1 integrales para calcular la probabilidad de respuesta. Este es un problema computacional que en situaciones prácticas se resuelve mediante algún método de simulación como el método de Monte-Carlo o mediante aproximaciones como el método de Clark.

Métodos de simulación

En la literatura (Daganzo, 1979) se han propuesto tres aproximaciones diferentes para el cálculo de las probabilidades de elección de un modelo probit multinomial.

1. Integración numérica

2. Simulación Monte Carlo

3. Aproximación numérica

Cada una de ellas presenta ciertas ventajas y desvantajas.

1. Integración numérica

Con este método se consigue reducir en un grado la integración necesaria para calcular las probabilidades del modelo probit multinomial. Por ello será muy adecuado en situaciones con tres alternativas de elección, ya que sólo será necesario calcular una integral. Para problemas con más de tres alternativas tendrá menor interés.

A continuación se presenta este método para una situación general de J alternativas de elección.

La probabilidad de elección en un modelo probit multinomial viene dada por (3-7):

$$P_{ij} = \int_{-\infty}^{+\infty}\int_{-\infty}^{V_{ij}-V_{i1}+\varepsilon_{ij}} \cdots \int_{-\infty}^{V_{ij}-V_{iJ}+\varepsilon_{ij}} \frac{1}{(2\pi)^{J/2}|\Omega_i|^{1/2}} e^{-\frac{1}{2}\vec{t}\,\Omega_i^{-1}\vec{t}'} d\vec{t}$$

Considerando la factorización de Cholesky que asegura que si una matriz es definida positiva y simétrica, su inversa puede expresarse como el producto de tres matrices, se tiene que:

$$\Omega_{i-j}^{-1} = L D L'$$

donde L es una matriz cuadrada triangular inferior de orden $(J-1)\times(J-1)$ con los elementos de la diagonal iguales a la unidad y D es una matriz diagonal positiva.

A partir de esta descomposición si se considera $M = L\sqrt{D}$ se puede escribir:

$$\Omega_{i-j}^{-1} = M M'$$

donde M será por construcción una matriz triangular inferior con los elementos de la diagonal positivos.

Realizando el cambio $w_{i-j} = \varepsilon_{i-j} M$ en la integral que aparece en la expresión de las probabilidades de elección se tendrá que se reduce en un grado la integración.

La función de densidad del vector ε_{i-j} adopta la expresión siguiente:

$$f_{\varepsilon_{i-j}}\left(\varepsilon_{i,1-j},.....,\varepsilon_{i,J-j}\right) = \left[(2\pi)^{J-1}\left|\Omega_{i-j}\right|\right]^{-1/2} exp\left[-\frac{1}{2}\varepsilon_{i-j}\,\Omega_{i-j}^{-1}\,\varepsilon'_{i-j}\right] =$$

$$(2\pi)^{-(J-1)/2}\,|M|\,exp\left[-\frac{1}{2}w_{i-j}\,w'_{i-j}\right] = \prod_{\substack{k=1\\k\neq j}}^{J}\phi(w_{ik})$$

siendo $w_{i-j} = \left(w_{i1},.....,w_{ij-1},w_{ij+1},.....,w_{iJ}\right)$.

Los límites de integración se transforman con el cambio de variable realizado. Como la matriz M es diagonal inferior, también lo será M^{-1} de donde se obtendrá que cada componente $\varepsilon_{ik} - \varepsilon_{ij}$ del vector ε_{i-j} dependerá de la correspondiente componente w_{ik} y de todas las siguientes, $(w_{ik+1},....,w_{iJ})$, por lo que los límites de integración inferior y superior serán respectivamente $-\infty$ y $W_{ik}(w_{ik+1},....,w_{iJ})$ siendo

$$W_{ik}(w_{ik+1},....,w_{iJ}) = -\left(V_{ik} - V_{ij} + \sum_{r=k+1}^{J-1} w_{ir}\, m_{kr}^{-1}\right)\frac{1}{m_{kk}^{-1}}$$

la transformación correspondiente, y $\dfrac{d}{d\varepsilon_{i-j}} w_{i-j} = |M| = |\Omega_{i-j}|^{-1/2}$.

Así las probabilidades de elección se calcularán como:

$$P_{ij} = \int_{-\infty}^{W_{iJ}} \int_{-\infty}^{W_{iJ-1}(w_{iJ})} \cdots \int_{-\infty}^{W_{i2}(w_{i3},....,w_{iJ})} \int_{-\infty}^{W_{i1}(w_{i2},...w_{iJ})} \prod_{\substack{k=1 \\ k\neq j}}^{J} \phi(w_{ik})\, dw_{iJ} \cdots dw_{i1} =$$

$$\int_{-\infty}^{W_{iJ}} \cdots \int_{-\infty}^{W_{i2}(w_{i3},....,w_{iJ})} \prod_{\substack{k=2 \\ k\neq j}}^{J} \phi(w_{ik}) \left[\int_{-\infty}^{W_{i1}(w_{i2},....,w_{iJ})} \phi(w_{i1})\, dw_{i1}\right] dw_{i2} \cdots dw_{iJ} =$$

$$\int_{-\infty}^{W_{iJ}} \cdots \int_{-\infty}^{W_{i2}(w_{i3},....,w_{iJ})} \Phi(W_{i1}(w_{i2},....,w_{iJ})) \prod_{\substack{k=2 \\ k\neq j}}^{J} \phi(w_{ik})\, dw_{i2} \cdots dw_{iJ}$$

2.Método de Simulación de Monte Carlo

Este método fue sugerido por Lerman y Manski (Daganzo, 1979). Su idea es simular el valor de la utilidad para todas las alternativas muchas veces y la probabilidad de elegir una alternativa j será la proporción de veces que se haya encontrado que su utilidad es la mayor de todas.

Para simular las utilidades de las alternativas se asume la relación $V_{ij} = x_{ij}'\beta, \quad j = 1,....,J$, y se utilizan las estimaciones máximo-verosímiles de los parámetros del modelo $(\hat{\beta}, \hat{\Omega})$ para generar una observación del vector aleatorio $\hat{U}_i = x_i'\hat{\beta} + \varepsilon_i$ que sigue una distribución Normal J-dimensional con

vector de medias $\hat{V}_i$ y matriz de varianzas-covariazas $\hat{\Omega}_i$. Cualquier programa de ordenador que genere números aleatorios puede servir a este propósito.

Una vez se observe el vector $\left(\hat{U}_{i1},\hat{U}_{i2},.....,\hat{U}_{iJ}\right)$ se anota la alternativa que presente el mayor valor de la utilidad.

Se repite el proceso un número de veces y la probabilidad de elegir la alternativa j será la proporción de veces que esta alternativa ha presentado el mayor valor de la utilidad.

El principal inconveniente de este método es que si hay alternativas cuya probabilidad de elección sea muy pequeña será necesario realizar la simulación de Monte Carlo un número muy elevado de veces para conseguir encontrar su valor. Además no hay ningún criterio específico que indique cuál es el número de veces que debe realizarse el experimento.

3.Aproximación numérica

Este método de obtención de las probabilidades de elección de un modelo probit multinomial mediante una aproximación numérica se basa en la idea de Clark (1961) quien demostró que la variable aleatoria definida como el máximo de variables aleatorias conjuntamente Normales sigue una distribución aproximadamente Normal.

Con este método se consigue calcular las probabilidades de elección del modelo probit multinomial con una única integral.

Utilizando las variables ε_{i-j}, las probabilidades de elección se pueden calcular como:

$$P_{ij} = P\left(U_{ik} - U_{ij} \leq 0 \;\; k \neq j \,, k = 1,2,...,J \,/\, x_i\right) =$$

$$P\left(\max_{\substack{k=1,.....J \\ k \neq j}} \left\{U_{ik} - U_{ij}\right\} \leq 0 \,/\, x_i \right) = P\left(\xi_j \leq 0 \,/\, x_i\right) = F_{\xi_j}(0)$$

siendo $F_{\xi_i}(\cdot)$ la función de distribución de la variable aleatoria unidimensional

$$\xi_j = \max_{\substack{k=1,.....J \\ k \neq j}} \left\{U_{ik} - U_{ij}\right\}$$

Puesto que ε_i sigue una distribución conjuntamente Normal se tiene que $(U_{i1},.....,U_{iJ})'$ también sigue una distribución conjuntamente Normal, con vector de medias $V_i = (V_{i1},.....,V_{iJ})'$ y matriz de varianzas-covarianzas Ω_i.

Del mismo modo el vector de las diferencias entre las utilidades, U_{i-j} sigue una distribución Normal J-1 dimensional con vector de medias dado por V_{i-j} y matriz de varianzas-covarianzas Ω_{i-j}.

Así, la variable ξ_j está definida como el máximo de J-1 variables con distribución conjunta Normal, y Clark (1961) sugiere aproximar la distribución de esta variable por la de una Normal. La media y la varianza de esta variable ξ_j, se obtienen mediante unas fórmulas propuestas por Clark.

Para clarificar este método de aproximación, se va a presentar la solución en un problema con tres alternativas.

Dadas tres variables conjuntamente Normales $(Y_1, Y_2, Y_3)'$, Clark calculó el primer y segundo momento de $\tilde{Y}_2 = max\{Y_1, Y_2\}$ y la covarianza entre $\tilde{Y}_2$ e Y_3 en función del vector de medias original $\vec{\mu} = (\mu_1, \mu_2, \mu_3)'$ y su matriz de varianzas-covarianzas $\Sigma_Y = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$.

Las fórmulas correspondientes son:

$$\mu_{\tilde{2}} = \mu_2 + (\mu_1 - \mu_2)\,\Phi(\alpha) + a\,\phi(\alpha)$$

$$\sigma_{\tilde{2}}^2 = \mu_2^2 + \sigma_2^2 + (\mu_1^2 + \sigma_1^2 - \mu_2^2 - \sigma_2^2)\,\Phi(\alpha) + (\mu_1 + \mu_2)\,a\,\phi(\alpha) - \mu_{\tilde{2}}^2$$

$$\sigma_{\tilde{2}3} = \sigma_{23} + (\sigma_{13} - \sigma_{23})\,\Phi(\alpha)$$

siendo $a = (\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})^{1/2}$ y $\alpha = \dfrac{\mu_1 - \mu_2}{a}$

Asumiendo que $\left(\tilde{Y}_2 , Y_3\right)'$ sigue una distribución Normal bivariante con vector de medias $\tilde{\mu} = \left(\mu_{\tilde{2}} , \mu_3\right)'$ y matriz de varianzas-covarianzas $\Sigma_{\tilde{Y}_2} = \begin{bmatrix} \sigma^2_{\tilde{2}} & \sigma_{\tilde{2}3} \\ \sigma_{\tilde{2}3} & \sigma^2_3 \end{bmatrix}$ se puede repetir el proceso para calcular la media y varianza de la variable definida ahora por

$$\tilde{Y}_3 = max\left\{\tilde{Y}_2 , Y_3\right\} = max\left\{Y_1 , Y_2 , Y_3\right\}$$

Este proceso se aplica de forma recursiva para la situación de las J-1 variables $U_{i.k-j} = U_{ik} - U_{ij}$ anteriores y se obtendrá que $\xi_j = \underset{\substack{k=1,\ldots,J \\ k\neq j}}{max}\left\{U_{ik} - U_{ij}\right\}$ sigue una distribución Normal cuya media y varianza aproximadas vendrán dadas por las fórmulas recursivas de Clark, que en el paso $r-1$ adoptan la expresión:

$$\mu_{\tilde{r}} = \mu_r + \left(\mu_{\tilde{r}-1} - \mu_r\right)\Phi(\alpha_r) + a_r\,\phi(\alpha_r)$$

$$\sigma^2_{\tilde{r}} = \mu^2_r + \sigma^2_r + \left(\mu^2_{\tilde{r}-1} + \sigma^2_{\tilde{r}-1} - \mu^2_r - \sigma^2_r\right)\Phi(\alpha_r) + \left(\mu_{\tilde{r}-1} + \mu_r\right)a_r\,\phi(\alpha_r) - \mu^2_{\tilde{r}}$$

$$\sigma_{\tilde{r}s} = \sigma_{rs} + \left(\sigma_{\tilde{r}-1,s} - \sigma_{rs}\right)\Phi(\alpha_r) \quad , \quad r \neq s$$

siendo $a_r = \left(\sigma^2_{\tilde{r}-1} + \sigma^2_r - 2\sigma_{\tilde{r}-1,r}\right)^{1/2}$ y $\alpha_r = \dfrac{\mu_{\tilde{r}-1} - \mu_r}{a_r}$

En total se necesita aplicar las fórmulas de Clark J-2 veces, lo que lleva a calcular J-1 veces las funciones $\phi(\cdot)$ y $\Phi(\cdot)$, 2(J-1) medias y $\dfrac{J(J-1)}{2}$ varianzas y covarianzas. Este es el principal inconveniente de este método, hay que realizar muchos cálculos, aunque son sencillos.

Modelo Probit Independiente

El problema de cálculo de las probabilidades de elección en un modelo probit multinomial aparece por el hecho de que entre las componentes del vector aleatorio ε_i no hay ninguna restricción en cuanto a la dependencia entre ellas. El problema computacional desaparece si se introduce la restricción de que todas las componentes del vector aleatorio ε_i son independientes entre sí. De esta forma la

integral múltiple que hay que resolver pasaría a calcularse como el producto de integrales simples. Se tendría en este caso que

$$P_{ij} = F_{\varepsilon_{i-j}}(V_{i-j}) = \prod_{\substack{r=1 \\ r \neq j}}^{J} F_{\varepsilon_{i,r-j}}(V_{ij} - V_{ir}) \qquad (3\text{-}8)$$

donde cada $F_{\varepsilon_{i,r-j}}(\cdot)$ es la función de distribución de la variable $\varepsilon_{i,r-j} = \varepsilon_{ir} - \varepsilon_{ij}$ con distribución Normal univariante de media 0 y varianza $\sigma_{ir}^2 + \sigma_{ij}^2$.

Bajo la hipótesis de independencia, el modelo probit multinomial recibe el nombre de probit independiente.

La ventaja del modelo probit independiente es la facilidad de cálculo de las probabilidades, sin embargo en muchas situaciones no será adecuado admitir la independencia entre estas variables.

La independencia anterior indicará que cada una de las componentes del vector ε_i recoge únicamente las características no observables correspondientes a la alternativa a la que representan y no puede tener en cuenta una situación en la que dos o más alternativas presentan rasgos comunes y tengan una cierta similaridad.

Por ejemplo en la elección entre adquirir una segunda vivienda en la montaña, en la playa de una localidad o en la playa de otra localidad este modelo no resulta adecuado, puesto que las dos últimas alternativas están relacionadas. En esta situación debería utilizarse un modelo probit multinomial.

3.1.4. Modelo Logit

Este modelo está asociado a la distribución de probabilidad del valor extremo Tipo I o Weibull. Se asume dicha distribución de probabilidad para las variables aleatorias ε_i y que todas son independientes entre sí.

Modelo Logit Binomial

En el caso binomial, el modelo se denomina logit binomial y la distribución utilizada lleva a la siguiente expresión para la probabilidad de respuesta:

$$P(y_i = 1 / x_i) = F(x_i'\beta) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \qquad (3\text{-}9)$$

Esta función de distribución univariante, que recibe el nombre de Logística presenta una serie de propiedades similares a las de la distribución Normal.

El modelo logit binomial se define a partir de la distribución Logística de media cero, $\mu = 0$ y varianza $\sigma^2 = \pi^2/3$. Estos valores para la media y la varianza no son restrictivos, ya que si en un problema concreto se cree más adecuado considerar un valor diferente de cero para la media, basta añadir un término constante en $x_i'\beta$.

Cuando se desee una varianza $\sigma^2 \neq \pi^2/3$ se modifica el modelo y la probabilidad de respuesta se calculará como:

$$P(y_i = 1 / x_i) = \frac{e^{\frac{x_i'\beta}{\sigma}\frac{\pi}{\sqrt{3}}}}{1 + e^{\frac{x_i'\beta}{\sigma}\frac{\pi}{\sqrt{3}}}}$$

La semejanza entre la distribución Normal univariante y la distribución Logística implica que las probabilidades de respuesta calculadas con un modelo probit binomial y un modelo logit binomial son muy similares.

No obstante, la distribución Logística posee mayor masa probabilística en las colas que la distribución Normal, lo que llevará a resultados diferentes cuando en un modelo se tengan muchas observaciones en las colas.

Esta semejanza en estos dos modelos binomiales desaparece al pasar a modelos multinomiales. Cuando hay más de dos alternativas de elección, el modelo probit multinomial lleva a resultados muy diferentes de los obtenidos con el modelo logit multinomial.

Modelo Logit Multinomial

Cuando hay más de dos alternativas, por ejemplo en la elección entre apartamento, chalet, bungalov o casa como vivienda secundaria, el modelo se

denomina logit multinomial y las probabilidades de respuesta vienen dadas por la expresión:

$$P_{ij} = \frac{e^{V_{ij}}}{\sum_{r=1}^{J} e^{V_{ir}}} \qquad (3\text{-}10)$$

El modelo logit multinomial se basa en la independencia entre las componentes aleatorias ε_{ij}. Esta hipótesis lleva a que la parte determinista de la utilidad correspondiente a la alternativa j, V_{ij}, no dependa de otras alternativas diferentes a la alternativa j. Este hecho hace que, al igual que el modelo probit independiente, el modelo logit multinomial sea poco adecuado en situaciones en las que existen ciertas relaciones entre las alternativas de elección.

Es fácil comprobar que esta hipótesis de independencia implica que en el modelo logit multinomial la razón de las probabilidades de elección de dos alternativas j y k depende únicamente de las características asociadas a las mismas:

$$\frac{P_{ij}}{P_{ik}} = \frac{e^{V_{ij}} \Big/ \sum_{r=1}^{J} e^{V_{ir}}}{e^{V_{ik}} \Big/ \sum_{r=1}^{J} e^{V_{ir}}} = \frac{e^{V_{ij}}}{e^{V_{ik}}} = e^{V_{ij} - V_{ik}}$$

La propiedad anterior se conoce en el contexto de los modelos de respuesta cualitativa con el nombre de Independencia de Alternativas Irrelevantes (IIA) y puede enunciarse alternativamente del siguiente modo.

Sea $A = \{1,2,.....,M\}$ un subconjunto de alternativas de elección y $P(A) = \sum_{r \in A} P_{ir}$ su probabilidad. Cuando se verifique que la probabilidad de que un individuo i elija la alternativa j del subconjunto A, $P_{ij/A}$, no depende de las alternativas que no se encuentren en el subconjunto A se verifica la propiedad IIA, ya que $P_{ij/A}$ sólo dependerá de las características asociadas al subconjunto A.

Aunque la propiedad IIA sea algo restrictiva, ofrece una atractiva ventaja de cálculo, ya que añadir o eliminar una alternativa, o varias, del conjunto de elección únicamente implica añadir o eliminar tantos términos de la forma $e^{V_{ir}}$ en

el denominador de la probabilidad de respuesta como alternativas se estén considerando, facilitando de esta forma los cálculos de las probabilidades en un conjunto de alternativas extendido o restringido.

Es bastante usual admitir una especificación lineal para la parte determinista de la función de utilidad, $V_{ij} = x'_{ij}\beta$. En algunas situaciones no existe ninguna característica específica de alternativas, es decir, el vector de características x'_{ij} está formado únicamente por las r características del individuo, que toman el mismo valor para todas las alternativas. En este caso, el modelo logit multinomial se planteará de forma equivalente como:

$$P_{ij} = \frac{e^{x_i'\beta_j}}{\sum_{k=1}^{J} e^{x_i'\beta_k}}$$

Otro planteamiento alternativo para encontrar las probabilidades de respuesta del modelo logit multinomial que no utiliza el modelo de variable latente ni la teoría de la maximización de la utilidad, fue seguido por Luce (1959) y su idea fue obtener estas probabilidades de respuesta a partir de la imposición de ciertas propiedades y restricciones sobre las mismas y que fueran coherentes con su planteamiento.

Luce exigió que sus probabilidades de respuesta verificaran que la razón entre las probabilidades de dos alternativas cualquiera no dependiera de una tercera alternativa. Este axioma propuesto es el que ahora se conoce como la Independencia de Alternativas Irrelevantes, mencionada anteriormente.

Su axioma puede expresarse formalmente de la siguiente manera. Sean dos alternativas j, k del conjunto de elección C, con probabilidades de ser elegidas $P_{ij} = P(j / x_i, C)$ y $P_{ik} = P(k / x_i, C)$ respectivamente. Estas alternativas deben verificar que:

$$\frac{P(j / x_i, C)}{P(k / x_i, C)} = \frac{P(j / x_i, \{j, k\})}{P(k / x_i, \{j, k\})} = \frac{P_{jk}}{P_{kj}}$$

donde

$$P_{jk} = P(j / x_i, \{j, k\}) \quad \text{y} \quad P_{kj} = P(k / x_i, \{j, k\})$$

Despejando se tiene:

$$P(k / x_i, C) = \frac{P_{kj}}{P_{jk}} P(j / x_i, C)$$

Se ha asumido que $P(k / x_i, C) > 0$ y por lo tanto $P_{kj} > 0$. Esto no supone ninguna pérdida de generalidad, ya que una probabilidad muy pequeña es indistinguible de una probabilidad nula.

Si en la relación anterior se suma para todas las alternativas $k \in C$ se obtiene que

$$P(j / x_i, C) = \frac{1}{\sum_{k \in C} \frac{P_{kj}}{P_{jk}}}$$

y puesto que dadas tres alternativas $j,k,r \in C$ la razón de probabilidades de las alternativas j y k se puede expresar como:

$$\frac{P_{kj}}{P_{jk}} = \frac{P(k / x_i, C)}{P(j / x_i, C)} = \frac{\frac{P_{kr}}{P_{rk}} P(r / x_i, C)}{\frac{P_{jr}}{P_{rj}} P(r / x_i, C)} = \frac{P_{kr} / P_{rk}}{P_{jr} / P_{rj}}$$

la probabilidad de la alternativa j se puede reescribir del siguiente modo:

$$P(j / x_i, C) = \frac{1}{\sum_{k \in C} \frac{P_{kr} / P_{rk}}{P_{jr} / P_{rj}}} = \frac{P_{jr} / P_{rj}}{\sum_{k \in C} P_{kr} / P_{rk}}$$

Denotando por $V(x_i, j, k)$ a la transformación $\ln\left(P_{jk} / P_{kj}\right)$ y admitiendo la descomposición $V(x_i, j, r) = v(x_i, j) - v(x_i, r)$ se obtendrá la siguiente expresión para las probabilidades de elección:

$$P(j / x_i, C) = \frac{e^{v(x_i,j)-v(x_i,r)}}{\sum_{k=1}^{J} e^{v(x_i,k)-v(x_i,r)}} = \frac{e^{v(x_i,j)}}{\sum_{k=1}^{J} e^{v(x_i,k)}}$$

que utilizando la relación $V_{ij} = v(x_i, j)$ coincide con la expresión (3-10) de las probabilidades de elección del modelo logit multinomial.

El razonamiento seguido por Luce llevó a la obtención del modelo logit multinomial, sin embargo el método seguido para la construcción de este modelo no garantiza su compatibilidad con la teoría de la maximización de la utilidad.

No obstante, McFadden (1981) proporciona un procedimiento para comprobar si un conjunto de probabilidades de elección es coherente o no con el planteamiento de la maximización de la utilidad.

La forma de comprobar si las probabilidades P_{ij} son compatibles con la maximización de la utilidad es mediante el Teorema de Williams, Daly y Zachary.

Este teorema asegura la existencia de una función de densidad que, asociada a las variables aleatorias ε_i de la utilidad, proporciona las probabilidades P_{ij} como un modelo de maximización de la utilidad. Además de las condiciones usuales de ser no negativas y que la suma total sea igual a la unidad, las probabilidades P_{ij} deben cumplir las siguientes condiciones para aplicar el teorema:

1. Las probabilidades son invariantes por traslación:

$$P_{ij} = P(y_{ij} = 1, y_{ir} = 0 \quad , \forall r \neq j / x_i) = g_j(V_{ij}) = g_j(V_{ij} + \alpha)$$

2. Las derivadas de las probabilidades cumplen:

$$\frac{d}{dV_{ik}} P_{ij} = \frac{d}{dV_{ij}} P_{ik}$$

Esta condición garantiza la integrabilidad de P_{ij}.

3. La derivada de orden J-1 de las probabilidades es no negativa:

$$\frac{d^{(J-1)}}{d\mathrm{V}_{i1}\cdots d\mathrm{V}_{ij-1}\,d\mathrm{V}_{ij+1}\cdots d\mathrm{V}_{iJ}}\mathrm{P}_{ij} \geq 0$$

Esta condición asegura que la función de densidad es propia.

Modelo Logit Universal

Los modelos probit y logit multinomial se han obtenido al asignar una distribución concreta a las variables aleatorias ε_i.

La utilización de una distribución de probabilidad u otra diferente lleva a la obtención de diferentes modelos de respuesta cualitativa. Es el propio investigador el que debe decidir el modelo más adecuado a su problema en función de sus objetivos de análisis y sus datos.

No obstante puede realizarse un planteamiento alternativo para elegir el modelo. Se desarrollará en primer lugar la situación con dos alternativas, es decir, para los modelos binomiales y después se verá la generalización a los modelos multinomiales.

En un modelo de respuesta cualitativa binomial la probabilidad de respuesta viene dada a partir de la función de distribución $\mathrm{F}(\cdot)$:

$$\mathrm{P}(y_i = 1/x_i) = \mathrm{F}(x_i'\beta)$$

El Teorema de Aproximación de Weierstrass garantiza la existencia de una aproximación lineal a cualquier función, en particular garantiza que la transformación

$$g(x_i'\beta) = \ln\frac{\mathrm{F}(x_i'\beta)}{1-\mathrm{F}(x_i'\beta)}$$

puede ser aproximada para cualquier nivel de precisión deseado por una función lineal $z_i'\gamma$.

En consecuencia se puede considerar que:

$$g(x_i'\beta) = \ln\frac{F(x_i'\beta)}{1-F(x_i'\beta)} = \ln\frac{P(y_i=1/x_i)}{1-P(y_i=1/x_i)} = z_i'\gamma$$

Operando en esta igualdad se obtiene:

$$e^{z_i'\gamma} = \frac{P(y_i=1/x_i)}{1-P(y_i=1/x_i)} \Rightarrow e^{z_i'\gamma}\left[1-P(y_i=1/x_i)\right] = P(y_i=1/x_i) \Rightarrow$$

$$P(y_i=1/x_i) = \frac{e^{z_i'\gamma}}{1+e^{z_i'\gamma}}$$

Es decir, las probabilidades de respuesta para una situación de elección tienen la forma de un modelo logit, pero utilizando como variable latente no la original $y_i^* = x_i'\beta - \varepsilon_i$, sino un modelo obtenido como una transformación de las características observadas, $z_i'\gamma$.

De esta forma, la elección de una determinada función de distribución para modelizar la variable respuesta no es un problema para el investigador, el problema se traslada a encontrar la transformación lineal de las características, $z_i'\gamma$, adecuada.

Si el modelo que debe analizarse es un modelo multinomial, el razonamiento seguido es el mismo. Sea $z_{ij}'\gamma$ una aproximación lineal a la transformación $\ln P_{ij}$, siendo P_{ij} las probabilidades de respuesta del modelo obtenidas con una función de distribución $F(\cdot)$.

A partir de la relación $\ln P_{ij} \approx z_{ij}'\gamma$ se puede obtener la siguiente expresión para las probabilidades:

$$\frac{e^{z_{ij}'\gamma}}{\sum_{r=1}^{J} e^{z_{ir}'\gamma}} = \frac{e^{\ln P_{ij}}}{\sum_{r=1}^{J} e^{\ln P_{ir}}} = \frac{P_{ij}}{\sum_{r=1}^{J} P_{ir}} = P_{ij}$$

Es decir, P_{ij} admite una expresión como la de un modelo logit. De nuevo el problema de encontrar la distribución de probabilidad adecuada se convierte en el problema de encontrar la transformación lineal $z_{ij}'\gamma$ adecuada.

Cuando se utilizan estas aproximaciones lineales el modelo logit que aparece recibe el nombre de modelo logit universal. A pesar de la semejanza en la

expresión funcional con el modelo logit multinomial, el modelo logit universal no verifica la propiedad IIA ya que en la aproximación lineal $z'_{ij}\gamma$ aparecen características de todas las alternativas.

No obstante el modelo logit universal no es muy utilizado en la práctica, ya que encontrar la transformación lineal $z'_{ij}\gamma$ adecuada al problema no es una tarea fácil.

3.1.5. Modelo del Valor Extremo Generalizado

En un modelo de respuesta cualitativa es muy conveniente la facilidad de cálculo de las probabilidades de elección P_{ij}, pero es necesario encontrar una relación capaz de reflejar las propias características del problema analizado.

En cuanto a la facilidad computacional, el modelo logit multinomial sería el más adecuado, pero la propiedad de la Independencia de Alternativas Irrelevantes lo hace inapropiado en muchas situaciones.

Un ejemplo para ilustrar este inconveniente del modelo logit multinomial es la elección de un local comercial para abrir un negocio. Supóngase que las alternativas que tiene un individuo son alquilar un local en un centro comercial o alquilarlo en una calle céntrica de la ciudad. Admitiendo la modelización de esta situación de elección mediante un modelo logit, si se incorpora una tercera alternativa la razón entre las probabilidades de alquilar en el centro comercial o en la zona céntrica de la ciudad seguirá manteniéndose igual, tanto si la tercera alternativa es alquilar en un nuevo centro comercial que se va a abrir en esa ciudad como si es alquilar un local en otra finca de la misma calle céntrica de la ciudad.

En el caso de que la tercera alternativa sea otro local en el centro de la ciudad, no parece lógico que se tenga que mantener la misma relación entre las probabilidades de alquilar en el centro comercial y alquilar en el centro de la ciudad que si esta tercera alternativa es alquilar el local en un nuevo centro comercial.

Esta circunstancia lleva a plantear la necesidad de encontrar modelos que contemplen la existencia de relaciones o similitudes entre las alternativas de elección.

El modelo probit multinomial podría ser una solución, ya que en su definición se permite que exista una covarianza diferente de cero entre las diferentes variables aleatorias ε_{ij}, pudiéndose reflejar esta similitud entre las alternativas.

En la práctica ésta no es la solución más adoptada ya que el modelo probit multinomial es muy complicado computacionalmente y deben utilizarse procedimientos de simulación o de aproximación para calcular sus probabilidades. Por el contrario, en el caso de utilizar el modelo probit independiente que es más fácil de cálculo, nos encontramos con un modelo que presenta una propiedad parecida a la IIA, con lo que también resultará inadecuado.

A continuación se desarrollará una familia de modelos de respuesta cualitativa que permitirán admitir la dependencia entre las alternativas del conjunto de elección y que McFadden (1978) obtuvo como una generalización del modelo logit multinomial, a través de la teoría de la maximización de la utilidad.

La propuesta es admitir que el vector aleatorio ε_i sigue una distribución del Valor Extremo Generalizado, cuya función de distribución tiene la siguiente expresión:

$$F(t_1,t_2,.....,t_J) = exp\left[-G\left(e^{-t_1},e^{-t_2},.....,e^{-t_J}\right)\right]$$

donde la función G debe de verificar las siguientes condiciones:

1. $G(s_1,.....,s_J)$ es no negativa y homogénea de grado 1 para cualquier vector $(s_1,.....,s_J) \geq 0$

2. $\lim_{s_j \to \infty} G(s_1,.....,s_J) = +\infty \qquad j = 1,2,.....,J$

3. Las derivadas parciales mixtas de *G* existen y son continuas, con derivadas parciales mixtas de orden impar no negativas y de orden par no positivas.

Especificaciones diferentes para la función *G* darán lugar a modelos diferentes que reciben el nombre genérico de modelos GEV, y cuyas probabilidades de elección, calculadas a partir de la maximización de la utilidad, adoptarán la siguiente forma funcional general:

$$P_{ij} = P\left(V_{ir} + \varepsilon_{ir} \leq V_{ij} + \varepsilon_{ij}\ ,\ \forall r \neq j\ , r \in C / x_i\right) = \frac{e^{V_{ij}}\ G_j\left(e^{V_{i1}},.....,e^{V_{iJ}}\right)}{G\left(e^{V_{i1}},.....,e^{V_{iJ}}\right)} \tag{3-11}$$

siendo $G_j\left(e^{V_{i1}},.....,e^{V_{iJ}}\right)$ la derivada de la función $G\left(e^{V_{i1}},.....,e^{V_{iJ}}\right)$ con respecto a la componente j-ésima.

La ventaja de los modelos GEV es su gran flexibilidad para describir una amplia variedad de situaciones de elección. Bastará considerar para cada situación la función $G(\cdot)$ adecuada.

Por ejemplo, si la función $G\left(s_1, s_2,.....,s_J\right)$ tiene una forma aditiva y considera por igual a todas las componentes, sin ponderar más una que otra, $G\left(s_1, s_2,.....,s_J\right) = \sum_{r=1}^{J} s_r$, se verifica que $G_j\left(s_1, s_2,.....,s_J\right) = 1$ y por lo tanto las probabilidades de elección tendrán la expresión

$$P_{ij} = \frac{e^{V_{ij}}}{\sum_{r=1}^{J} e^{V_{ir}}}$$

que corresponde al logit multinomial (3-10).

Modelo Logit Multinomial Anidado

Supóngase ahora la situación de la elección del alquiler del local comercial en la que las alternativas disponibles son alquilar en el centro comercial, alquilar en una finca céntrica de la ciudad o alquilar en otra finca de la misma zona céntrica. El modelo logit multinomial no era adecuado por la similitud entre las dos últimas alternativas. Puede considerarse ahora una función $G(\cdot)$ que trate de agrupar en un mismo tratamiento a las alternativas de alquiler en el centro de la ciudad y de forma separada considere a la alternativa de alquilar en el centro comercial. Una posibilidad es considerar que la función $G(\cdot)$ tiene la forma siguiente

$$G\left(e^{V_{i1}}, e^{V_{i2}}, e^{V_{i3}}\right) = e^{V_{i1}} + \left(e^{V_{i2}/1-\sigma} + e^{V_{i3}/1-\sigma}\right)^{1-\sigma}$$

siendo la alternativa 1 la correspondiente a alquilar en el centro comercial y la 2 y la 3 son el alquiler en las dos fincas del centro de la ciudad, respectivamente. Con el coeficiente σ, se pretende dar una medida de la similitud entre las dos últimas alternativas.

Cuando se tenga una situación como ésta, en la que hay ciertos parecidos o relaciones entre algunas alternativas, el modelo GEV puede considerar grupos de alternativas similares, con un coeficiente de similaridad σ_k, de forma que la función $G(\cdot)$ venga definida como:

$$G\left(e^{V_{i1}}, e^{V_{i2}}, \ldots\ldots, e^{V_{iJ}}\right) = \sum_{k=1}^{m} \left[\sum_{r=1}^{J_k} e^{V_{ir}^k / 1-\sigma_k} \right]^{1-\sigma_k}$$

siendo m el total de grupos de alternativas, J_k el número de alternativas de cada grupo k, y con el superíndice k, V_{ij}^k, se denota el grupo k al que pertenece la alternativa r.

En este caso

$$G_j\left(e^{V_{i1}}, e^{V_{i2}}, \ldots\ldots, e^{V_{iJ}}\right) = e^{V_{ij}^k \frac{\sigma_k}{1-\sigma_k}} \left[\sum_{r=1}^{J_k} e^{V_{ir}^k / 1-\sigma_k} \right]^{-\sigma_k}$$

y las probabilidades de elección serán:

$$P_{ij} = \frac{e^{V_{ij}^k \frac{1}{1-\sigma_k}} \left[\sum_{r=1}^{J_k} e^{V_{ir}^k / 1-\sigma_k} \right]^{-\sigma_k}}{\sum_{s=1}^{m} \left[\sum_{r=1}^{J_s} e^{V_{ir}^s / 1-\sigma_s} \right]^{1-\sigma_s}} \qquad (3\text{-}12)$$

En este último modelo se ha considerado un grado σ_k de similaridad entre las alternativas del grupo k y todos los grupos se han ponderado de la misma forma.

Este modelo recibe el nombre de modelo logit multinomial anidado y McFadden llegó al mismo utilizando un razonamiento alternativo al presentado aquí y que sería comparable al seguido por Luce para la obtención del modelo logit multinomial.

A continuación se expone el planteamiento de McFadden para obtener el modelo logit multinomial anidado.

Cuando entre las alternativas de elección exista algún tipo de relación es fácil considerar grupos de alternativas, de forma que las alternativas más semejantes se encuentren en un grupo y los grupos se diferencien entre sí. En el ejemplo de la elección del local a alquilar, un grupo lo formaría la alternativa alquilar en el centro comercial y el otro grupo contendría a las alternativas de alquiler en las dos fincas céntricas de la ciudad. De esta forma la elección puede plantearse de forma secuencial, eligiendo primero el grupo de alternativas y en segundo lugar la alternativa dentro del grupo elegido.

Con este planteamiento McFadden razonó la obtención de las probabilidades a partir de la relación entre la probabilidad de elegir una alternativa y las probabilidades de elegir un grupo y la alternativa dentro del grupo. Obsérvese que si se denota por P_i^k la probabilidad de elegir un grupo k de entre todos los que hay y por $P_{ij/k}$ la probabilidad de elegir una alternativa j dentro del grupo k, la probabilidad de una alternativa j se puede expresar como $P_{ij} = P_i^k \, P_{ij/k}$.

Buscando la flexibilidad del modelo logit multinomial, se considera que las probabilidades intermedias (probabilidad del grupo y de la alternativa dentro del grupo) tienen la forma funcional de ese modelo.

Es decir, se asume que:

$$P_{ij/k} = \frac{e^{V_{ij}^k / 1-\sigma_k}}{\sum_{r=1}^{J_k} e^{V_{ir}^k / 1-\sigma_k}}$$

siendo V_{ij}^k la parte determinista que recoge las características observables del individuo i y de las alternativas que están dentro del grupo k, σ_k es un coeficiente que representa la similaridad entre las alternativas del grupo k y J_k es el número total de alternativas que hay en el grupo k.

La probabilidad de elegir el grupo k adopta la siguiente expresión:

$$P_i^k = \frac{e^{I_k(1-\sigma_k)}}{\sum_{s=1}^{m} e^{I_s(1-\sigma_s)}}$$

siendo m el total de grupos de alternativas que hay en el problema.

El valor I_k es denominado por McFadden como valor inclusivo y Train (1986), entre otros, lo identifica como la utilidad media que el individuo i puede esperar de las alternativas del grupo k. Este valor inclusivo se define como:

$$I_k = \ln\left[\sum_{r=1}^{J_k} e^{V_{ir}^k/1-\sigma_k}\right]$$

Ambas probabilidades definidas tienen la forma funcional del logit multinomial, pero cada una está definida para un nivel diferente: grupos en un caso y alternativas en otro caso. La combinación de las dos probabilidades dará lugar a la siguiente expresión de la probabilidad de elección que coincide con la ecuación (3-12):

$$P_{ij} = P_i^k \, P_{ij/k} = \frac{e^{I_k(1-\sigma_k)}}{\sum_{s=1}^{m} e^{I_s(1-\sigma_s)}} \frac{e^{V_{ij}^k/1-\sigma_k}}{\sum_{r=1}^{J_k} e^{V_{ir}^k/1-\sigma_k}} = \frac{e^{V_{ij}^k/1-\sigma_k} \, e^{-I_k\sigma_k}}{\sum_{s=1}^{m} e^{I_s(1-\sigma_s)}} =$$

$$\frac{\left[\sum_{r=1}^{J_k} e^{V_{ir}^k/1-\sigma_k}\right]^{1-\sigma_k} e^{V_{ij}^k/1-\sigma_k}}{\sum_{s=1}^{m}\left[\sum_{r=1}^{J_k} e^{V_{ir}^k/1-\sigma_s}\right]^{1-\sigma_s} \sum_{r=1}^{J_k} e^{V_{ir}^k/1-\sigma_k}} = \frac{e^{V_{ij}^k/1-\sigma_k}\left[\sum_{r=1}^{J_k} e^{V_{ir}^k/1-\sigma_k}\right]^{-\sigma_r}}{\sum_{s=1}^{m}\left[\sum_{r=1}^{J_s} e^{V_{ir}^s/1-\sigma_s}\right]^{1-\sigma_s}}$$

para cualquier alternativa de elección $j = 1,2,.....,J_k, \quad k = 1,2,.....,m.$

Aunque únicamente se ha supuesto la existencia de un nivel de agrupamiento, alternativa dentro del grupo y el grupo, el modelo logit multinomial anidado permite considerar muchos más niveles de anidamiento.

El modelo logit multinomial anidado obtenido de esta forma no tiene garantizada la compatibilidad con la teoría de la maximización de la utilidad, ya que se ha llegado a él a través de supuestos e hipótesis sobre las alternativas, como la estructura jerárquica de elección y la forma funcional logit multinomial para las probabilidades intermedias.

Sin embargo, al igual que en el planteamiento de Luce para obtener las probabilidades de elección del modelo logit multinomial, también aquí se puede

encontrar la compatibilidad con la maximización de la utilidad desde el Teorema de Williams, Daly y Zachary.

Las probabilidades definidas con el modelo logit multinomial anidado verifican sin ningún problema las exigencias de ese teorema. La única condición algo más restrictiva es la última, que es equivalente a exigir que el coeficiente de similaridad σ_k sea una cantidad acotada en el intervalo unidad, $0 \leq \sigma_k \leq 1$, para cualquier grupo k considerado en el problema.

Si se consiguen probabilidades que verifiquen esta restricción, el teorema de Williams-Daly-Zachary asegura que existe una función de distribución F asociada a las variables ε_i de la componente aleatoria de la utilidad y que permite llegar a las expresiones anteriores de las probabilidades de elección utilizando la teoría de la maximización de la utilidad.

El modelo logit multinomial anidado, obtenido como un caso particular de un modelo GEV, permite una pequeña variación si los grupos k se ponderan de forma diferente mediante un coeficiente a_k.

Sea la siguiente función $G(\cdot)$

$$G\left(e^{V_{i1}},.....,e^{V_{iJ}}\right)=\sum_{k=1}^{m} a_k\left[\sum_{r=1}^{J_k} e^{V_{ir}^k/1-\sigma_k}\right]^{1-\sigma_k}$$

donde a_k es la ponderación o peso asociado al grupo k. Estos coeficientes $a_1,.....,a_m$ pueden ser todos diferentes entre sí, pero las condiciones que necesita la función $G(\cdot)$ exigen que sean estrictamente positivos. Por supuesto, también ahora será necesario que σ_k esté en el intervalo unidad para mantener la consistencia con la teoría de la maximización de la utilidad.

Las probabilidades de elección en este caso son:

$$P_{ij}=\frac{a_k\, e^{V_{ij}^k/1-\sigma_k}\left[\sum_{r=1}^{J_k} e^{V_{ir}^k/1-\sigma_k}\right]^{-\sigma_k}}{\sum_{s=1}^{m} a_s\left[\sum_{r=1}^{J_s} e^{V_{ir}^s/1-\sigma_s}\right]^{1-\sigma_s}} \qquad j=1,.....,J_k \qquad k=1,.....,m \quad (3\text{-}13)$$

La interpretación que dió McFadden para el logit multinomial anidado a partir de una estructura jerárquica, eligiendo primero el grupo k y en segundo lugar la alternativa j dentro de este grupo, sigue manteniéndose ahora.

Sea la descomposición de la probabilidad de elección como el producto de las dos probabilidades, $P_{ij} = P_i^k \, P_{ij/k}$.

En este caso, es fácil comprobar que la probabilidad de elegir una alternativa j dentro del grupo k es igual a la correspondiente probabilidad asociada al modelo logit multinomial anidado .

La forma de calcularlas es considerando que tanto P_i^k como $P_{ij/k}$ son de la forma GEV.

Tomando la siguiente función

$$G^k\left(e^{V_{i1}^k}, \ldots\ldots, e^{V_{iJ_k}^k}\right) = \left[\sum_{r=1}^{k} e^{V_{ir}^k / 1-\sigma_k}\right]^{1-\sigma_k}$$

para la elección de la alternativa dentro del grupo se obtiene que

$$P_{ij/k} = \frac{e^{V_{ij}^k / 1-\sigma_k} \left[\sum_{r=1}^{J_k} e^{V_{ir}^k / 1-\sigma_k}\right]^{-\sigma_k}}{\left[\sum_{r=1}^{J_k} e^{V_{ir}^k / 1-\sigma_k}\right]^{1-\sigma_k}} = \frac{e^{V_{ij}^k / 1-\sigma_k}}{\sum_{r=1}^{J_k} e^{V_{ir}^k / 1-\sigma_k}} = \frac{e^{V_{ij}^k / 1-\sigma_k}}{e^{I_k}}$$

Y para la elección del grupo k se considerará la función

$$G(a_1, \ldots\ldots, a_m) = \sum_{k=1}^{m} a_k \left[\sum_{r=1}^{J_k} e^{V_{ir}^k / 1-\sigma_k}\right]^{1-\sigma_k}$$

de donde

$$P_i^k = \frac{a_k \left[\sum_{r=1}^{J_k} e^{V_{ir}^k / 1-\sigma_k}\right]^{1-\sigma_k}}{\sum_{s=1}^{m} a_s \left[\sum_{r=1}^{J_s} e^{V_{ir}^s / 1-\sigma_s}\right]^{1-\sigma_s}} = \frac{a_k \, e^{I_k(1-\sigma_k)}}{\sum_{s=1}^{m} a_s \, e^{I_s(1-\sigma_s)}}$$

En consecuencia la probabilidad de elección es de nuevo (3-13):

$$P_{ij} = \frac{a_k\, e^{I_k(1-\sigma_k)}}{\sum_{s=1}^{m} a_s\, e^{I_s(1-\sigma_s)}} \frac{e^{V_{ij}^k/1-\sigma_k}}{e^{I_k}} = \frac{a_k\, e^{V_{ij}^k/1-\sigma_k}\, e^{-I_k\sigma_k}}{\sum_{s=1}^{m} a_s\, e^{I_s(1-\sigma_s)}}$$

Para $a_k = a, \quad k = 1,2,.....,m$ se obtendrá el logit multinomial anidado.

Con lo anterior se puede observar que la familia de modelos GEV es bastante general y que permite obtener modelos para una amplia variedad de situaciones de elección.

Modelo del Valor Extremo Generalizado Ordenado

Otro caso particular es el que se encuentra cuando las alternativas de elección están ordenadas, es decir, existe un orden natural entre las alternativas.

Podría tenerse una situación en la que las características no observables de las alternativas pueden afectar de forma similar a dos o más de ellas. En este caso, además de disponer de la ordenación entre alternativas, existirá un cierto grado de correlación entre las componentes aleatorias de la utilidad de las alternativas que estén más próximas en la ordenación. Small (1987) habla de una situación de covarianzas próximas, y puede utilizarse un modelo GEV que considere esta estructura de covarianzas.

Una situación empírica analizada por Small es el retraso o adelanto en la llegada de un trabajador a su puesto de trabajo. Las alternativas son períodos de 5 minutos, considerando desde 40 minutos antes hasta 15 minutos después de la hora de comienzo de la jornada.

Para estas situaciones se van a considerar grupos de alternativas ponderados de forma diferente en la función $G(\cdot)$ según el grado de relación que tengan, pero no serán grupos excluyentes entre sí, sino que una alternativa pertenecerá a más de un grupo.

Supóngase que cada alternativa j está correlacionada con las M alternativas adyacentes a ella en la ordenación y que estarán por debajo y por arriba de ella en el orden natural.

Para cada alternativa j se puede definir el conjunto de alternativas que están relacionadas con ella como:

$$B_r = \{j \in C / r - M \leq j \leq r\} \qquad r = 1,.....,J+M$$

Se tendrán $J+M$ conjuntos que, por su definición, pueden contener hasta $M+1$ alternativas y están solapados.

Por ejemplo, para $M=3$ y $J=5$, los conjuntos B serán:

$$B_1 = \{1\}; \quad B_2 = \{1,2\}; \quad B_3 = \{1,2,3\}; \quad B_4 = \{1,2,3,4\}; \quad B_5 = \{2,3,4,5\};$$

$$B_6 = \{3,4,5\}; \quad B_7 = \{4,5\}; \quad B_8 = \{5\}$$

En general puede definirse la función $G(\cdot)$ como:

$$G(e^{V_{i1}},.....,e^{V_{iJ}}) = \sum_{r=1}^{J+M}\left[\sum_{j\in B_r} w_{r-j}\, e^{V_{ij}/1-\sigma_r}\right]^{1-\sigma_r}$$

siendo, igual que antes, σ_r una cantidad acotada en el intervalo unidad que mide el grado de similaridad dentro del grupo B, w_{r-j} son constantes no negativas tales que $\sum_{s=0}^{M} w_s = 1$ de forma que en un mismo grupo cada alternativa recibe un peso diferente según la correlación existente y de un grupo a otro cambia la ponderación de la misma alternativa.

Las probabilidades de elección adoptan la siguiente expresión:

$$P_{ij} = \sum_{s=j}^{j+M} \frac{w_{s-j}\, e^{V_{ij}/1-\sigma_s}\left[\sum_{k\in B_s} w_{s-k}\, e^{V_{ik}/1-\sigma_s}\right]^{-\sigma_s}}{\sum_{r=1}^{J+M}\left[\sum_{k\in B_r} w_{r-k}\, e^{V_{ik}/1-\sigma_r}\right]^{1-\sigma_r}} \qquad j = 1,2,.....,J \qquad (3\text{-}14)$$

Este modelo recibe el nombre de modelo del valor extremo generalizado ordenado (OGEV) y, por la construcción de los conjuntos de alternativas B_r, se tiene que cada alternativa pertenece exactamente a $M+1$ grupos, hecho que se refleja en la expresión de las probabilidades con la suma respecto a todos los grupos a los que pertenece.

El modelo OGEV refleja la situación de covarianza próxima mediante las ponderaciones w_{r-j}, ya que la covarianza entre dos alternativas recibe una contribución diferente desde cada subconjunto al que pertenecen en común.

También ahora puede considerarse la elección de la alternativa j como un proceso secuencial, eligiendo primero el grupo y en segundo lugar eligiendo la alternativa dentro de este grupo. Puesto que cada alternativa j pertenece a $M+1$ grupos, las probabilidades de elección serán:

$$P_{ij} = \sum_{s=j}^{j+M} P_{ij/B_s} P_i^{B_s}$$

y cada probabilidad intermedia adopta una forma GEV.

En este caso

$$P_{ij/B_s} = \frac{w_{s-j} e^{V_{ij}/1-\sigma_s}}{\sum_{k \in B_s} w_{s-k} e^{V_{ik}/1-\sigma_s}}$$

es la probabilidad de elegir la alternativa j dentro del subconjunto B_s, y

$$P_i^{B_s} = \frac{\sum_{k \in B_s} w_{s-k} e^{V_{ik}/1-\sigma_s}}{\sum_{r=1}^{J+M} \left[\sum_{k \in B_r} w_{r-k} e^{V_{ik}/1-\sigma_r} \right]^{1-\sigma_r}}$$

es la probabilidad de elegir el grupo B_s.

Small (1987) propone modelos OGEV particulares al considerar ponderaciones iguales para todas las alternativas de un mismo grupo $w_s = \frac{1}{M+1}$ y por admitir que el coeficiente de similaridad para todos los grupos es el mismo, $\sigma_r = \sigma, \quad r = 1,.....,J+M$.

3.1.6. Modelos Ordenados

Los modelos ordenados se utilizan en situaciones en las que las alternativas presentan un orden natural. Como ejemplos se pueden mencionar el nivel de educación de un individuo, el grado de enfermedad de un paciente, el tamaño de la vivienda de una familia o el número de aparatos de televisión que posee.

Un modelo adecuado a esta situación es el modelo OGEV comentado antes como un caso particular de modelo GEV.

Sin embargo, cuando no se considere una situación de covarianzas próximas, sino que la única relación entre las alternativas es el hecho de ser consecutivas, hay una posibilidad de encontrar una solución más fácil computacionalmente que la que proporciona el modelo OGEV.

Sea J el número de alternativas, que presentan una ordenación, y_i la variable respuesta y x_i las variables explicativas. En estos modelos, aunque la variable dependiente tiene J modalidades o respuestas, el vector de la variable latente está formado por una única componente.

Se definen $I_j = [\alpha_j, \alpha_{j-1}[$, $j = 1,2,.....,J$ subintervalos en el dominio de la variable latente, con α_j, $j = 1,2,.....,J-1$, valores de la recta real, y $\alpha_0 = +\infty$, $\alpha_J = -\infty$. Desde estos subintervalos se define la variable respuesta a partir del modelo de variable latente $y_i^* = x_i'\beta - \varepsilon_i$ como:

$$y_i = \begin{cases} 1 & sii \quad y_i^* \in I_1 \\ 2 & sii \quad y_i^* \in I_2 \\ . & \\ . & \\ . & \\ J & sii \quad y_i^* \in I_J \end{cases}$$

La probabilidad de que el decisor elija la alternativa j se calculará del siguiente modo:

$$P_{ij} = P(y_i = j / x_i) = P(y_i^* \in I_j / x_i) = P(\alpha_j \leq x_i'\beta - \varepsilon_i < \alpha_{j-1} / x_i) =$$

$$P(x_i'\beta - \alpha_{j-1} < \varepsilon_i \leq x_i'\beta - \alpha_j / x_i) = F_{\varepsilon_i}(x_i'\beta - \alpha_j) - F_{\varepsilon_i}(x_i'\beta - \alpha_{j-1}) \tag{3-15}$$

donde $F_{\varepsilon_i}(\cdot)$ representa la función de distribución univariante de la variable aleatoria, ε_i, del modelo de variable latente.

Las probabilidades de elección de las alternativas 1 y J (alternativas extremas en la ordenación) vienen dadas por:

$$P(y_i = 1 / x_i) = F_{\varepsilon_i}(x_i'\beta - \alpha_1)$$

y

$$P(y_i = J / x_i) = 1 - F_{\varepsilon_i}(x_i'\beta - \alpha_{J-1}) \tag{3-16}$$

respectivamente.

Según la especificación distribucional que se admita para la variable aleatoria ε_i, se obtendrán diferentes modelos de respuesta ordenada. Estos modelos presentan una ventaja computacional respecto a los otros modelos multinomiales, la variable aleatoria es una variable univariante. La desventaja que tienen estos modelos es la pérdida de flexibilidad, la utilización de los mismos está muy limitada a situaciones específicas que cumplan las condiciones exigidas por el modelo.

3.1.7. Modelos de Eliminación

En algunas situaciones prácticas, el razonamiento que sigue el decisor para elegir la alternativa es buscar aquella que sea menos mala para él. El decisor considera que no se decantará por alternativas que presenten ciertos rasgos, así pues en un primer paso elimina todas aquellas alternativas que tengan esa característica no deseada. En un segundo paso hay otros rasgos de las alternativas que tampoco le interesan y por ello descartará otro subconjunto de alternativas, y así sucesivamente hasta que en el proceso de eliminación se queda con una única alternativa.

Los modelos de respuesta cualitativa que reflejan este proceso de elección del decisor en el que la decisión se toma mediante un proceso secuencial de eliminación de alternativas, se denominan modelos de eliminación.

Formalmente, este proceso de eliminación secuencial sería del siguiente modo: del conjunto total de las alternativas, C, se va pasando a subconjuntos B_r, cada vez con un número de alternativas inferior:

$$C \rightarrow B_1 \rightarrow B_2 \rightarrow \cdots\cdots \rightarrow \{j\}$$

Para elegir la alternativa $j \in C$ hay tantas posibilidades como secuencias de subconjuntos que contengan a dicha alternativa. La probabilidad de esa alternativa se calcula como la suma de las probabilidades para cada secuencia de subconjuntos:

$$\mathrm{P}_{ij} = \sum_{B \in \wp(C)} \mathrm{P}(j / B, x_i)\, Q(B / C, x_i) \qquad (3\text{-}17)$$

donde el sumatorio se extiende sobre todos los posibles subconjuntos de C, es decir, sobre $B \in \wp(C)$; $Q(B / C, x_i)$ es una probabilidad de transición y representa la probabilidad de eliminar las alternativas que están en el subconjunto $C - B$ y $\mathrm{P}(j / B, x_i)$ representa la probabilidad de elegir la alternativa j en el subconjunto B.

Las probabilidades de transición deben verificar las siguientes propiedades:

1. $Q(B / C, x_i) \geq 0$

2. $\sum_{B \in \wp(C)} Q(B / C, x_i) = 1$

3. $Q(\varnothing / C, x_i) = 0$

4. $Q(C / C, x_i) < 1$

La propiedad 3 se exige para evitar eliminar todas las alternativas y la 4 para asumir que el decisor es capaz de eliminar alternativas.

Si el subconjunto B está formado únicamente por la alternativa j la probabilidad de elegir esa alternativa en B es igual a 1 y si la alternativa j no pertenece al subconjunto B, entonces la probabilidad correspondiente es nula. En

el caso que B contenga más de una alternativa, la probabilidad se obtiene repitiendo el proceso anterior; es decir, calculando

$$P(j / B, x_i) = \sum_{A \in \wp(B)} P(j / A, x_i)\, Q(A / B, x_i)$$

Un tipo de estos modelos, propuesto por Tversky (1972a), son los modelos de eliminación por aspectos o más comunmente denominados modelos EBA.

En el modelo EBA la elección de la alternativa se realiza mediante la selección de un aspecto o característica que describe a la alternativa. Cada aspecto tiene asociada una probabilidad de ser seleccionado equivalente a su peso, entendiendo por peso la importancia de este aspecto frente a los demás. En el proceso, el decisor selecciona un aspecto y se eliminan todas aquellas alternativas que no lo poseen. Una vez eliminado el primer subconjunto de alternativas, se selecciona un nuevo aspecto y se repite el proceso hasta que queda una única alternativa.

Por ejemplo una situación en la que se desea adquirir una vivienda secundaria. Los aspectos considerados por el hogar son la disponibilidad de una zona de esparcimiento cercana a la vivienda, el precio y el tamaño. En primer lugar se eliminan todas aquellas viviendas que no tienen zona de esparcimiento, a continuación se reduce el número de viviendas según el precio y en último lugar se considera el tamaño de las viviendas.

La forma de cuantificar las preferencias por un subconjunto frente a cualquier otro subconjunto de su mismo nivel se realiza mediante las probabilidades de transición.

Sea un subconjunto de alternativas B y sea A un subconjunto cualquiera tal que $A \subseteq B$. La probabilidad de pasar de B a A, si se prefieren las características comunes de las alternativas que están en A frente a las características de los otros subconjuntos de B, se define como:

$$Q(A / B, x_i) = \frac{\upsilon_A}{\sum_{D \subseteq B} \upsilon_D} \qquad para \quad A \subseteq B$$

siendo υ_A un valor escala no negativo que representa las características comunes del subconjunto A. Éste se puede interpretar como la probabilidad de extraer un aspecto que es único para A y común para todas las alternativas dentro de A.

El modelo EBA presenta la dificultad de especificar los valores escala υ_A. Estos valores deben ser función de las características de las alternativas que pertenecen al subconjunto A. Si se representa a dichas características por x_A, se tiene que $\upsilon_A = v(x_A)$.

Asumiendo una relación log-lineal en los parámetros, $\ln \upsilon_A = x'_A \beta_A$, se tendrá que las probabilidades de transición tienen la forma funcional de las probabilidades logit multinomial:

$$Q(A / B, x_i) = \frac{e^{x'_A \beta_A}}{\sum_{D \subseteq B} e^{x'_D \beta_D}}$$

y las probabilidades de elección del modelo EBA serán sumas de productos de las probabilidades anteriores.

Todos los modelos comentados anteriormente podrían obtenerse como casos especiales del modelo EBA. Por ejemplo, el modelo logit multinomial se obtiene considerando $\upsilon_A = 0$ para todos los subconjuntos A con más de una alternativa.

Modelos de Eliminación Jerárquicos

Cuando las alternativas pueden representarse gráficamente en una estructura de árbol, denominado árbol de preferencias, cada alternativa del conjunto de elección viene determinada biunívocamente por una única secuencia de subconjuntos:

$$\{j\} = B_{j_0} \subset \cdots\cdots \subset B_{j_{T-1}} \subset B_{j_T} = C$$

La elección de una alternativa j se realiza mediante un proceso de eliminación jerárquico. El decisor elige un nodo del árbol (cada nodo identifica todas las alternativas que están por debajo de él) y después otro nodo situado inferiormente en el arbol, hasta llegar a la rama correspondiente a la alternativa j. La probabilidad de elección se obtiene como:

$$\mathrm{P}(j / C, x_i) = Q(B_{j_0} / B_{j_1}, x_i)\, Q(B_{j_1} / B_{j_2}, x_i) \cdots\cdots Q(B_{j_{T-1}} / B_{j_T}, x_i) \tag{3-18}$$

Este tipo de modelos de respuesta cualitativa recibe el nombre de modelos de eliminación jerárquicos. Según la especificación de las probabilidades de transición se obtienen los diferentes modelos de eliminación.

Si las alternativas, en una estructura jerárquica, son eliminadas por sus aspectos, el modelo de eliminación jerárquico recibe el nombre de modelo HEBA.

3.1.8. Modelos Secuenciales

En este epígrafe se van a presentar modelos de respuesta cualitativa que consideran que el individuo elige una alternativa mediante un proceso secuencial de toma de decisiones intermedias.

Para ilustrar este procedimiento de elección considérese el siguiente problema: un individuo debe decidir entre comprar o alquilar una segunda vivienda y la situación geográfica (montaña o playa).

Una posibilidad es considerar cuatro alternativas de elección (comprar en montaña, comprar en playa, alquilar en montaña, alquilar en playa) y utilizar cualquiera de los modelos de respuesta cualitativa multinomiales ya presentados.

Supóngase que, por el contrario, el individuo realiza primero la elección entre el área geográfica (montaña o playa), y una vez tomada esta decisión se plantea la elección entre comprar o alquilar.

Los modelos que reflejan esta procedimiento de elección reciben el nombre de modelos secuenciales.

Admitiendo la hipótesis de independencia entre las diferentes etapas del proceso, las probabilidades de elección se calcularán como el producto de las probabilidades de elección intermedias.

En el ejemplo propuesto, la probabilidad de comprar una vivienda en la playa se calcularía como el producto de la probabilidad de elegir playa frente a montaña y la probabilidad de comprar frente a alquilar.

La ventaja de los modelos secuenciales frente a otros modelos de respuesta cualitativa multinomiales es que, al admitir independencia entre las diferentes etapas del proceso de elección, las probabilidades de elección se calcularán como

el producto de las probabilidades intermedias que se plantean sobre situaciones de elección más simples (se puede llegar a situaciones de elección binarias en todas las etapas).

La especificación de las probabilidades intermedias dará lugar al modelo secuencial correspondiente. Dependiendo de las alternativas de elección y del proceso planteado se llegará a diferentes modelos secuneciales.

Modelos Secuenciales para Alternativas Ordenadas

Si para analizar situaciones de elección con alternativas ordenadas se considera un proceso secuencial para elegir la alternativa j tal que la elección de esa alternativa j implicará que se prefiere ésta a todas las alternativas de rango inferior, se hablará de modelo secuencial para alternativas ordenadas.

Estos modelos asumen que cuando el decisor elige una alternativa j, expresa que todas las alternativas de orden inferior han sido elegidas previamente.

El individuo elige en un primer paso entre las dos primeras alternativas. Si elige la alternativa mayor, en un segundo paso se plantea la elección entre esta alternativa y la inmediatamente superior; y así sucesivamente. El proceso finaliza cuando en la elección entre dos alternativas el individuo decide quedarse con la alternativa de menor rango. En este proceso secuencial se asume independencia entre cada etapa y las probabilidades de elección se obtienen como el producto de todas las probabilidades de las elecciones binarias realizadas.

Para la obtención de las probabilidades de elección se deben considerar dos postulados:

1. Ninguna alternativa puede ser elegida sin implicar que todas las de orden inferior hayan sido elegidas. Si una alternativa no es elegida ninguna de rango superior puede ser elegida.

2. La diferencia de utilidades de alternativas consecutivas en el conjunto de elección son variables aleatorias independientes.

Considerando estos postulados y la teoría de la maximización de la utilidad, la probabilidad de elegir una alternativa j se calcula según el desarrollo presentado a continuación.

Sea I el conjunto de índices de las alternativas, que contiene a todos los enteros ordenados según el orden de las alternativas y sean I' e I'' dos subconjuntos mutuamente excluyentes tales que

$$I' = \{\forall k < j, k \in I\} \quad \text{e} \quad I'' = \{\forall k > j, k \in I\}$$

Utilizando el postulado 2 se tiene que la probabilidad de la alternativa j viene dada como:

$$P_{ij} = P(U_{ij} \geq U_{ik}, \forall k \in I) = P(U_{ij} \geq U_{ik}, \forall k \in I') P(U_{ij} \geq U_{ik}, \forall k \in I'') \quad (3\text{-}19)$$

Por el postulado 1 se obtiene:

$$P(U_{ij} \geq U_{ik}, \forall k \in I') = \left[\prod_{k=3}^{j} P(U_{ik} \geq U_{ik-1} / U_{ik-1} \geq U_{ik-2})\right] P(U_{i2} \geq U_{i1})$$

que por la independencia enunciada en el postulado 2 se puede escribir:

$$P(U_{ij} \geq U_{ik}, \forall k \in I') = \prod_{k=2}^{j} P(U_{ik} \geq U_{ik-1})$$

y por otro lado se tiene que:

$$P(U_{ij} \geq U_{ik}, \forall k \in I'') = P(U_{ij} \geq U_{ij+1})$$

Considerando estos resultados en la ecuación (3-19) original:

$$P_{ij} = P(U_{ij} \geq U_{ij+1}) \prod_{k=2}^{j} P(U_{ik} \geq U_{ik-1}) \quad (3\text{-}20)$$

Como se comentaba anteriormente, las probabilidades de elección se calculan a partir de probabilidades de elección binaria. Estas probabilidades binarias se pueden calcular utilizando cualquiera de los modelos binomiales (logit, probit,...).

Estos modelos presentan una ventaja muy interesante que es la facilidad de cálculo. Sin embargo, sólo pueden ser utilizados cuando exista independencia entre las diferentes etapas del proceso.

3.2. Modelos con variable dependiente limitada

Los modelos de variable dependiente limitada, que se comentan a continuación, están muy relacionados con los modelos de elección discreta o modelos de respuesta cualitativa.

La característica principal de estos modelos es que la variable dependiente es una variable continua, pero no totalmente observada. La diferencia con los modelos de respuesta cualitativa es que en éstos se consideraba una variable continua denominada variable latente de la cual se observaba únicamente su signo, mientras que en los modelos de variable dependiente limitada se dispone de una variable continua que sí que es observable en algunas ocasiones.

Supóngase una situación en la que se desea analizar el dinero que un hogar gasta mensualmente en reparaciones en su vivienda (albañilería, carpintería, aparatos eléctricos, etc...).

La variable de interés, y_i^*, es una variable continua que indica cuál es el gasto realizado por esta familia, pero únicamente es observable cuando el hogar ha realizado alguna reparación, $y_i^* > 0$.

Un modelo de variable dependiente limitada se define como:

$$y_i = \begin{cases} y_i^* & sii \quad y_i^* > 0 \\ 0 & en\ otro\ caso \end{cases} \tag{3-21}$$

donde $y_i^* = x_i'\beta + \varepsilon_i$

Al igual que en los modelos de respuesta cualitativa en general, el vector x_i recoge todas las características observables del individuo, β es un vector de parámetros desconocidos y ε_i es la perturbación aleatoria y corresponde a todas las características no observables.

Bajo la hipótesis de que la variable aleatoria ε_i sigue una distribución Normal, este modelo de variable dependiente limitada recibe el nombre de modelo tobit.

Además del modelo tobit existen otros tipos de modelos de variable dependiente limitada, cuya definición está supeditada a la combinación de diferentes variables, una variable continua y otra discreta o varias variables continuas con algún punto de truncamiento.

En algunas ocasiones un individuo se encuentra en la situación de tomar dos decisiones o elecciones interrelacionadas. Si las dos decisiones son de tipo cualitativo el problema puede transformarse en uno de elección discreta combinando las posibles respuestas. Un ejemplo de esta situación la presenta Train (1986). Un individuo elige cuantos coches tener y el modo de ir al trabajo. Supóngase por simplicidad que el número de coches posibles es 0, 1, ó 2 y que el modo de ir al trabajo puede ser coche o autobús. El problema de las dos elecciones se transforma en un único problema donde el decisor elige entre cinco alternativas: 0 coches y va en autobús, 1 coche y va en coche, 1 coche y va en autobús, 2 coches y va en coche o 2 coches y va en autobús. Un modelo GEV servirá para analizar este problema.

Sin embargo en algunas situaciones no son las dos elecciones de tipo cualitativo, sino que una variable respuesta es de tipo discreto y la otra de tipo continuo. A continuación se van a presentar algunas de las posibles situaciones que pueden ser analizadas con un modelo de variable dependiente limitada.

Sean dos variables continuas $y_{1i}^{*} = x_{1i}'\beta_1 + \varepsilon_{1i}$ e $y_{2i}^{*} = x_{2i}'\beta_2 + \varepsilon_{2i}$, de la primera únicamente se observa su signo y la segunda es observada según el signo de la primera. En este caso se define la variable respuesta como:

$$y_{2i} = \begin{cases} y_{2i}^{*} & sii \quad y_{1i}^{*} > 0 \\ 0 & sii \quad y_{1i}^{*} \leq 0 \end{cases} \tag{3-22}$$

Como ejemplo se podría considerar una situación en la que se desea analizar el dinero que un individuo se gasta si decide realizar un viaje en sus vacaciones.

La variable $y_{2i}^{*} = x_{2i}'\beta_2 + \varepsilon_{2i}$ representa el gasto realizado por el individuo en el viaje, y evidentemente sólo es observada cuando dicho individuo ha tomado la decisión de realizar un viaje.

En primer lugar el individuo decide, en base a las características x_{1i}, realizar o no el viaje. Esta decisión puede modelizarse mediante una variable latente $y_{1i}^{*} = x_{1i}'\beta_1 + \varepsilon_{1i}$, de la que únicamente se observa el signo, considerando que la respuesta es afirmativa (realiza el viaje) cuando $y_{1i}^{*} > 0$.

También se considera ahora que las variables ε_{1i} y ε_{2i} siguen una distribución conjuntamente Normal con medias cero, varianzas σ_1^2 y σ_2^2 respectivamente y covarianza igual a σ_{12}.

Otra posibilidad sería considerar tres variables continuas:

$$y_{1i}^* = x_{1i}'\beta_1 + \varepsilon_{1i}$$

$$y_{2i}^* = x_{2i}'\beta_2 + \varepsilon_{2i}$$

$$y_{3i}^* = x_{3i}'\beta_3 + \varepsilon_{3i}$$

de forma que las variables respuesta que se observan son:

$$y_{1i} = \begin{cases} y_{1i}^* & sii \quad y_{1i}^* > 0 \\ 0 & sii \quad y_{1i}^* \leq 0 \end{cases}$$

$$y_{2i} = \begin{cases} y_{2i}^* & sii \quad y_{1i}^* > 0 \\ 0 & sii \quad y_{1i}^* \leq 0 \end{cases}$$

$$y_{3i} = \begin{cases} y_{3i}^* & sii \quad y_{1i}^* \leq 0 \\ 0 & sii \quad y_{1i}^* > 0 \end{cases}$$

(3-23)

Es decir, el signo de la variable y_{1i}^* es el que determina cual de las otras dos variables, y_{2i}^* ó y_{3i}^*, es observada, pero a su vez, la variable y_{1i}^* también es observada según su signo.

Una situación que se modeliza adecuadamente con este modelo es la siguiente. Una familia estudia las posibilidades que tiene de realizar una instalación de aire acondicionado en su vivienda. Sea y_{1i}^* la variable que representa el coste de la instalación, y_{2i}^* el gasto en mantenimiento del aire acondicionado e y_{3i}^* el gasto en otro tipo de refrigeración. La variable y_{2i}^* únicamente es observable cuando la familia ha decidido instalar el aire acondicionado y por lo tanto realiza un gasto en su compra, $y_{1i}^* > 0$ y la variable y_{3i}^* se observa cuando no se decide instalar el aire acondicionado.

Análogamente a los modelos anteriores, se asume la hipótesis de Normalidad para el vector aleatorio $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i})'$, con vector de medias cero y matriz de varianzas-covarianzas

$$\Omega_i = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

Dentro de este esquema se pueden considerar los modelos de variable dependiente continua con selección muestral, comúnmente denominados modelos continuos-discretos, y que están definidos a partir de variables continuas observables únicamente para ciertos individuos.

Un ejemplo de estos modelos se enmarca dentro del mercado de la vivienda. El decisor debe elegir entre comprar o alquilar su vivienda y además cuanto dinero gastar en ella. La primera decisión es de tipo cualitativo y cualquiera de los modelos de elección discreta estudiados puede ser adecuado para analizarla. La segunda decisión es de tipo continuo y será necesario recurrir a otro tipo de modelos para analizarla.

Estas situaciones son las denominadas "continuas-discretas" y se caracterizan porque la elección continua depende de la respuesta dada en el modelo discreto.

En términos generales el modelo se puede plantear como el siguiente sistema de ecuaciones simultáneas:

$$\begin{aligned} y_{1i} &= x_{1i}'\beta + \varepsilon_{1i} \quad sii \quad I_i = 1 \\ y_{2i} &= x_{2i}'\beta + \varepsilon_{2i} \quad sii \quad I_i = 0 \end{aligned} \tag{3-24}$$

con

$$I_i = \begin{cases} 1 & sii \quad I_i^* = z_i'\gamma - \varepsilon_i > 0 \\ 0 & en\ otro\ caso \end{cases}$$

Es fácil ver que la última ecuación corresponde a un modelo discreto binomial, que puede ser planteado como un modelo probit o logit, o cualquier otro modelo de respuesta cualitativa binomial.

Las dos primeras ecuaciones son las correspondientes a la variable continua, la primera de ellas, y_{1i} es la correspondiente a los individuos que tienen por respuesta del modelo binomial el valor 1 y la variable y_{2i} corresponde a los individuos cuya respuesta del modelo binomial es 0.

La variable continua y, de interés, tiene dos especificaciones diferentes según cual sea la elección que realiza el individuo. Es decir, el modelo continuo está determinado por el modelo de respuesta cualitativa que aparece en el sistema de ecuaciones.

Una característica común de los modelos tobit y sus variantes comentadas es la hipótesis de Normalidad que se asume para las variables aleatorias que aparecen en las ecuaciones.

No obstante, Maddala (1983), argumenta que esta hipótesis no es imprescindible y que puede relajarse utilizando transformaciones normalizadoras o bien considerar distribuciones no Normales, como plantean Amemiya y Boskin (1974) al suponer una distribución Log-Normal.

3.3. Estimación del vector de parámetros de un modelo de respuesta cualitativa

En este epígrafe se van a exponer técnicas para la estimación del vector de parámetros desconocidos que aparecen en los modelos de respuesta cualitativa. Este vector de parámetros se denotará genéricamente por θ y está formado tanto por los pesos asociados a las variables explicativas, β, como por cualquier otro parámetro identificativo de la distribución de probabilidad asociada al modelo correspondiente.

En la estimación del vector paramétrico θ hay que diferenciar dos supuestos distintos. En primer lugar se considerará que el vector de características x son cantidades observadas fijas, que no tienen carácter aleatorio. En segundo lugar, el planteamiento es que dicho vector de características x es un vector de variables aleatorias, observables, y cuya distribución de probabilidad se representará como $\mathrm{P}(x)$.

3.3.1. Estimación cuando las variables explicativas son fijas

Cuando se admite que las variables explicativas son cantidades fijas observadas por el investigador, la aleatoriedad del proceso inferencial radica únicamente en la variable dependiente, y_i.

El investigador realizará un procedimiento de muestreo sobre la población de la variable respuesta y_i. Puesto que para cada individuo i se conoce el valor de las variables x_i, la distribución de probabilidad asociada a las observaciones y_i es una distribución de probabilidad condicionada a los valores de dichas variables, y que estará en función del vector paramétrico θ desconocido.

Dadas las observaciones $y_i = j$, con distribución de probabilidad que se denotará como $\mathrm{P}(y_i = j / x_i, \theta) = \mathrm{P}(j_i / x_i, \theta)$, se extrae una muestra de N individuos, $i = 1,2,.....,N$, y desde ella se estima el vector de parámetros θ que determina la probabilidad de respuesta dados los valores x_i de las variables explicativas.

La selección muestral se realiza mediante un proceso de muestreo aleatorio. Dentro de este tipo de muestreo se puede considerar el proceso de muestreo

aleatorio simple y el proceso de muestreo estratificado. En este apartado no se diferencian ambos, puesto que las verosimilitudes correspondientes conducen a problemas de maximización equivalentes y proporcionan la misma estimación para el vector de parámetros desconocidos θ.

Considérense N individuos, $i = 1,2,.....,N$, que pueden elegir una alternativa j de un conjunto de elección que consta de J posibles alternativas. Se representa por x_i el vector de variables explicativas, por j_i la respuesta del individuo y por θ el vector de parámetros desconocidos. La función de verosimilitud de ese vector θ viene dada por:

$$L(\theta) = \prod_{i=1}^{N} P(j_i / x_i, \theta) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[P(y_{ij} = 1 / x_i, \theta) \right]^{y_{ij}} \qquad (3\text{-}25)$$

La estimación máximo-verosímil del vector paramétrico θ se obtiene maximizando la función de verosimilitud anterior (o equivalentemente su logaritmo) respecto de dicho parámetro θ. El logaritmo de la función de verosimilitud adopta la expresión siguiente:

$$LL(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} \ln P(y_{ij} = 1 / x_i, \theta) \qquad (3\text{-}26)$$

Y la estimación máximo-verosímil se obtiene resolviendo la ecuación siguiente:

$$\frac{d}{d\theta} LL(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} \frac{1}{P(y_{ij} = 1 / x_i, \theta)} \frac{d}{d\theta} P(y_{ij} = 1 / x_i, \theta) = 0$$

Como esta ecuación no es lineal en el vector de parámetros θ será necesario utilizar algún procedimiento iterativo para resolverla. Existen muchos métodos iterativos que permiten llegar a una solución, en estos modelos los más utilizados son el método de Newton-Raphson y el método de "Scoring".

A continuación se presenta la fórmula recursiva que utiliza cada uno de los dos métodos iterativos:

(a) Método de Newton-Raphson

Partiendo de una estimación inicial θ_0, el valor de la estimación r-ésima viene dado por:

$$\hat{\theta}_r = \hat{\theta}_{r-1} - \left[\frac{d^2}{d\theta\, d\theta'} \mathrm{LL}(\theta)\Big|_{\hat{\theta}_{r-1}} \right]^{-1} \left[\frac{d}{d\theta} \mathrm{LL}(\theta)\Big|_{\hat{\theta}_{r-1}} \right] \qquad (3\text{-}27)$$

donde $\frac{d}{d\theta}\mathrm{LL}(\theta)$ y $\frac{d^2}{d\theta\, d\theta'}\mathrm{LL}(\theta)$, son la primera y segunda derivadas del logaritmo de la función de verosimilitud y $\hat{\theta}_{r-1}$ es la estimación máximo-verosímil en el paso anterior del proceso.

(b) Método de "Scoring"

La estimación máximo-verosímil por este procedimiento se obtiene sustituyendo los valores correspondientes en la siguiente expresión:

$$\hat{\theta}_r = \hat{\theta}_{r-1} - \left[\mathrm{E}\left[\frac{d^2}{d\theta\, d\theta'} \mathrm{LL}(\theta)\Big|_{\hat{\theta}_{r-1}} \right] \right]^{-1} \left[\frac{d}{d\theta} \mathrm{LL}(\theta)\Big|_{\hat{\theta}_{r-1}} \right] \qquad (3\text{-}28)$$

donde $\mathrm{E}\left[\frac{d^2}{d\theta\, d\theta'} \mathrm{LL}(\theta) \right]$ representa la esperanza de la segunda derivada del logaritmo de la función de verosimilitud y $\hat{\theta}_{r-1}$ es la estimación en el paso $r-1$.

Para ambos métodos iterativos la estimación final, después de alcanzar el criterio de convergencia se representará por $\hat{\theta}$.

Un desarrollo más detallado de estos métodos iterativos y algunos otros pueden encontrarse en Goldfeld y Quandt (1972).

Sustituyendo los valores correspondientes a la primera y segunda derivada y a la esperanza de la segunda derivada para los modelos de respuesta cualitativa más usuales, desarrollados en el epígrafe anterior, se obtienen las estimaciones máximo-verosímiles para los vectores paramétricos desconocidos.

En el apéndice A se encuentra la demostración de las propiedades de consistencia y eficiencia asintótica de estos estimadores.

3.3.2. Estimación cuando las variables explicativas son aleatorias

En este apartado se va a admitir que las variables explicativas x son variables aleatorias observables con una distribución de probabilidad $\mathrm{P}(x)$.

Considerando las variables x como aleatorias, la estimación del vector de parámetros desconocidos θ puede plantearse desde varios puntos, llegando a diferentes estimaciones. Ahora, el carácter aleatorio de las variables x lleva a considerar una población multi-dimensional formada por las alternativas y las características.

Diseños muestrales

Sea un universo de individuos i, sobre los cuales se observa la alternativa elegida y sus características (características propias del individuo y de las alternativas). Sea $C \times Z$ el espacio producto de las alternativas y las características.

En un diseño muestral aleatorio simple se admite un único modelo generador de datos en la población $C \times Z$, caracterizada por la probabilidad de la alternativa j condicionada al vector de características x, $\mathrm{P}(j/x)$, y por la distribución de dichas características, que se especificará con la función $\mathrm{P}(x)$.

Desde las funciones anteriores, $\mathrm{P}(j/x)$ y $\mathrm{P}(x)$ se determina la distribución conjunta de alternativa y vector de características

$$\mathrm{P}(j,x) = \mathrm{P}(j/x)\mathrm{P}(x)$$

También se puede considerar la distribución de las características x condicionada a la alternativa j, $\mathrm{P}(x/j)$ y la distribución marginal de las alternativas, $\mathrm{P}(j)$.

En un diseño de muestreo estratificado se divide el universo de individuos en un número finito o numerable de estratos, con B el conjunto de estratos, y se supone un modelo generador de datos diferente para cada una de las subpoblaciones.

Cada una de estas subpoblaciones no es más que un subconjunto de observaciones, $A_b \subset C \times Z$, que define al estrato b como el conjunto de individuos cuya observación (j_i, x_i) está en A_b.

Hay que destacar la diferencia existente entre una muestra estratificada y una muestra representativa. Mientras en el muestreo estratificado se suponen modelos generadores de datos diferentes para cada subpoblación, en una muestra representativa se asume un único modelo generador de datos, pero los individuos del universo aparecen clasificados por algún factor y al extraer la muestra se obliga a extraer un número determinado de individuos de cada tipo.

En el muestreo estratificado, dentro de cada estrato b, y por lo tanto, en su subpoblación asociada A_b, se asume un modelo generador de datos caracterizado por la función de probabilidad de la alternativa j condicionada al vector de características x, $\mathrm{P}(j/x,b)$, y por la distribución de dichas características dentro de la subpoblación A_b, $\mathrm{P}(x/b)$.

Desde las funciones $\mathrm{P}(j/x,b)$ y $\mathrm{P}(x/b)$ se determina la distribución de probabilidad de la observación (j,x) condicionada al estrato b:

$$\mathrm{P}((j,x)/b) = \mathrm{P}(j/x,b)\,\mathrm{P}(x/b)$$

Además asignando una distribución de probabilidad sobre el conjunto de estratos B, $H(b)$, se calcula la distribución conjunta de la observación (j,x) y el estrato b:

$$\mathrm{P}((j,x),b) = \mathrm{P}((j,x)/b)\,H(b)$$

En los modelos de respuesta cualitativa se asume que la relación entre las alternativas y las características de un individuo no varía para cada una de las subpoblaciones, es decir, $\mathrm{P}(j/x)$ coincide para cualquier estrato b. Así en un diseño estratificado las diferencias en los modelos generadores de datos son debidas a la distribución de probabilidad que tenga el vector de características dentro del estrato. La distribución de probabilidad del par (j,x) condicionada al estrato b viene dada por:

$$\mathrm{P}((j,x)/b) = \mathrm{P}(j/x)\,\mathrm{P}(x/b)$$

Nótese que el muestreo aleatorio simple puede considerarse como un muestreo estratificado con un único estrato, y por lo tanto con un único modelo generador de datos.

Pueden considerarse dos casos particulares de muestreo estratificado: muestreo exógeno y muestreo endógeno o basado en la elección. El muestreo basado en la elección se utiliza en situaciones con alguna alternativa poco usual para conseguir que en la muestra aparezcan individuos que eligen esta alternativa (en un muestreo aleatorio simple sería posible no obtener individuos con dicha elección).

En el muestreo exógeno los subconjuntos A_h de observaciones se obtienen particionando el espacio de atributos Z en subconjuntos Z_h, es decir, $A_h = C \times Z_h$; y en el muestreo basado en la elección los subconjuntos de observaciones A_h se obtienen particionando el espacio de alternativas C en subconjuntos C_h, es decir, $A_h = C_h \times Z$.

En todo este apartado se considerará una muestra de tamaño N extraída por alguno de los diseños muestrales descritos anteriormente: muestra aleatoria simple o muestra estratificada (exógena o basada en la elección).

Función de verosimilitud

En los modelos de respuesta cualitativa el interés radica en la obtención de la probabilidad de respuesta, $\mathrm{P}(j/x)$, cuya forma funcional va a admitirse totalmente conocida a excepción de un vector paramétrico θ. Por lo tanto, de ahora en adelante se representará la dependencia de ese vector paramétrico en la probabilidad de respuesta

$$\mathrm{P}(j/x)=\mathrm{P}(j/x,\theta)$$

En un proceso de muestreo aleatorio simple, la verosimilitud de una observación viene dada por:

$$\mathrm{L}((j,x)) = \mathrm{P}(j/x,\theta)\,\mathrm{P}(x) \qquad (3\text{-}29)$$

En un proceso de muestreo estratificado general, si Z es finito, la verosimilitud de una observación (j,x) y un estrato b se puede obtener como:

$$L((j,x),b) = P((j,x)/b)\,H(b) = \frac{P(j,x)}{\sum_{(k,z)\in A_h} P(k,z)} H(b) = \frac{P(j/x,\theta)\,P(x)}{\sum_{(k,z)\in A_h} P(k/z,\theta)\,P(z)} H(b) \tag{3-30}$$

donde $P((j,x)/b)$ se obtiene como la probabilidad de la observación (j,x) frente a la probabilidad del resto de observaciones del estrato b:

$$P((j,x)/b) = \begin{cases} 0 & sii \quad (j,x) \notin A_h \\ \dfrac{P(j,x)}{\sum_{(k,z)\in A_h} P(k,z)} & sii \quad (j,x) \in A_h \end{cases}$$

En la notación de la distribución de probabilidad de la observación (j,x) se suprime la referencia al estrato b porque está incluida en la propia definición de la probabilidad.

Análogamente en la descomposición $P(j,x) = P(j/x,\theta)\,P(x)$ también se suprime la referencia al estrato b, aunque esto no presupone que la distribución de probabilidad de las características es la misma para cada estrato, sino que puede variar a través de los estratos.

Si Z no es finito, la verosimilitud de una observación (j,x) y un estrato b viene dada como:

$$L((j,x),b) = \frac{P(j/x,\theta)\,P(x)}{\sum_{C_h} \int_{Z_h} P(k/z,\theta)\,P(z)\,dz} H(b)$$

Si el muestreo estratificado es un muestreo exógeno, la verosimilitud de una observación (j,x) y un estrato b es:

$$L((j,x),b) = \frac{P(j/x,\theta)\,P(x)}{\sum_{z\in Z_h} P(z)} H(b) \tag{3-31}$$

ya que:

$$\sum_{(k,z)\in A_b} \mathrm{P}(k/z,\theta)\,\mathrm{P}(z) = \sum_{z\in Z_b}\sum_{k\in C} \mathrm{P}(k/z,\theta)\,\mathrm{P}(z) = \sum_{z\in Z_b} \mathrm{P}(z) \sum_{k\in C} \mathrm{P}(k/z,\theta) = \sum_{z\in Z_b} \mathrm{P}(z)$$

Y en el muestreo basado en la elección o endógeno, la verosimilitud de una observación (j,x) y un estrato b es:

$$\mathrm{L}\big((j,x),b\big) = \frac{\mathrm{P}(j/x,\theta)\,\mathrm{P}(x)}{\sum_{(k,z)\in C_b\times Z} \mathrm{P}(k/z,\theta)\,\mathrm{P}(z)}\,H(b) = \frac{\mathrm{P}(j/x,\theta)\,\mathrm{P}(x)}{\mathrm{P}(b/\theta)}\,H(b) \quad (3\text{-}32)$$

Dentro del muestreo basado en la elección se va a considerar el caso particular en el que cada subconjunto C_b está definido por una única alternativa, así cada estrato b está identificado por la alternativa j que constituye el subconjunto C_b, y la verosimilitud de una observación (j,x) y un estrato b será:

$$\mathrm{L}\big((j,x),b\big) = \frac{\mathrm{P}(j/x,\theta)\,\mathrm{P}(x)}{\sum_{(k,z)\in C_b\times Z} \mathrm{P}(k/z,\theta)\,\mathrm{P}(z)}\,H(b) = \frac{\mathrm{P}(j/x,\theta)\,\mathrm{P}(x)}{\sum_{Z} \mathrm{P}(j/z,\theta)\,\mathrm{P}(z)}\,H(b) =$$

$$\frac{\mathrm{P}(j/x,\theta)\,\mathrm{P}(x)}{\mathrm{P}(j)}\,H(j) \quad (3\text{-}33)$$

Si Z es finito en el muestreo exógeno también se puede considerar una partición fina de Z, es decir, que cada subconjunto Z_b esté formado por un único valor x del vector de características, y denotando $H(b)$ como $g(x)$, la verosimilitud de una observación (j,x) y un estrato b se puede expresar como:

$$\mathrm{L}\big((j,x),b\big) = \frac{\mathrm{P}(j/x,\theta)\,\mathrm{P}(x)}{\sum_{Z_b} \mathrm{P}(z)}\,H(b) = \frac{\mathrm{P}(j/x,\theta)\,\mathrm{P}(x)}{\mathrm{P}(x)}\,g(x) = \mathrm{P}(j/x,\theta)\,g(x) \quad (3\text{-}34)$$

Se observa que la verosimilitud adopta una expresión distinta según el tipo de muestreo efectuado.

Esta diferencia en las expresiones de la función de verosimilitud se refleja en el proceso de obtención de las estimaciones máximo-verosímiles, siendo un proceso bastante complejo para algunas de ellas y menos para otras.

A continuación se planteará la obtención de las estimaciones por el método de la máxima-verosimilitud para cada uno de los diferentes tipos de muestreo, analizando las particularidades que vayan apareciendo en cada caso y desarrollando las soluciones alternativas que vayan surgiendo para aquellas situaciones en las que no exista solución directa o bien que ésta sea de difícil obtención.

Antes de comenzar con el análisis de las estimaciones se enuncian las hipótesis de regularidad e identificabilidad que se exigirán en todo el desarrollo, y que son necesarias para garantizar las propiedades de consistencia y Normalidad asintótica de los correspondientes estimadores (la demostración de las mismas se encuentra en el apéndice A).

Hipótesis en modelos de elección cualitativa

En todo el proceso de obtención de estimadores, así como en la demostración de sus propiedades estadísticas, se van a utilizar las siguientes hipótesis:

Hipótesis 1 (positividad)

Para cada par $(j, x) \in C \times Z$ se verifica que

$$\mathrm{P}(j/x, \theta) > 0 \quad \forall \theta \in \Theta$$

o bien se verifica que

$$\mathrm{P}(j/x, \theta) = 0$$

Hipótesis 2 (identificabilidad)

Para cada $\theta \in \Theta$ tal que $\theta \neq \theta^*$ existe $A \subset C \times Z$ tal que:

$$\sum_A \mathrm{P}(k/z, \theta)\,\mathrm{P}(z) \neq \sum_A \mathrm{P}(k/z, \theta^*)\,\mathrm{P}(z)$$

Además el proceso de muestreo satisface las condiciones:

$$\bigcup_{h \in B} A_h = C \times Z$$

para cada $b \in B$

$$\sum_{A_b} \mathrm{P}(k/z,\theta)\,\mathrm{P}(z) > 0$$

Hipótesis 3 (espacio paramétrico)

El espacio paramétrico $\Theta \subset R^k$ es compacto.

Además existe un conjunto abierto Θ' en R^k al cual pertenece el verdadero valor del parámetro desconocido

$$\theta^* \in \Theta' \subset \Theta$$

Hipótesis 4 (espacio de atributos)

El espacio $Z \subset R^M$ es compacto.

Hipótesis 5 (regularidad)

La probabilidad de elección $\mathrm{P}(j/x,\theta)$ es continua en $C \times Z \times \Theta$.

Además para cada par $(j,x) \in C \times Z$ tal que $\mathrm{P}(j/x,\theta^*) > 0$, esta función es tres veces diferenciable para todo θ en un entorno de θ^* y se denota por R a una matriz $k \times J$ con columnas

$$\sum_z \frac{d}{d\theta}\mathrm{P}(j/x,\theta^*)\,\mathrm{P}(x)$$

cuyo rango es el mínimo valor de k y $J-1$, siendo k la dimensión del vector paramétrico θ.

La hipótesis 1, necesaria para utilizar métodos estándar en la demostración de la consistencia, implica que el soporte de la función de verosimilitud es independiente de θ, tanto en muestras exógenas como en muestras basadas en la elección.

Nótese que esta hipótesis proporciona una forma de tratar con alternativas j que no están disponibles para los decisores. Para estos pares (j,x) simplemente se considera $\mathrm{P}(j/x,\theta) = 0 \quad \forall\theta \in \Theta$.

Con la hipótesis 2 se asegura que el modelo con probabilidades de elección $P(j/x,\theta^*)$ es observacionalmente distinguible de todos los otros modelos de la forma $P(j/x,\theta)$ $\forall\theta\neq\theta^*$, y que el proceso muestral es tal que θ^* es identificable.

La primera parte de la hipótesis 3, la hipótesis 4 y la primera parte de la hipótesis 5 se utilizan en las demostraciones de la consistencia. Estas hipótesis pueden ser sustancialmente debilitadas imponiendo estructuras adicionales en $P(j/x,\theta)$ y $P(x)$. En particular pueden desarrollarse demostraciones asumiendo que $\Theta = R^k$ y $Z = R^M$.

La hipótesis 5 y la segunda parte de la hipótesis 3 son inocuas generalmente, aunque son utilizadas en la demostración de la Normalidad Asintótica de algunos estimadores.

Estimación máximo-verosímil bajo muestreo aleatorio simple y muestreo exógeno

Las estimaciones máximo-verosímiles del vector de parámetros θ se plantean para la solución más general de los modelos multinomiales. Los modelos binomiales se obtienen como un caso particular de los multinomiales, ya que basta considerar que $J = 2$.

Tras esta solución teórica su aplicación a modelos determinados, como el logit, probit, etc., se consigue utilizando la correspondiente expresión de las probabilidades $P(j/x,\theta)$. La estimación se obtiene utilizando cualquier método iterativo de los propuestos anteriormente, en el apartado 3.3.1.

A. La función de verosimilitud en un proceso de muestreo aleatorio simple para una muestra de tamaño N viene dada desde (3-29) por la expresión:

$$L(\theta) = \prod_{i=1}^{N} P(j_i,x_i) = \prod_{i=1}^{N} P(j_i/x_i,\theta)\,P(x_i)$$

y su logaritmo será:

$$LL(\theta) = \sum_{i=1}^{N} \ln P(j_i / x_i, \theta) + \sum_{i=1}^{N} \ln P(x_i)$$

La estimación máximo-verosímil será la solución al problema:

$$\max_{\theta \in \Theta} \sum_{i=1}^{N} \ln P(j_i / x_i, \theta) \tag{3-35}$$

B. Bajo un proceso de muestreo exógeno se tiene la siguiente función de verosimilitud:

$$L(\theta) = \prod_{i=1}^{N} P((j_i, x_i), b) = \prod_{i=1}^{N} \frac{P(j_i / x_i, \theta) P(x_i)}{\sum_{z \in Z_b} P(z)} H(b)$$

y su logaritmo:

$$LL(\theta) = \sum_{i=1}^{N} \left[\ln P(j_i / x_i, \theta) + \ln P(x_i) - \ln \sum_{z \in Z_b} P(z) + \ln H(b) \right]$$

Con este proceso de muestreo exógeno, la estimación máximo-verosímil que se obtiene es diferente según la información conocida a priori sobre las distribuciones marginales $P(x)$ y $P(j)$, ya que éstas deben verificar la siguiente relación:

$$P(j) = \sum_{Z} P(j / z, \theta) P(z)$$

Esta relación implica al vector de parámetros θ, por lo que debe de considerarse en la maximización.

Se verán varias situaciones con informaciones a priori diferentes. Cuando las distribuciones marginales $P(x)$ y $P(j)$ son ambas desconocidas ó cuando es conocida alguna de las dos distribuciones marginales y desconocida la otra, la estimación máximo-verosímil del vector de parámetros θ es la solución al problema:

$$\max_{\theta \in \Theta} \sum_{i=1}^{N} \ln P(j_i / x_i, \theta)$$

donde Θ es el espacio paramétrico y coincide con (3-35).

La última situación es cuando tanto $P(x)$ como $P(j)$ son conocidas, pero en este caso hay que considerar el hecho de que ambas distribuciones marginales están relacionadas:

$$P(j) = \sum_{z \in Z} P(j/z, \theta) P(z) \quad , j \in C$$

Ahora hay que tener en cuenta la relación existente entre las dos distribuciones marginales, por ello la estimación máximo-verosímil del vector de parámetros desconocidos θ será la solución al siguiente problema de maximización con restricciones:

$$\max_{\theta \in \Theta_0} \sum_{i=1}^{N} \ln P(j_i / x_i, \theta) \tag{3-36}$$

donde:

$$\Theta_0 = \left\{ \theta \in \Theta / P(j) = \sum_{z \in Z} P(j/z, \theta) P(z) \quad , j \in C \right\}$$

En cualquier caso el estimador obtenido desde (3-35) ó (3-36) es consistente y asintóticamente Normal. En el apéndice A se encuentran las correspondientes demostraciones.

Estimación máximo-verosímil bajo muestreo basado en la elección

En muestreo basado en la elección para el desarrollo de la estimación máximo-verosímil del vector de parámetros θ se plantea únicamente la solución teórica general, como en muestreo exógeno.

Se distinguen también varios casos según la información disponible sobre las distribuciones marginales $P(x)$ y $P(j)$.

La función de verosimilitud bajo muestreo basado en la elección es:

$$L(\theta) = \prod_{i=1}^{N} P((j_i, x_i), b) = \prod_{i=1}^{N} \frac{P(j_i / x_i, \theta) P(x_i)}{P(j_i)} H(j_i)$$

y el logaritmo de la función de verosimilitud:

$$LL(\theta) = \sum_{i=1}^{N} \ln P\left((j_i, x_i), b\right) =$$

$$\sum_{i=1}^{N} \left[\ln P(j_i / x_i, \theta) + \ln P(x_i) + \ln H(j_i) - \ln P(j_i) \right]$$

A continuación se analizan las soluciones específicas a cada situación informativa planteada.

a) $P(x)$ **y** $P(j)$ **conocidas**

La estimación máximo-verosímil del vector de parámetros θ es la solución al problema:

$$\max_{\theta \in \Theta} \sum_{i=1}^{N} \ln P(j_i / x_i, \theta)$$

teniendo en cuenta que

$$P(j) = \sum_{z \in Z} P(j/z, \theta) P(z)$$

Equivalentemente:

$$\max_{\theta \in \Theta_0} \sum_{i=1}^{N} \ln P(j_i / x_i, \theta)$$

donde

$$\Theta_0 = \left\{ \theta \in \Theta / P(j) = \sum_{z} P(j/z, \theta) P(z) \quad , j \in C \right\}$$

que coincide con (3-36).

b) $P(x)$ conocida y $P(j)$ desconocida

Ahora la distribución de probabilidad $P(x)$ no afectará a la maximización por ser conocida, pero $P(j)$ está relacionada con el parámetro desconocido ya que debe de verificar la relación

$$P(j) = \sum_{z \in Z} P(j/z, \theta) P(z)$$

y por lo tanto la estimación máximo-verosímil del vector de parámetros θ se obtendrá como la solución a:

$$\max_{\theta \in \Theta} \left[\sum_{i=1}^{N} \ln P(j_i / x_i, \theta) - \sum_{i=1}^{N} \ln \left\{ \sum_{Z} P(j_i / z, \theta) P(z) \right\} \right] \qquad (3\text{-}37)$$

c) $P(x)$ desconocida y $P(j)$ conocida

En este caso no se conoce la distribución marginal de las características observables, $P(x)$, pero sí que se dispone de información respecto a la distribución de probabilidad de las alternativas.

La estimación máximo-verosímil del vector de parámetros es la solución al problema:

$$\max_{\theta \in \Theta} \sum_{i=1}^{N} \left[\ln P(j_i / x_i, \theta) P(x_i) - \ln \sum_{Z} P(j_i / z, \theta) P(z) \right]$$

sujeto a

$$P(j) = \sum_{z \in Z} P(j/z, \theta) P(z)$$

Como $P(z)$ es desconocida será necesario encontrar una estimación de dicha distribución.

Cuando la distribución de probabilidad $P(x)$ de las características observables, esté caracterizada por un vector finito de parámetros desconocidos, es decir,

cuando se conozca la familia a la que pertenece, pero se desconocen los parámetros que la identifican, se estimará conjuntamente con el vector paramétrico desconocido. Es decir, hay que resolver el siguiente problema de maximización:

$$\max_{(\theta,\tilde{P})\in\Theta\times\wp_0} \sum_{i=1}^{N}\left[\ln \mathrm{P}(j_i/x_i,\theta)\,\tilde{P}(x_i)-\ln\sum_{z}\mathrm{P}(j_i/z,\theta)\,\tilde{P}(z)\right]$$

sujeto a

$$\mathrm{P}(j)=\sum_{z\in Z}\mathrm{P}(j/z,\theta)\,\tilde{P}(z)$$

donde $\wp_0$ es el espacio, de dimensión finita, de distribuciones de probabilidad.

Si por el contrario no puede asumirse ninguna familia de distribuciones para esta distribución marginal, $\mathrm{P}(x)$, y no se dispone de un espacio de distribuciones de dimensión finita, las soluciones al problema de maximización no tienen porque existir ni ser computacionalmente tratables, e incluso en el caso de obtener los estimadores correspondientes puede que no gocen de las propiedades asintóticas convenientes.

Para esta situación, en la literatura de los modelos de respuesta cualitativa se han propuesto varias soluciones alternativas, buscando siempre la obtención de la estimación del vector de parámetros desconocidos θ pero sin implicar la distribución marginal de las características $\mathrm{P}(x)$.

A continuación se van a presentar algunas de estas soluciones.

1. La primera solución fue la aportada por Manski y Lerman (1977), quienes en su artículo proponen utilizar el estimador máximo-verosímil bajo muestreo exógeno ponderado por los pesos w_i definidos como:

$$w(j)=\frac{\mathrm{P}(j)}{H(j)}$$

es decir, estos autores proponen como estimación del vector de parámetros θ la solución al problema

$$\max_{\theta\in\Theta}\sum_{i=1}^{N} w(j_i)\ln \mathrm{P}(j_i/x_i,\theta) \qquad (3\text{-}38)$$

Aunque este estimador esté construido para el muestreo exógeno, Manski y Lerman demostraron (ver apéndice A) que bajo las condiciones usuales de regularidad este estimador es fuertemente consistente y asintóticamente Normal en el caso de utilizar muestreo basado en la elección con $P(j)$ conocida y $P(x)$ desconocida.

2. Basándose en que la función de distribución empírica es la estimación máximo-verosímil de una función de distribución cualquiera, Cosslett (1981(a)) plantea cambiar la distribución de probabilidad $P(x)$ por los pesos w_i asociados a cada uno de los valores x_i observados. De esta forma se ha discretizado el espacio Z y el logaritmo de la función de verosimilitud será ahora:

$$LL(\theta) = \sum_{i=1}^{N} \ln P(j_i / x_i, \theta) + \sum_{i=1}^{N} \ln w_i + \sum_{i=1}^{N} \ln P(j_i) - \sum_{i=1}^{N} \ln H(j_i)$$

y la estimación máximo-verosímil es la solución al problema:

$$\max_{(\theta, w) \in \Theta \times W} \left[\sum_{i=1}^{N} \ln P(j_i / x_i, \theta) + \sum_{i=1}^{N} \ln w_i \right]$$

sujeto a la restricción

$$P(j) = \sum_{i=1}^{N} P(j / x_i, \theta) w_i$$

donde $W = \left\{ w: w_i \geq 0, \sum_{i=1}^{N} w_i = 1 \right\}$

Este problema de optimización puede transformarse en otro equivalente utilizando la función Lagrangiana y el dual, y el problema de obtener la estimación máximo-verosímil se reduce a encontrar $\hat{\theta}_N$ y $\hat{\lambda}_N$ tales que:

$$\mathcal{L}^2(\hat{\theta}_N, \hat{\lambda}_N) = \max_{\theta \in \Theta} \left\{ \min_{\lambda \in \Delta_2} \mathcal{L}^2(\theta, \lambda) \right\} \qquad (3\text{-}39)$$

siendo

$$\mathcal{L}^2(\theta,\lambda) = \sum_{i=1}^{N} \ln \frac{P(j_i / x_i, \theta)}{\sum_C \lambda(k) P(k / x_i, \theta)}$$

y

$$\Delta_2 = \left\{ \lambda / \sum_C \lambda(k) P(k) = 1, \quad \sum_C \lambda(k) P(k / x_i, \theta) > 0, i = 1, ..., N \right\}$$

Cosslett (1981a) hace todo el desarrollo de estas equivalencias y demuestra la consistencia y Normalidad asintótica de su estimador.

3. Para obtener un tercer estimador del vector de parámetros θ que no implique la distribución marginal de las características, $P(x)$, se considera la siguiente igualdad sobre las distribuciones condicionadas:

$$P(x) = \sum_C P(j) P(x / j)$$

donde $P(x / j)$ es la distribución de las características condicionada a la alternativa j y se considera que esta distribución condicionada es desconocida por serlo $P(x)$.

Nótese que en el caso de ser una distribución conocida se podría resolver el problema de conocer el vector de parámetros desconocidos sin utilizar datos muestrales, ya que sería la solución a las ecuaciones:

$$P(j / x, \theta) = \frac{P(j) P(x / j)}{\sum_C P(k) P(x / k)} \quad para \quad (j,x) \in C \times Z$$

Carroll y Relles (1976) suponen que esta distribución es una Normal multivariante, y se limitan a estimar sus parámetros. Sin embargo, McFadden (1976) razona que no es adecuado en general considerar $P(x / j)$ como una Normal multivariante. Por lo tanto se asumirá que esta distribución es totalmente desconocida.

En un muestreo basado en la elección se observan pares (k,x) extraídos aleatoriamente de la subpoblación $\{k\} \times Z$ con distribución $P(x / k)$. La

probabilidad de elección, $P(j/x,\theta)$ puede verse como una transformación de pares de la forma (j,x) y $\sum_z P(j/x,\theta)P(x/k)$ puede considerarse como la esperanza de $P(j/x,\theta)$.

Por otra parte, sea $N(k)$ el subconjunto de individuos de la muestra que eligen la alternativa k y N_k el cardinal del subconjunto.

Desde la muestra de observaciones (j,x_i) cuando los valores de las características x_i se obtienen según la distribución $P(x/k)$ y considerando la transformación $P(j/x_i,\theta)$, puede calcularse la media muestral de las observaciones independientes $P(j/x_i,\theta)$ que viene dada por la expresión:

$$\frac{1}{N_k}\sum_{m\in N(k)} P(j/x_m,\theta)$$

Por la ley fuerte de los grandes números la media muestral converge a la media poblacional, de donde

$$\frac{1}{N_k}\sum_{m\in N(k)} P(j/x_m,\theta) \xrightarrow{c.s.} \sum_z P(j/x,\theta)P(x/k)$$

y por tanto

$$\sum_{k\in C}\frac{P(k)}{N_k}\sum_{m\in N(k)} P(j/x_m,\theta) \xrightarrow{c.s.} \sum_{k\in C} P(k)\sum_z P(j/x,\theta)P(x/k)$$

Operando en la última expresión se obtiene el siguiente resultado:

$$\sum_{k\in C} P(k)\sum_z P(j/x,\theta)P(x/k) = \sum_z P(j/x,\theta)P(x) = P(j)$$

es decir

$$\sum_{k\in C}\frac{P(k)}{N_k}\sum_{m\in N(k)} P(j/x_m,\theta) \xrightarrow{c.s.} P(j)$$

Desde esa relación se pueden obtener dos estimadores para el vector de parámetros desconocidos. Como $P(j)$ es conocida y su expresión no está

relacionada con el primer término del problema de maximización el primer estimador se obtiene como solución al problema:

$$\max_{\theta\in\Theta} \sum_{i=1}^{N} \ln P(j/x_i,\theta)$$

sujeto a

$$P(j) = \sum_{k\in C} \frac{P(k)}{N_k} \sum_{m\in N(k)} P(j/x_m,\theta) \qquad j \in C \tag{3-40}$$

Esta maximización puede reformularse como un problema de Lagrange:

$$\max_{\theta\in\Theta} \min_{\lambda\in R^J} \sum_{i=1}^{N} \left[\ln P(j_i/x_i,\theta) - \lambda(j_i) \ln \left[\sum_{C} \frac{P(k)}{N_k P(j_i)} \sum_{m\in N(k)} P(j_i/x_m,\theta) \right] \right] \tag{3-41}$$

Para obtener el segundo estimador se considera también la aproximación anterior para la probabilidad de la alternativa:

$$P(j) = \sum_{k\in C} \frac{P(k)}{N_k} \sum_{m\in N(k)} P(j/x_m,\theta)$$

Aunque la distribución marginal $P(j)$ sea conocida al considerar la aproximación anterior puede verse como desconocida y el estimador viene dado por:

$$\max_{\theta\in\Theta} \sum_{i=1}^{N} \left[\ln P(j_i/x_i,\theta) - \ln \sum_{k\in C} \frac{P(k)}{N_k} \sum_{m\in N(k)} P(j_i/x_m,\theta) \right] \tag{3-42}$$

Este criterio puede reescribirse como el problema de Lagrange anterior (3-41) tomando como valor fijo de $\lambda(j)$ la unidad.

4. Una cuarta solución para estimar el vector de parámetros θ sin involucrar a la distribución de las características $P(x)$, se obtiene al considerar para cada par observado (j,x) de la muestra, el vector de características x fijo y calcular lo verosímil que es observar la alternativa j frente al resto de alternativas:

$$P(j/x) = \frac{P(j,x)}{\sum_{C} P(k,x)}$$

El muestreo es el basado en la elección, por lo que la verosimilitud de la observación (j,x) de la muestra viene dada por la siguiente expresión:

$$P(j,x) = \frac{P(j/x,\theta)P(x)H(j)}{P(j)}$$

y así la verosimilitud de observar la alternativa j condicionada al valor x del vector de características será:

$$P(j/x) = \frac{P(j/x,\theta)P(x)^{H(j)/P(j)}}{\sum_{C} \frac{P(k/x,\theta)P(x)H(k)}{P(k)}} = \frac{P(j/x,\theta)^{H(j)/P(j)}}{\sum_{C} P(k/x,\theta)^{H(k)/P(k)}}$$

Esta última expresión no depende de la distribución $P(x)$.

Se puede calcular la estimación máximo-verosímil del vector de parámetros desconocidos θ utilizando la verosimilitud condicionada anterior. Dicha estimación será la solución al problema:

$$\max_{\theta\in\Theta} \sum_{i=1}^{N} \frac{P(j_i/x_i,\theta)^{H(j_i)/P(j_i)}}{\sum_{C} P(k/x_i,\theta)^{H(k)/P(k)}} \qquad (3\text{-}43)$$

d) $P(x)$ y $P(j)$ ambas desconocidas

En esta situación, la estimación máximo-verosímil del vector de parámetros desconocidos θ se obtiene como la solución al problema:

$$\max_{\theta\in\Theta} \sum_{i=1}^{N} \left[\ln P(j_i/x_i,\theta) P(x_i) - \ln \sum_{Z} P(j_i/z,\theta) P(z) + \ln H(j_i) \right]$$

o equivalentemente

$$\max_{\theta\in\Theta}\sum_{i=1}^{N}\left[\ln P(j_i/x_i,\theta)P(x_i)-\ln\sum_{z}P(j_i/z,\theta)P(z)\right]$$

Al igual que en el caso anterior, como la distribución marginal de las características, $P(x)$, es desconocida será necesario estimar dicha distribución. Así, si $P(x)$ está caracterizada por un vector finito de parámetros desconocidos, el estimador conjunto viene dado como solución a:

$$\max_{(\theta,\tilde{P})\in\Theta\times\wp_0}\sum_{i=1}^{N}\left[\ln P(j_i/x_i,\theta)\,\tilde{P}(x_i)-\ln\sum_{z}P(j_i/z,\theta)\,\tilde{P}(z)\right]$$

donde $\wp_0$ es el espacio, de dimensión finita, de distribuciones de probabilidad de las características.

Es fácil observar que la solución planteada en este caso no es más que la del apartado anterior pero considerando que ahora también $P(j)$ es desconocida.

Recordando las soluciones obtenidas cuando $P(x)$ es desconocida y $P(j)$ conocida pueden plantearse las mismas para la situación en la que $P(j)$ también es desconocida. Bastará una pequeña modificación sobre las soluciones ya conocidas para adecuarlas a la situación actual.

La primera de las soluciones del caso anterior es la ya comentada, que resuelve el problema en el caso de que la distribución marginal $P(x)$ está restringida a un espacio de dimensión finita.

Cuando esta distribución marginal no esté restringida a un espacio de dimensión finita deben de buscarse soluciones alternativas que no impliquen dicha distribución. Se van a plantear las mismas soluciones que en el caso anterior adecuadamente modificadas para reflejar el desconocimiento de la distribución marginal $P(j)$, aunque alguna lleve a estimadores inconsistentes (ver apéndice A para la demostración).

1. La primera solución que no implicaba la distribución de probabilidad marginal $P(x)$ era la propuesta por Manski y Lerman. Aplicándola a esta situación, puesto que $P(j)$ es también desconocida, se tendrá que la estimación del vector de parámetros θ es la solución al problema:

$$\max_{(\theta,\ddot{P})\in\Theta\times\Pi} \frac{1}{N}\sum_{i=1}^{N} \frac{\ddot{P}(j_i)}{H(j_i)} \ln P(j_i/x_i,\theta) \qquad (3\text{-}44)$$

donde

$$\Pi = \left\{ \ddot{P} / \sum_{C} \ddot{P}(j) = 1 \right\}$$

En este caso hay que estimar también la distribución marginal de las alternativas y se considera como un conjunto de parámetros libres.

2. El estimador propuesto por Cosslett, que implicaba considerar una discretización del espacio de las características Z, modificado para la situación en la que ambas distribuciones son desconocidas, es el que aparece como la solución al problema:

$$\max_{(\theta,w)\in\Theta\times W} \left[\sum_{i=1}^{N} \ln P(j_i/x_i,\theta) + \sum_{i=1}^{N} \ln w_i - \sum_{i=1}^{N} \ln \sum_{k=1}^{N} w_k P(j_i/x_k,\theta) \right]$$

siendo w_i los pesos asociados a los valores de x observados en la muestra.

Igual que en la situación en la que $P(j)$ era conocida, esta solución se puede replantear como un problema de Lagrange, y a través del problema dual, Cosslett (1981a) demuestra que el problema anterior puede resolverse equivalentemente y de una forma más sencilla si se considera la siguiente función de pseudoverosimilitud:

$$LL(\theta,\lambda) = \sum_{i=1}^{N} \ln \frac{\lambda(j_i)P(j_i/x_i,\theta)}{\sum_{j=1}^{J} \lambda(j)P(j/x_i,\theta)}$$

ya que ahora el problema de encontrar la estimación máximo-verosímil del vector de parámetros desconocidos se reduce al problema de encontrar $\hat{\theta}_N$ y $\hat{\lambda}_N$, tal que:

$$LL(\hat{\theta}_N, \hat{\lambda}_N) = \max_{\theta\in\Theta} \max_{\lambda\in\Delta_0} LL(\theta,\lambda) \qquad (3\text{-}45)$$

siendo

$$\Delta_\theta = \left\{ \lambda / \lambda(k) \geq 0 \quad y \quad \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sum_{C} \lambda(k) \mathrm{P}(k / x_i, \theta)} = 1 \right\}$$

3. Otro estimador aparece al considerar la tercera solución del problema de estimar el vector de parámetros desconocidos θ cuando $\mathrm{P}(x)$ es desconocido y $\mathrm{P}(j)$ conocida. En este caso, al ser desconocida la distribución marginal de las alternativas también debe de ser estimada, y considerándola como un conjunto de parámetros libres, la estimación máximo-verosímil será la solución al problema:

$$\max_{(\theta, \ddot{P}) \in \Theta \times \Pi} \sum_{i=1}^{N} \left[\frac{1}{N} \ln \mathrm{P}(j_i / x_i, \theta) - \frac{1}{N} \ln \left[\sum_{k \in C} \frac{\ddot{\mathrm{P}}(k)}{N_k} \sum_{m \in N(k)} \mathrm{P}(j_i / x_m, \theta) \right] \right] \tag{3-46}$$

4. El estimador que se va a presentar ahora es el que aparece al modificar el último estimador de la situación anterior para que tenga en cuenta el desconocimiento de la distribución marginal de las alternativas, $\mathrm{P}(j)$. La estimación máximo-verosímil del vector de parámetros desconocidos y la de la distribución marginal $\mathrm{P}(j)$ se obtiene como solución al problema:

$$\max_{(\theta, \ddot{P}) \in \Theta \times \Pi} \left[\frac{1}{N} \sum_{i=1}^{N} \ln \frac{\mathrm{P}(j_i / x_i, \theta)^{H(j_i) / \ddot{\mathrm{P}}(j_i)}}{\sum_{k \in C} \mathrm{P}(k / x_i, \theta)^{H(k) / \ddot{\mathrm{P}}(k)}} \right] \tag{3-47}$$

Observar que este estimador es un caso particular del propuesto por Cosslett para ambas distribuciones desconocidas. Cosslett demostró que los parámetros $\lambda(k)$ que intervienen en su estimador convergen a la razón $H(j) / \mathrm{P}(j)$, adoptando por lo tanto la misma expresión considerada ahora.

Este estimador (3-47) y el de Cosslett, (3-45), son los únicos que son consistentes. Ni el de Manski y Lerman, (3-44), ni la tercera solución, (3-46), verifican esta propiedad. En el apéndice A se encuentran todas estas demostraciones.

3.4. Estimación en modelos de variable dependiente limitada

En este epígrafe se van a presentar las estimaciones correspondientes al vector de parámetros asociados a los modelos de variable dependiente limitada.

Todos los modelos de variable dependiente limitada que se han planteado en el epígrafe 3.2. presentan la misma estructura y el método de estimación que se propone aquí es aplicable a todos ellos.

Por este motivo sólo se desarrollará el proceso de estimación para el modelo tobit y para los modelos de variable continua con separación muestral.

3.4.1. Estimación en el modelo Tobit

El modelo tobit estandar definido por (3-21) mediante una variable continua $y_i^* = x_i'\beta + \varepsilon_i$ observable únicamente cuando $y_i^* > 0$ estará completamente identificado al conocer el vector de parámetros β y la varianza σ^2 de la variable aleatoria ε_i para la que se asume una distribución Normal, $N[0, \sigma^2]$.

Amemiya (1984) en su artículo propone métodos alternativos para estimar los parámetros desconocidos. El primero se basa en la función de verosimilitud, que en el modelo tobit viene dada por:

$$L(\theta) = \prod_{i/y_i^* \le 0} \left[1 - \Phi(x_i'\beta/\sigma)\right] \prod_{i/y_i^* > 0} \sigma^{-1} \phi\left((y_i - x_i'\beta)/\sigma\right)$$

donde $\prod_{i/y_i^* \le 0}$ representa el producto sobre aquellos individuos i para los cuales la variable dependiente es no positiva, $y_i^* \le 0$, y por lo tanto no observable, y el simbolo $\prod_{i/y_i^* > 0}$ representa el producto sobre los individuos i con $y_i^* > 0$.

Desde esta función de verosimilitud se buscará la estimación de los parámetros desconocidos mediante un proceso de maximización. No pueden estimarse separadamente β y σ, y se estima $\alpha = \frac{\beta}{\sigma}$ por maximizar esta función de

verosimilitud utilizando cualquiera de los métodos iterativos ya comentados, como el de Newton-Raphson o el método Scoring dada la no linealidad en los parámetros de las ecuaciones a resolver.

Para facilitar esta búsqueda del máximo, Amemiya propone considerar la siguiente expresión para la función de verosimilitud que es equivalente a la anterior:

$$L(\theta) = \prod_{i/y_i^* \leq 0} \left[1 - \Phi(x_i'\beta/\sigma)\right] \prod_{i/y_i^* > 0} \Phi(x_i'\beta/\sigma) \prod_{i/y_i^* > 0} \Phi^{-1}(x_i'\beta/\sigma)\,\sigma^{-1}\phi\left((y_i - x_i'\beta)/\sigma\right)$$

Los dos primeros factores de este producto constituyen la función de verosimilitud de un modelo probit binomial y se estima $\alpha = \frac{\beta}{\sigma}$ utilizando únicamente estos dos factores como se propone en el epígrafe 3.3 dedicado a los modelos de respuesta cualitativa. Esta alternativa simplifica el proceso de maximización de la función de verosimilitud, pero proporciona estimadores no eficientes (ver apéndice A) al no utilizar toda la información para obtener dichos estimadores.

Otra posibilidad para encontrar estimaciones de los parámetros desconocidos es utilizar el hecho de que la variable dependiente está construida como un modelo de regresión lineal y buscar la estimación por mínimos cuadrados ordinarios. En este caso se plantea la estimación del modelo lineal $y_i^* = x_i'\beta + \varepsilon_i$ pero utilizando como muestra únicamente aquellos individuos para los cuales se ha observado esta variable, es decir, para aquellos individuos que verifican que $y_i^* > 0$. No obstante, esta solución no es adecuada directamente, ya que el hecho de que la variable y_i^* sea observable únicamente cuando es positiva lleva a que se tenga la relación siguiente:

$$E[y_i^* / y_i^* > 0] = x_i'\beta + E[\varepsilon_i / \varepsilon_i > -x_i'\beta]$$

y el último término generalmente no es cero como requiere la teoría de la estimación por mínimos cuadrados, llevando a estimaciones inconsistentes.

Se puede demostrar fácilmente que bajo la hipótesis de Normalidad de la variable aleatoria ε_i se tiene que:

$$E[\varepsilon_i / \varepsilon_i > -x_i'\beta] = \sigma\,\lambda(x_i'\beta/\sigma)$$

siendo $\lambda(\cdot)=\dfrac{\phi(\cdot)}{\Phi(\cdot)}$ la inversa del ratio de Mills y está definida a partir de las funciones de densidad $\phi(\cdot)$ y de distribución $\Phi(\cdot)$ de una variable aleatoria con distribución $N[0,1]$.

Con esto se puede reescribir el modelo de regresión lineal como:

$$y_i^* = x_i'\beta+\sigma\,\lambda\left(x_i'\beta/\sigma\right)+\eta_i \qquad (3\text{-}48)$$

siendo η_i una variable aleatoria con distribución Normal de media cero.

Ahora se dispone de un modelo lineal que sí que cumple las condiciones usuales para la estimación por mínimos cuadrados ordinarios, ya que se ha eliminado el problema del sesgo de selección. Sin embargo antes de aplicar este método de estimación será necesario realizar la estimación del coeficiente $\lambda\left(x_i'\beta/\sigma\right)$ introducido para eliminar el sesgo.

Para obtener la estimación de los parámetros del modelo tobit se seguirá un proceso en dos etapas. En primer lugar se estima $\alpha=\dfrac{\beta}{\sigma}$ del modelo probit comentado antes por máxima-verosimilitud. Se sustituye $\alpha=\dfrac{\beta}{\sigma}$ por su estimación $\hat{\alpha}$ en la ecuación lineal (3-48) anterior:

$$y_i^* = x_i'\beta+\sigma\,\lambda\left(x_i'\hat{\alpha}\right)+\eta_i \qquad (3\text{-}49)$$

y se estiman β y σ por mínimos cuadrados (x_i' y $\lambda\left(x_i'\hat{\alpha}\right)$ serán conocidas) utilizando únicamente las observaciones positivas de la variable dependiente, es decir, $y_i^*>0$.

Con la inversa del ratio de Mills se consigue eliminar el sesgo introducido al estimar por mínimos cuadrados el modelo lineal utilizando en la estimación la submuestra formada por los individuos que poseen un valor $y_i^*>0$. Este procedimiento en dos etapas elimina la inconsistencia del estimador de mínimos cuadrados ordinarios obtenido sin corregir el sesgo de selección.

Se ha comentado antes la posible dificultad de encontrar la estimación máximo-verosímil en el modelo tobit. El problema está en encontrar buenos

valores iniciales para conseguir que el proceso de búsqueda del máximo de la función de verosimilitud sea eficiente.

Una solución a este problema es considerar las estimaciones obtenidas en el proceso en dos etapas como estimaciones iniciales del proceso de máxima-verosimilitud.

3.4.2. Estimación en el modelo de variable dependiente continua con separación muestral

En esta sección se planteará la estimación del vector de parámetros desconocidos para los modelos del tipo (3-24):

$$y_{1i} = x'_{1i}\beta_1 + \varepsilon_{1i} \quad sii \quad I_i = 1$$

$$y_{2i} = x'_{2i}\beta_2 + \varepsilon_{2i} \quad sii \quad I_i = 0$$

con

$$I_i = \begin{cases} 1 & sii \quad z'_i\gamma - \varepsilon_i > 0 \\ 0 & en\ otro\ caso \end{cases}$$

Estos modelos que reciben el nombre de continuos-discretos involucran una decisión discreta, representada por la tercera ecuación y una decisión continua, y_{1i} o y_{2i}, que dependerá de la decisión discreta tomada.

En este modelo el vector de parámetros θ está formado por los parámetros β_1, β_2 y γ de las ecuaciones lineales y por los parámetros necesarios para especificar las distribuciones de probabilidad de las variables aleatorias ε_{1i}, ε_{2i}, ε_i.

La estimación de este vector θ se podría realizar estimando las dos primeras ecuaciones y_{1i}, y_{2i} por mínimos cuadrados ordinarios utilizando, en la primera, una muestra formada por los individuos de la muestra total que realmente han tomado la elección $I_i = 1$, y para la segunda ecuación se utilizaría la submuestra de los individuos que han elegido la segunda alternativa, $I_i = 0$.

No obstante al igual que ocurría con el modelo tobit, si se realiza la estimación con este procedimiento se encuentran estimaciones de los parámetros

inconsistentes. La razón es que la decisión continua está condicionada por la elección discreta, y por tanto $E[\varepsilon_{1i} / I_i] \neq 0$ y $E[\varepsilon_{2i} / I_i] \neq 0$. Es decir, los valores medios de los errores de las ecuaciones de la variable continua no son cero como se requiere para la estimación por mínimos cuadrados.

Este problema origina el llamado sesgo de selección nombrado antes en el modelo tobit, y será necesario eliminarlo para obtener buenas estimaciones de los parámetros del modelo. Para resolver el problema se introducen las siguientes transformaciones:

$$\varepsilon_{1i} = E[\varepsilon_{1i} / I_i] + \eta_{1i} \quad \text{y} \quad \varepsilon_{2i} = E[\varepsilon_{2i} / I_i] + \eta_{2i}$$

siendo η_{1i} y η_{2i} dos perturbaciones aleatorias de media cero y que están incorrelacionadas con la ecuación discreta. De esta forma el sistema de ecuaciones continuas del modelo se transforma en el siguiente sistema equivalente:

$$y_{1i} = x'_{1i}\beta_1 + E[\varepsilon_{1i} / I_i = 1] + \eta_{1i}$$

$$y_{2i} = x'_{2i}\beta_2 + E[\varepsilon_{2i} / I_i = 0] + \eta_{2i}$$

Ambas ecuaciones pueden ser estimadas ahora por mínimos cuadrados ordinarios siempre que previamente se encuentre una estimación consistente del factor introducido para la corrección del sesgo de selección, $E[\varepsilon_{1i} / I_i = 1]$ y $E[\varepsilon_{2i} / I_i = 1]$.

Considerando diferentes hipótesis distribucionales sobre el modelo original se pueden derivar las expresiones de los términos de corrección de selectividad.

Dubin y McFadden (1984) han demostrado que si se utiliza un modelo logit binomial para la elección discreta y admitiendo que las variables ε_{1i}, ε_{2i} siguen una distribución Normal se obtienen las siguientes expresiones:

$$E[\varepsilon_{1i} / I_i = 1] = \frac{\sqrt{6\sigma^2}}{\pi}\left[\rho_{0\varepsilon}\frac{P_0 \ln P_0}{1-P_0} - \rho_{1\varepsilon}\ln P_1\right]$$

$$E[\varepsilon_{2i} / I_i = 0] = \frac{\sqrt{6\sigma^2}}{\pi}\left[\rho_{1\varepsilon}\frac{P_1 \ln P_1}{1-P_1} - \rho_{0\varepsilon}\ln P_0\right]$$

siendo $P_j = P(I_i = j / x_i)$, $j = 0, 1$, $\rho_{1\varepsilon}$ el coeficiente de correlación entre las variables ε_{1i} y ε_i y $\rho_{0\varepsilon}$ el coeficiente de correlación entre las variables ε_{2i} y ε_i.

Dado que las alternativas 1 y 0 son mutuamente excluyentes se podría razonar que $\rho_{0\varepsilon} = -\rho_{1\varepsilon}$ de donde:

$$\rho_{0\varepsilon}\frac{P_0 \ln P_0}{1-P_0} - \rho_{1\varepsilon}\ln P_1 = -\rho_{1\varepsilon}\left[\frac{P_0 \ln P_0}{1-P_0} + \ln P_1\right] = -\rho_{1\varepsilon}\, c_{1\varepsilon}$$

$$\rho_{1\varepsilon}\frac{P_1 \ln P_1}{1-P_1} - \rho_{0\varepsilon}\ln P_0 = -\rho_{0\varepsilon}\left[\frac{P_1 \ln P_1}{1-P_1} + \ln P_0\right] = -\rho_{0\varepsilon}\, c_{2\varepsilon}$$

y las ecuaciones originales podrán ahora escribirse como:

$$\begin{aligned} y_{1i} &= x'_{1i}\beta_1 + \lambda_1 c_{1\varepsilon} + \eta_{1i} \\ y_{2i} &= x'_{2i}\beta_2 + \lambda_2 c_{2\varepsilon} + \eta_{2i} \end{aligned} \qquad (3\text{-}50)$$

con $\lambda_1 = -\frac{\sqrt{6\sigma^2}}{\pi}\rho_{1\varepsilon}$ y $\lambda_2 = -\frac{\sqrt{6\sigma^2}}{\pi}\rho_{2\varepsilon}$

Para estimar el sistema completo se estimará en primer lugar el modelo logit binomial y las estimaciones obtenidas se utilizarán para calcular $\hat{c}_{1\varepsilon}$, $\hat{c}_{2\varepsilon}$. Sustituyendo estas estimaciones en las ecuaciones lineales (3-50), se obtendrán los siguientes modelos:

$$\begin{aligned} y_{1i} &= x'_{1i}\beta_2 + \lambda_1 \hat{c}_{1i} + \tilde{\eta}_{1i} \\ y_{2i} &= x'_{1i}\beta_2 + \lambda_2 \hat{c}_{2i} + \tilde{\eta}_{2i} \end{aligned} \qquad (3\text{-}51)$$

que se estimarán por mínimos cuadrados ordinarios utilizando las muestras separadas, para y_{1i} se considera la muestra de N_1 individuos que toman la elección 1 y para y_{2i} la muestra de N_2 individuos que toman la elección 0.

Lee y Trost (1978) obtuvieron las expresiones asociadas al factor corrector del sesgo de selección para cada ecuación asumiendo que el modelo de elección discreta era un modelo probit binomial y que los tres términos de error seguían

una distribución conjunta Normal con vector de medias $\vec{\mu} = (0,0,0)'$ y matriz de varianzas-covarianzas

$$\Omega = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{1\varepsilon} \\ \sigma_{12} & \sigma_2^2 & \sigma_{2\varepsilon} \\ \sigma_{1\varepsilon} & \sigma_{2\varepsilon} & \sigma_\varepsilon^2 \end{bmatrix}$$

Considerando, sin pérdida de generalidad $\sigma_\varepsilon^2 = 1$, las expresiones del corrector de la selectividad son:

$$E[\varepsilon_{1i} / I_i = 1] = -\sigma_{1\varepsilon} \frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)}$$

$$E[\varepsilon_{2i} / I_i = 1] = \sigma_{2\varepsilon} \frac{\phi(z_i'\gamma)}{1-\Phi(z_i'\gamma)}$$

Obteniéndose el siguiente sistema de ecuaciones:

$$\begin{aligned} y_{1i} &= x_{1i}'\beta_1 - \sigma_{1\varepsilon} \frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)} + \eta_{1i} = x_{1i}'\beta_1 + \lambda_1 c_{1i} + \eta_{1i} \\ y_{2i} &= x_{2i}'\beta_2 + \sigma_{2\varepsilon} \frac{\phi(z_i'\gamma)}{1-\Phi(z_i'\gamma)} + \eta_{2i} = x_{2i}'\beta_2 + \lambda_2 c_{2i} + \eta_{2i} \end{aligned} \qquad (3\text{-}52)$$

con

$$\lambda_1 = -\sigma_{1\varepsilon}, \quad \lambda_2 = \sigma_{2\varepsilon}, \quad c_{1i} = \frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)}, \quad c_{2i} = \frac{\phi(z_i'\gamma)}{1-\Phi(z_i'\gamma)}$$

El proceso de estimación propuesto por Lee y Trost consiste, igual que el ya comentado, en dos pasos:

1° estimar el modelo probit binomial

$$I_i = \begin{cases} 1 & sii \quad I_i^* = z_i'\gamma - \varepsilon_i \geq 0 \\ 0 & en\ otro\ caso \end{cases}$$

por máxima-verosimilitud.

2° sustituir en las ecuaciones continuas (3-52) la estimación $\hat{\gamma}$ obtenida en el paso anterior y estimar por mínimos cuadrados las nuevas ecuaciones:

$$y_{1i} = x'_{1i}\beta_1 + \lambda_{1i}\hat{c}_{1i} + \tilde{\eta}_{1i}$$
$$y_{2i} = x'_{2i}\beta_2 + \lambda_{2i}\hat{c}_{2i} + \tilde{\eta}_{2i} \qquad (3\text{-}53)$$

utilizando la submuestra de N_1 individuos que responden 1 en el modelo de elección discreta para estimar los parámetros asociados a la variable y_{1i} y la submuestra de individuos que responden 0 en el modelo probit para estimar los parámetros correspondientes a y_{2i}, siendo

$$\hat{c}_{1i} = \frac{\phi(z'_i\hat{\gamma})}{\Phi(z'_i\hat{\gamma})} \quad \text{y} \quad \hat{c}_{2i} = \frac{\phi(z'_i\hat{\gamma})}{1-\Phi(z'_i\hat{\gamma})}$$

De nuevo se plantea ahora la posibilidad de considerar las estimaciones mediante el proceso en dos etapas como valores iniciales para el método iterativo de maximizar la función de verosimilitud conjunta del modelo.

3.5. Contraste de hipótesis

En este trabajo se ha planteado un proceso inferencial acerca de la estructura de los modelos de elección discreta y de variable dependiente limitada. Sin embargo, este proceso no estará finalizado si no se realiza una constatación de la adecuación de los resultados obtenidos a los datos del problema analizado.

En los epígrafes 3.1 y 3.2 se ha realizado una revisión metodológica de los modelos de respuesta discreta así como de algunos de variable dependiente limitada. Todos estos modelos están identificados a partir de un vector de parámetros desconocidos, y en los epígrafes 3.3 y 3.4 se han desarrollado los procesos de estimación más usuales para el mismo.

Al plantear un modelo econométrico cualquiera y realizar la estimación de los parámetros correspondientes, el investigador siempre puede intentar averiguar si el modelo utilizado es adecuado al problema que analiza o si por el contrario existen errores de mala especificación, bien sea por omitir variables explicativas o por asumir una distribución de probabilidad para el modelo que no es adecuada a los datos.

Una posible respuesta a estas cuestiones implicaría utilizar medidas de bondad de ajuste que indiquen si el modelo utilizado es coherente con las observaciones. Aunque ésta es una solución aceptable no hay que olvidar que las medidas de bondad de ajuste, al igual que el R^2 de un modelo de regresión lineal, dan una medida del grado de adecuación del modelo a los datos del problema indicando si el modelo es adecuado o no mediante un valor numérico. Pero en caso de resultar un modelo poco adecuado, estas medidas no permiten encontrar cual es el fallo cometido, y en muchas ocasiones estas medidas por si solas no permiten sacar conclusiones acerca de la bondad del ajuste y es necesario hacer comparaciones para elegir el mejor modelo.

Para resolver estos problemas la literatura estadística ofrece un amplio abanico de posibilidades con los denominados contrastes de hipótesis. Se plantearán contrastes de hipótesis que puedan indicar si el modelo es correcto o no, y en el caso de no serlo, se podrá concluir si el problema es la no significatividad de algún parámetro o una mala especificación del modelo, bien sea por variables omitidas o por suponer una distribución de probabilidad que no es adecuada.

Cuando se plantea un contraste de hipótesis se deben especificar correctamente las dos hipótesis (nula y alternativa). En un contraste de hipótesis el rechazo de la hipótesis nula lleva a la conclusión de que la hipótesis alternativa es correcta y esta afirmación puede no ser cierta. El problema es tener que escoger entre dos alternativas pero decidiendo sólo sobre una de ellas.

Supóngase que en la hipótesis nula se plantea que el modelo está especificado correctamente, mientras que en la alternativa se plantea un modelo más general de la misma familia. Si la hipótesis nula se rechaza tal vez no sea debido a una mala especificación distribucional, sino a un problema de variables omitidas. En este caso, tampoco la hipótesis alternativa sería la correcta, pero al realizar el contraste, los datos observados han llevado a la conclusión de que es preferible cambiar de modelo antes que mantenerlo con variables omitidas.

El contraste de hipótesis es la principal herramienta de trabajo cuando se pretende confrontar la teoría con los fenómenos observables. Para poder realizar esta constatación será necesario que los modelos teóricos sean reducidos a hipótesis contrastables.

La gran complejidad de la mayoría de los modelos estadísticos lleva a que el método de estimación más utilizado sea el de la máxima-verosimilitud, y también en el contraste de hipótesis esta función va a jugar un importante papel en la definición de los tests correspondientes.

Una gran parte de los contrastes analizados se resuelven mediante los tests clásicos: el de Wald, el test de la razón de verosimilitudes o el test de los multiplicadores de Lagrange. Los tres tienen cierta similitud en su planteamiento, aunque la forma de razonar sea diferente. Mientras el test de los multiplicadores de Lagrange empieza en la hipótesis nula y mira hacia la alternativa buscando una mejora, el test de Wald lo hace al revés y el de la razón de verosimilitudes compara las dos hipótesis directamente.

3.5.1. Test de Wald, razón de verosimilitudes y multiplicadores de Lagrange

En este apartado se van a presentar las expresiones generales de los tests de Wald, la razón de verosimilitudes y los multiplicadores de Lagrange. En todo el desarrollo se asumirá que la función de verosimilitud satisface las condiciones usuales de regularidad y que la matriz de información es no singular, así como que los parámetros son identificables.

Hipótesis nula simple

El problema de contraste más simple es el que asume bajo la hipótesis nula un único valor del parámetro desconocido, de forma que los datos estarán generados por la distribución de probabilidad $f(y,\theta_0)$ y para la hipótesis alternativa se considera una hipótesis compuesta, que vendrá formalizada por la distribución de probabilidad $f(y,\theta)$, $\theta \in R^k$.

Para resolver este contraste de hipótesis nula simple e hipótesis alternativa compuesta se utiliza la función de verosimilitud, que para una muestra aleatoria simple de tamaño N viene definida por:

$$L(\theta) = \prod_{i=1}^{N} f(y_i,\theta)$$

Sea $LL(\theta)$ el logaritmo de la función de verosimilitud a partir del cual se puede calcular el estimador máximo-verosímil. Sea $\hat{\theta}$ el valor que maximiza el logaritmo de la función de verosimilitud y como tal verifica que

$$\frac{d}{d\theta} LL(\theta)\Big|_{\theta=\hat{\theta}} = 0$$

La varianza del estimador se calcula como la inversa de la matriz de información de Fisher:

$$V\left[\hat{\theta}\right] = \frac{1}{N} J^{-1}(\theta) = \frac{1}{N}\left[-E\left[\frac{d^2 LL(\theta)}{d\theta\, d\theta'}\right]\frac{1}{N}\right]^{-1}$$

Para resolver el contraste de hipótesis nula H_0:$\theta = \theta_0$ y de hipótesis alternativa H_1:$\theta \neq \theta_0$ $\theta \in R^k$ el estadístico de Wald viene definido como:

$$W = N\left(\hat{\theta} - \theta_0\right)' J\left(\hat{\theta}\right)\left(\hat{\theta} - \theta_0\right) \qquad (3\text{-}54)$$

siendo $J\left(\hat{\theta}\right)$ una estimación consistente de la matriz de información $J(\theta)$.

Si el estimador máximo-verosímil $\hat{\theta}$ es asintóticamente Normal, el estadístico de Wald, W, sigue una distribución asintótica χ^2 con los grados de libertad igual a la dimensión del parámetro, k, cuando la hipótesis nula es cierta.

El test asociado con el estadístico de Wald no es más que la aproximación asintótica de los usuales tests de la t y de la F en modelos lineales. Evidentemente la región crítica viene definida por aquellas muestras para las cuales el valor del estadístico W supera al correspondiente valor de la distribución χ_k^2 para un nivel de significación α prefijado.

El segundo de los tests es el de la razón de verosimilitudes que se basa en una comparación entre el máximo de la función de verosimilitud bajo la hipótesis nula y el correspondiente máximo bajo la hipótesis alternativa. Este test se define a partir de la diferencia de los logaritmos de la función de verosimilitud, es decir se basa en el estadístico:

$$LR = -2\left[\mathrm{LL}(\theta_0) - \mathrm{LL}\left(\hat{\theta}\right)\right] \tag{3-55}$$

que bajo condiciones generales sigue una distribución asintótica χ_k^2 cuando la hipótesis nula es cierta.

El test de la razón de verosimilitudes rechaza la hipótesis nula H_0 cuando el estadístico LR es superior al cuantil α de una distribución χ_k^2.

El tercer test viene definido por el principio de los multiplicadores de Lagrange y se basa en la derivada o "score" del logaritmo de la función de verosimilitud que se denota como $s(\theta)$.

El test de los multiplicadores de Lagrange se deriva de un principio de maximización condicionada. Maximizando la función de verosimilitud sujeta a la condición que especifica la hipótesis nula, $\theta = \theta_0$, la función lagrangiana es:

$$\mathrm{LL}(\theta) - \lambda'(\theta - \theta_0)$$

donde λ es el conjunto de los multiplicadores de Lagrange y se puede interpretar como el coste de la restricción. Si el coste es elevado, la condición considerada, $\theta = \theta_0$, será rechazada como inconsistente con los datos.

Para buscar el máximo se calculan las condiciones de primer orden obteniéndose que

$$\frac{d}{d\theta}\mathrm{LL}(\theta) = \lambda \qquad \text{y} \qquad \theta = \theta_0$$

así que $\lambda = s(\theta_0)$.

El estadístico de los multiplicadores de Lagrange se define a partir de estos resultados como:

$$LM = \frac{1}{N} s(\theta_0)' \, J^{-1}(\hat{\theta}) \, s(\theta_0) \qquad (3\text{-}56)$$

Aplicando el Teorema Central del Límite a los "scores" se puede afirmar que la distribución de probabilidad del estadístico LM es asintóticamente χ^2 con k grados de libertad bajo la hipótesis nula.

De nuevo se tendrá que el test rechaza la hipótesis nula cuando el valor del estadístico es superior al cuantil de una χ^2_k, para un nivel de significación α.

Los tres estadísticos que se han comentado, aunque son diferentes entre sí, están basados en el mismo principio: medir la distancia entre las dos hipótesis (nula y alternativa).

El test que genera el estadístico de Wald se basa en la diferencia entre el estimador máximo-verosímil y el valor del parámetro bajo la hipótesis nula, es decir, se define a partir de la diferencia $\hat{\theta} - \theta_0$.

El estadístico de la razón de verosimilitudes define un test que está formulado en términos de la diferencia entre los valores del logaritmo de la función de verosimilitud para la hipótesis nula y para la alternativa, es decir, se basa en la diferencia $\mathrm{LL}(\theta_0) - \mathrm{LL}(\hat{\theta})$.

Por el contrario, el test de los multiplicadores de Lagrange viene definido a partir de la derivada del logaritmo de la función de verosimilitud, $s(\theta_0)$.

Aunque, en general, los tres estadísticos están definidos de forma diferente, Engle (1984) demuestra que bajo ciertas condiciones de la función de verosimilitud (su logaritmo es, aproximadamente, una forma cuadrática), estos tres estadísticos son equivalentes.

La elección entre ellos para un problema concreto vendrá en función de las características particulares del mismo, ya que la única diferencia entre los tres tests está en términos computacionales.

Hipótesis nula compuesta

Si el problema de contraste que se pretende resolver tiene por hipótesis nula una hipótesis compuesta que establece determinadas relaciones entre los parámetros desconocidos, los tres estadísticos anteriores siguen siendo adecuados.

Supóngase la siguiente descomposición del vector de parámetros $\theta' = (\theta_1', \theta_2')$ y que la hipótesis nula únicamente hace referencia al subvector θ_1 dejando el subvector θ_2 sin restringir, $H_0: \theta_1 = \theta_1^0$.

Sea $\tilde{\theta}_2$ la estimación máximo-verosímil del subvector θ_2 bajo la hipótesis nula y sea $\tilde{\theta}' = (\theta_1^0, \tilde{\theta}_2')$.

Para definir el estadístico de Wald únicamente se necesita considerar una partición de la matriz de información J, de forma que J^{11} es la inversa particionada de J tal que:

$$(J^{11})^{-1} = J_{11} - J_{12}\, J_{22}^{-1}\, J_{21}$$

El test de Wald, aplicando (3-54) únicamente al subvector θ_1, es sencillamente

$$W = N\left(\hat{\theta}_1 - \theta_1^0\right)' (J^{11})^{-1} \left(\hat{\theta}_1 - \theta_1^0\right)$$

que tiene una distribución asintótica χ^2 con k_1 grados de libertad bajo H_0, donde k_1 es la dimensión de θ_1.

Por otro lado el test de la razón de verosimilitudes está basado en la diferencia entre el máximo de la función de verosimilitud en la hipótesis nula y el máximo no restringido, es decir, en el estadístico dado por (3-55):

$$LR = -2\left[\mathrm{LL}(\tilde{\theta}) - \mathrm{LL}(\hat{\theta})\right]$$

y, al igual que en el caso anterior, mantiene la misma distribución de probabilidad que el estadístico de Wald.

Para obtener el test que genera el estadístico de los multiplicadores de Lagrange es necesario, al igual que antes, definir la función Lagrangiana y calcular las condiciones de primer orden.

Para maximizar la función de verosimilitud hay que considerar la restricción que impone la hipótesis nula y en consecuencia la Lagrangiana es:

$$\mathrm{LL}(\theta) - \lambda'\left(\theta_1 - \theta_1^{0}\right)$$

Las condiciones de primer orden serán análogas a las del caso anterior, pero dividiendo el vector de parámetros, tal que:

$$\frac{d}{d\theta_1}\mathrm{LL}(\theta) = \lambda \qquad \text{y} \qquad \frac{d}{d\theta_2}\mathrm{LL}(\theta) = 0$$

Por lo tanto, se obtendrá que $\theta_1 = \theta_1^{0}$ y el estadístico de los multiplicadores de Lagrange vendrá definido desde (3-56) como:

$$LM = \frac{1}{N} s(\tilde{\theta})' J^{-1}(\tilde{\theta}) s(\tilde{\theta})$$

Recordando que $\tilde{\theta}' = \left(\theta_1^{0}, \tilde{\theta}_2'\right)$ y que la derivada del logaritmo de la función de verosimilitud, según las condiciones de primer orden, vale cero en el subvector $\tilde{\theta}_2$ por ser la estimación máximo-verosímil de θ_2, este estadístico puede reescribirse como:

$$LM = \frac{1}{N} s_1(\tilde{\theta})' J^{11} s_1(\tilde{\theta})$$

siendo $s_1(\tilde{\theta})$ el subvector correspondiente a la derivada del logaritmo de la función de verosimilitud para el subvector θ_1.

De nuevo se tiene que el estadístico de los multiplicadores de Lagrange LM sigue una distribución asintótica χ^2 con k_1 grados de libertad cuando la hipótesis nula es cierta.

Al igual que ocurre cuando la hipótesis nula es simple, también ahora los tres estadísticos anteriores llevan al mismo test cuando la función de verosimilitud cumple que su logaritmo es una forma cuadrática.

3.5.2. Test equivalente al de los multiplicadores de Lagrange para los modelos de respuesta cualitativa binomiales

El contraste que se desea realizar es si se han omitido variables, para lo cual se estima el modelo considerando únicamente las variables x_{i2}.

Particionando el vector de parámetros $\beta = (\beta_1', \beta_2')'$ y las variables explicativas consecuentemente a esta partición, se desea contrastar la hipótesis nula $H_0 : \beta_1 = 0$.

En el modelo de elección binaria, la respuesta está medida por una variable dependiente y que toma el valor 1 con probabilidad p y el valor 0 con probabilidad $1-p$. Para cada observación estas probabilidades son diferentes bien sea por la propia naturaleza de la elección o bien por el propio decisor, ya que para un individuo i , su probabilidad de respuesta viene dada por la relación siguiente: $P_i = F(x_i'\beta)$, donde x_i son las características observadas sobre el individuo y las alternativas de elección.

Esta función $F(\cdot)$ es una función de distribución. La función de verosimilitud, para una muestra aleatoria simple, viene dada por:

$$L(\beta) = \prod_{i=1}^{N} F(x_i'\beta)^{y_i} \left(1 - F(x_i'\beta)\right)^{1-y_i}$$

y su logaritmo

$$LL(\beta) = \sum_{i=1}^{N} \left[y_i \ln F(x_i'\beta) + (1-y_i) \ln\left(1 - F(x_i'\beta)\right) \right]$$

Se denotará por $\tilde{\beta}_2$ a las estimaciones de β_2 bajo la hipótesis nula y las probabilidades calculadas con estas estimaciones se denotarán como $\tilde{P}_i$.

El "score" del modelo viene dado por:

$$\frac{d}{d\beta} LL(\beta) = \sum_{i=1}^{N} \frac{y_i - P_i}{P_i(1-P_i)} f(x_i'\beta)\, x_i'$$

que es una función de los residuos $y_i - P_i$ y la matriz de información será:

$$J(\beta) = E\left[\frac{d}{d\beta} LL(\beta) \frac{d}{d\beta'} LL(\beta) \right] = \sum_{i=1}^{N} \frac{f^2(x_i'\beta)}{P_i(1-P_i)} x_i x_i'$$

donde $f(\cdot)$ es la derivada de $F(\cdot)$.

Evaluando estos estadísticos bajo la hipótesis nula, el test de los multiplicadores de Lagrange se define a partir del estadístico:

$$LM = \frac{d}{d\beta'} LL(\theta)\, J(\tilde{\beta})^{-1} \frac{d}{d\beta} LL(\theta) = \tilde{u}'\tilde{x}(\tilde{x}'\tilde{x})^{-1}\tilde{x}'\tilde{u} \qquad (3\text{-}57)$$

donde

$$\tilde{\beta} = (0, \tilde{\beta}_2)' \quad , \quad \tilde{u} = (\tilde{u}_1, ..., \tilde{u}_N)' \quad , \quad \tilde{x} = (\tilde{x}_1, ..., \tilde{x}_N)'$$

y $\tilde{u}_i = \dfrac{y_i - \tilde{P}_i}{\sqrt{\tilde{P}_i(1-\tilde{P}_i)}}$, $\tilde{x}_i = \dfrac{x_i f(x_{i2}'\tilde{\beta})}{\cdot \sqrt{\tilde{P}_i(1-\tilde{P}_i)}}$

Puesto que se verifica la relación:

$$plim \frac{\tilde{u}'\tilde{u}}{N} = 1$$

el estadístico LM anterior es asintóticamente equivalente a NR_0^2, donde R_0^2 es el coeficiente R^2 de la regresión de $\tilde{u}$ sobre $\tilde{x}$.

En el caso especial del modelo logit binomial se tiene que:

$$P_i = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \quad , \quad f(x_i'\tilde{\beta}) = \tilde{P}_i(1-\tilde{P}_i)$$

por lo que las expresiones anteriores se simplifican bastante, obteniendo una expresión muy simple del estadístico (3-57) correspondiente:

$$LM = \left[\sum_{i=1}^{N}(y_i - \tilde{P}_i)x_i'\right]\left[\sum_{i=1}^{N}\tilde{P}_i(1-\tilde{P}_i)x_i x_i'\right]^{-1}\left[\sum_{i=1}^{N}(y_i - \tilde{P}_i)x_i'\right]$$

Para el modelo probit binomial se tiene que $F(\cdot)$ es la función de distribución asociada a una variable aleatoria con distribución de probabilidad $N[0,1]$ y el factor $f(x_i'\beta) = e^{x_{i2}'\tilde{\beta}_2}$, ya que el factor de proporcionalidad se cancela en las expresiones anteriores:

$$LM = \left[\sum_{i=1}^{N}\frac{(y_i - \tilde{P}_i)}{\tilde{P}_i(1-\tilde{P}_i)}e^{x_{i2}'\tilde{\beta}_2}x_i'\right]\left[\sum_{i=1}^{N}\frac{e^{2x_{i2}'\tilde{\beta}_2}x_i x_i'}{\tilde{P}_i(1-\tilde{P}_i)}\right]^{-1}\left[\sum_{i=1}^{N}\frac{(y_i - \tilde{P}_i)}{\tilde{P}_i(1-\tilde{P}_i)}e^{x_{i2}'\tilde{\beta}_2}x_i'\right]$$

En consecuencia, el test de los multiplicadores de Lagrange es muy fácil de calcular basado en las estimaciones del modelo bajo la hipótesis nula.

De forma análoga a como se han planteado y resuelto los contrastes de hipótesis nula H_0:$\beta_1 = 0$ para modelos de respuesta discreta binomiales, también está la posibilidad de considerar modelos de variable dependiente limitada como el modelo tobit o el modelo continuo con separación muestral.

Para ello únicamente hay que utilizar la función de verosimilitud asociada a cada modelo y a partir de ella calcular la expresión del estadístico *LM* correspondiente.

Engle (1984) plantea un modelo continuo con autoselectividad y obtiene la expresión correspondiente al estadístico.

3.5.3. Test de la razón de verosimilitudes para contrastar la simultaneidad en los modelos de variable dependiente continua con separación muestral

En el proceso de estimación del vector de parámetros desconocidos en los modelos de variable dependiente limitada se ha razonado la necesidad de introducir un factor en el modelo lineal para corregir el sesgo debido a la selección muestral, de forma que la estimación se realiza en dos etapas. En primer lugar se estima el modelo de elección discreta asociado al modelo de variable dependiente limitada y en segundo lugar se estima por mínimos cuadrados ordinarios el modelo lineal con el corrector de selectividad.

En el modelo tobit la misma variable dependiente continua es la que selecciona los valores observados y los que no lo son según su signo, ecuación (3-21):

$$y_i = \begin{cases} y_i^* & sii \quad y_i^* = x_i'\beta + \varepsilon_i > 0 \\ 0 & sii \quad en\ otro\ caso \end{cases}$$

Por el contrario en el modelo continuo-discreto las variables dependientes continuas tienen sus observaciones limitadas según el signo de una tercera variable (3-24):

$$y_{1i} = x_{1i}'\beta_1 + \varepsilon_{1i} \qquad sii \quad I_i = 1$$
$$y_{2i} = x_{2i}'\beta_2 + \varepsilon_{2i} \qquad sii \quad I_i = 0$$

$$I_i = \begin{cases} 1 & sii \quad z_i'\gamma - \varepsilon_i > 0 \\ 0 & sii \quad en\ otro\ caso \end{cases}$$

En el proceso de estimación se ha supuesto de nuevo la existencia de una endogeneidad entre las tres ecuaciones de forma que la estimación por mínimos cuadrados ordinarios de cada variable continua, a partir de las observaciones correspondientes, lleva a estimaciones inconsistentes.

En esta sección se propone utilizar un test clásico, como el de la razón de verosimilitudes para contrastar si en el modelo de tres ecuaciones simultáneas se verifica o no la endogeneidad que se asume en el proceso de estimación, es decir, se planteará contrastar si la ecuación asociada al proceso de selección muestral actúa endógenamente en las otras dos ecuaciones.

Si esto es cierto, es decir, si hay simultaneidad, la estimación por mínimos cuadrados ordinarios con muestras separadas de las dos variables continuas llevará a las estimaciones inconsistentes. Por el contrario, si no existe simultaneidad, estas estimaciones serán correctas y no se deberían utilizar las estimaciones obtenidas con el método en dos etapas que se ha propuesto.

El que exista simultaneidad o no, es debido a que entre las variables aleatorias de los dos modelos lineales continuos, ε_{1i} , ε_{2i}, y la variable aleatoria del modelo de elección discreta, ε_i, que origina la separación muestral hay relación, es decir que no son variables incorrelacionadas. Así, una forma de contrastar la existencia de simultaneidad es contrastar que ambos coeficientes de correlación, $\rho_{1\varepsilon}$, $\rho_{0\varepsilon}$, son nulos (o equivalentemente, que las covarianzas son nulas).

Tras estas consideraciones, el contraste se puede plantear en los siguientes términos:

$$\begin{cases} H_0\colon\ \sigma_{1\varepsilon} = \sigma_{2\varepsilon} = 0 \\ H_1\colon\ \quad no\ H_0 \end{cases}$$

La forma de resolver este contraste paramétrico es utilizando cualquiera de los tests clásicos. En particular puede considerarse el test (3-55) de la razón de verosimilitudes, $LR = -2\left[\mathrm{LL}\left(\hat{\theta}_0\right) - \mathrm{LL}\left(\hat{\theta}\right)\right]$, siendo en este caso $\hat{\theta}$ la estimación máximo-verosímil conjunta de todos los parámetros del modelo con tres ecuaciones simultáneas y $\hat{\theta}_0$ la estimación máximo-verosímil, bajo la hipótesis nula. El estadístico LR seguirá una distribución asintótica χ^2 con dos grados de libertad.

Para obtener el valor del estadístico LR se calcula la estimación máximo-verosímil conjunta de las tres ecuaciones, $\hat{\theta}$, y para calcular $\hat{\theta}_0$ se considera la estimación máximo-verosímil del parámetro γ del modelo de variable discreta y la de los parámetros de los dos modelos lineales por separado, que coincidirán con las estimaciones por mínimos cuadrados ordinarios con separación muestral. Así, bajo la hipótesis nula de no existencia de simultaneidad, la verosimilitud vendrá dada como el producto de las verosimilitudes de los tres modelos estimados:

$$\mathrm{L}\left(\hat{\theta}_0\right) = \mathrm{L}(\hat{\gamma})\,\mathrm{L}\left(\hat{\beta}_1\right)\mathrm{L}\left(\hat{\beta}_2\right)$$

Se rechaza la hipótesis nula de existencia de simultaneidad si el estadístico LR es mayor que el correspondiente cuantil de una χ^2_2.

Es fácil observar que si en el proceso de estimación en dos etapas se considera que la covarianza entre cada variable continua y entre la variable aleatoria del modelo de respuesta cualitativa es nula, el coeficiente asociado al ratio de Mills que corregía el sesgo de selección es nulo.

De esta forma también podría plantearse contrastar la existencia de simultaneidad mediante un contraste de significación sobre los coeficientes asociados al corrector de selectividad en ambas ecuaciones.

Sin embargo esta solución, según Lee y Trost (1978), no es completamente correcta, ya que únicamente permite razonar la endogeneidad de la variable de la separación muestral en cada una de las ecuaciones continuas, pero no en las dos conjuntamente, aunque sí que puede servir como un indicador de la posible existencia de simultaneidad.

3.5.4. Tests equivalentes a los tests clásicos

En este epígrafe se van a presentar otros procedimientos de contraste alternativos al test de Wald, la razón de verosimilitudes y los multiplicadores de Lagrange para resolver los contrastes $\begin{cases} H_0: \theta_1 = \theta_1^{\,0} \\ H_1: \theta_1 \neq \theta_1^{\,0} \end{cases}$

siendo θ_1 un subvector del vector paramétrico, $\theta' = (\theta_1', \theta_2')$. La hipótesis nula hace referencia al subvector θ_1 sin contrastar ninguna restricción sobre el subvector θ_2.

El primer procedimiento de contraste que se va a presentar es una generalización directa del test de los multiplicadores de Lagrange. La idea es buscar, mediante un desarrollo en serie, una expresión alternativa para el "score" y además permitirá encontrar una estimación consistente aunque ineficiente de los parámetros que no están restringidos en la hipótesis nula, es decir, de θ_2.

Sea $\tilde{\tilde{\theta}}_2$ una estimación de θ_2 consistente y $\tilde{\tilde{\theta}} = \left(\theta_1^{\,0}, \tilde{\tilde{\theta}}'\right)'$. Desarrollando la derivada del logaritmo de la función de verosimilitud, o "score", evaluado en θ alrededor de la estimación máximo-verosímil bajo la hipótesis nula $\tilde{\theta} = \left(\theta_1^{\,0}, \tilde{\theta}_2'\right)'$ se obtiene, particionando el vector de derivadas según el parámetro:

$$\frac{d}{d\theta} LL\left(\tilde{\tilde{\theta}}\right) = \begin{bmatrix} \dfrac{d}{d\theta_1} LL\left(\tilde{\theta}\right) \\ 0 \end{bmatrix} + \begin{bmatrix} \dfrac{d^2}{d\theta_1 d\theta_2'} LL\left(\bar{\theta}\right)\left(\tilde{\tilde{\theta}}_2 - \tilde{\theta}_2\right) \\ \dfrac{d^2}{d\theta_2 d\theta_2'} LL\left(\bar{\theta}\right)\left(\tilde{\tilde{\theta}}_2 - \tilde{\theta}_2\right) \end{bmatrix}$$

siendo

$$\bar{\theta} \in \left(\tilde{\theta}, \tilde{\tilde{\theta}}\right) \qquad \text{y} \qquad \frac{d}{d\theta_2} \mathrm{LL}\left(\tilde{\theta}\right) = 0.$$

Desde la ecuación vectorial anterior se puede escribir que:

$$\frac{d}{d\theta_1} \mathrm{LL}\left(\tilde{\theta}\right) = \frac{d}{d\theta_1} \mathrm{LL}\left(\tilde{\tilde{\theta}}\right) - \frac{d^2}{d\theta_1 d\theta_2'} \mathrm{LL}\left(\bar{\theta}\right) \left[\frac{d^2}{d\theta_2 d\theta_2'} \mathrm{LL}\left(\bar{\theta}\right) \right]^{-1} \frac{d}{d\theta_2} \mathrm{LL}\left(\tilde{\tilde{\theta}}\right)$$

o equivalentemente:

$$\frac{d}{d\theta_1} \mathrm{LL}\left(\tilde{\theta}\right) = s_1\left(\tilde{\tilde{\theta}}\right) - J_{12}\left(\tilde{\tilde{\theta}}\right) J_{22}^{-1}\left(\tilde{\tilde{\theta}}\right) s_2\left(\tilde{\tilde{\theta}}\right) \qquad (3\text{-}58)$$

Ésta es una expresión alternativa del "score" utilizado en el test de los multiplicadores de Lagrange.

Se define el test $C(\alpha)$ de Neyman como el test de los multiplicadores de Lagrange que aparece al utilizar el "score" con la expresión (3-58) anterior.

Es fácil ver que este ajuste es un paso de la iteración de Newton-Raphson, dada por la ecuación (3-27), para encontrar una estimación eficiente de θ_2 basado en una estimación inicial consistente. Este resultado puede significar una gran simplificación computacional en algunas situaciones.

Nótese que este procedimiento de contraste utiliza estimaciones diferentes de los parámetros, aunque se basa en el principio de los multiplicadores de Lagrange.

Un segundo procedimiento de contraste alternativo es el propuesto por Durbin (1970) y también está basado en estimaciones diferentes de los parámetros, pero a diferencia del anterior trabaja con el subvector paramétrico θ_1.

Durbin sugiere calcular la estimación máximo-verosímil de θ_1 suponiendo para θ_2 el valor de su estimación máximo-verosímil bajo la hipótesis nula. Sean $\tilde{\tilde{\theta}}_1$ y $\tilde{\theta}_2$ ambas estimaciones.

Al igual que el procedimiento $C(\alpha)$ de Neyman, se hace un desarrollo en serie del "score" con respecto al parámetro θ_1 sobre la estimación máximo-verosímil $\tilde{\tilde{\theta}}_1$ manteniendo para θ_2 el valor $\tilde{\theta}_2$. El resultado obtenido es:

$$\frac{d}{d\theta_1}\mathrm{LL}(\tilde{\theta}) = -\frac{d^2}{d\theta_1\, d\theta_1'}\mathrm{LL}(\bar{\theta})\left(\tilde{\tilde{\theta}}_1 - \theta_1^{\,0}\right) \qquad (3\text{-}59)$$

ya que el primer término del desarrollo es cero por la propia definición del estimador $\tilde{\tilde{\theta}}_1$ como el máximo-verosímil.

Puesto que el Hessiano se supone no singular es fácil ver que cualquier test basado en la diferencia $\left(\tilde{\tilde{\theta}}_1 - \theta_1^{\,0}\right)$ tendrá la misma región crítica que un test basado en el "score", directamente. Así, el test de Durbin definido como el test de los multiplicadores de Lagrange utilizando la anterior expresión para el "score", es equivalente al test $C(\alpha)$ de Neyman.

Para ver el tercer procedimiento de contraste que se va a proponer se utiliza un razonamiento diferente.

Este nuevo procedimiento se basa en el principio de Hausman y su diferencia con respecto a los anteriores está en que ahora el parámetro de interés no es θ_1, sino que la idea de Hausman se centra en el subvector paramétrico θ_2 que no está restringido en la hipótesis nula.

El objetivo es restringir el espacio paramétrico teniendo en cuenta que el vector considerado en la hipótesis nula, θ_1, es igual a algún valor preasignado sin que esto cambie la consistencia de las estimaciones de θ_2, una posibilidad es suponer $\theta_1 = \theta_1^{\,0}$.

El test de Hausman se define a partir de la diferencia entre dos estimadores del subvector paramétrico θ_2 , uno que es eficiente bajo la hipótesis nula y que puede ser $\tilde{\theta}_2$ y otro estimador consistente, pero probablemente ineficiente bajo la hipótesis alternativa y que se denotará por $\hat{\hat{\theta}}_2$.

La idea original de Hausman implica hacer algunas hipótesis acerca de las propiedades de este segundo estimador $\hat{\hat{\theta}}_2$. No obstante, en un trabajo posterior,

Hausman y Taylor (1981) modifican estas afirmaciones y proponen utilizar la estimación máximo-verosímil de θ_2 bajo la hipótesis alternativa, $\hat{\theta}_2$.

De nuevo se considera un desarrollo en serie del "score" alrededor de la estimación máximo-verosímil $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2\right)'$ y evaluándolo en el valor de $\tilde{\theta} = \left(\theta_1^0, \tilde{\theta}_2\right)'$ se obtiene que:

$$\frac{d}{d\theta} LL(\tilde{\theta}) = \frac{d^2}{d\theta\, d\theta'} LL(\bar{\theta})(\tilde{\theta} - \hat{\theta}) \qquad (3\text{-}60)$$

siendo $\bar{\theta} \in (\hat{\theta}, \tilde{\theta})$.

Esta expresión puede escribirse equivalentemente como:

$$\begin{bmatrix} \theta_1^0 - \hat{\theta}_1 \\ \tilde{\theta}_2 - \hat{\theta}_2 \end{bmatrix} = \left[\frac{d^2}{d\theta\, d\theta'} LL(\bar{\theta})\right]^{-1} \begin{bmatrix} \frac{d}{d\theta_1} LL(\tilde{\theta}) \\ 0 \end{bmatrix}$$

Se puede observar ahora que los tres tests anteriores se basan en la matriz J que es no singular, y que las regiones críticas de los tres son la misma. En el caso particular del test de Hausman, la diferencia se basa en J^{21} veces el score. Si esta matriz es no singular, los tests serán equivalentes asintóticamente.

3.5.5. Test de exactitud y utilidad

Los contrastes de hipótesis planteados hasta el momento permiten discernir entre un modelo con restricciones en los parámetros y el modelo estimado. Todos los contrastes de este tipo se resuelven utilizando los tests clásicos como la razón de verosimilitudes, el de Wald o el de los Multiplicadores de Lagrange.

No obstante, no hay que dejar de lado que con los modelos de respuesta cualitativa se están dando predicciones sobre las probabilidades de elección $P_{ij} = P(y_{ij} = 1 / x_i)$, mientras que lo que se pretende analizar es la respuesta del individuo, es decir, la secuencia de valores 1 y 0 que toma la variable respuesta.

Así, en un proceso de elegir el modelo de respuesta cualitativa que mejor explique el problema no bastará con realizar un contraste de significación del modelo propuesto frente a un modelo con algunos, o todos, los parámetros nulos, sino que deberá completarse con tests capaces de comparar las predicciones hechas con el modelo.

Hauser (1978) propone completar el test de significación con un test de utilidad y uno de exactitud para conseguir seleccionar el mejor modelo, derivando ambos desde la Teoría de la Información.

Cuando se habla de utilidad de un modelo se está haciendo referencia al porcentaje de incertidumbre explicada por la información que proporciona el modelo, mientras que la exactitud se centra en determinar si las observaciones 0 y 1 son razonables, bajo la hipótesis de que el modelo es correcto.

A continuación se desarrollan los tests que propone Hauser.

Sea $C = \{1,2,.....,J\}$ el conjunto de alternativas, x_i las variables explicativas observadas sobre los individuos, $i = 1,2,.....,N$.

La probabilidad a priori de elegir la alternativa j, sin observar las variables explicativas x_i es $\mathrm{P}(y_{ij} = 1)$.

La probabilidad a posteriori de elegir la alternativa j, observadas las variables x_i, es la que proporciona el modelo $\mathrm{P}(y_{ij} = 1 / x_i)$.

La entropía a priori que mide la total incertidumbre antes de observar x_i será:

$$H(C) = -\sum_j \mathrm{P}(y_{ij} = 1) \log \mathrm{P}(y_{ij} = 1)$$

La entropía a posteriori que mide la incertidumbre que queda después de observar x_i será:

$$H(C / x) = -\sum_{x_i} \sum_j \mathrm{P}(y_{ij} = 1, x_i) \log \mathrm{P}(y_{ij} = 1 / x_i)$$

siendo $\mathrm{P}(y_{ij} = 1, x_i)$ la probabilidad conjunta de elegir la alternativa j y observar el vector x_i.

La información proporcionada por las variables x_i para la alternativa j viene dada por:

$$I(j, x_i) = \log \frac{P(y_{ij} = 1 / x_i)}{P(y_{ij} = 1)}$$

Asumiendo que el espacio de características es discreto y que está formado únicamente por los valores observados, la distribución de probabilidad $P(x_i)$ puede tomarse como la proporción muestral, $\frac{1}{N}$, y considerando que la probabilidad conjunta de la alternativa j y la observación x_i puede descomponerse como $P(y_{ij} = 1, x_i) = P(y_{ij} = 1 / x_i) P(x_i)$ la información esperada del modelo adopta la expresión:

$$E\, I(C,X) = \sum_{x_i} \sum_{j} P(y_{ij} = 1, x_i)\, I(j, x_i) =$$

$$\sum_{x_i} \sum_{j} P(x_i) P(y_{ij} = 1 / x_i) \log \frac{P(y_{ij} = 1 / x_i)}{P(y_{ij} = 1)} =$$

$$\sum_{i} \sum_{j} \frac{1}{N} P(y_{ij} = 1 / x_i) \log \frac{P(y_{ij} = 1 / x_i)}{P(y_{ij} = 1)} \qquad (3\text{-}61)$$

Por otro lado, la información empírica para el conjunto de alternativas y observaciones es:

$$I(C, X) = \sum_{i} \sum_{j} \frac{1}{N} y_{ij} \log \frac{P(y_{ij} = 1 / x_i)}{P(y_{ij} = 1)} \qquad (3\text{-}62)$$

Antes de desarrollar los tests de utilidad y exactitud, si el investigador se plantea realizar un contraste de significación del modelo puede utilizar el test de la razón de verosimilitudes, como se ha comentado en los epígrafes anteriores. A continuación se desarrollará la expresión del estadístico LR en términos de la información empírica antes definida.

El estadístico LR se obtiene desde (3-55) como: $LR = -2\left[LL(\hat{\theta}_0) - LL(\hat{\theta})\right]$ donde $\hat{\theta}$ y $\hat{\theta}_0$ son las estimaciones máximo-verosimiles (pueden utilizarse otros

estimadores con la misma distribución asintótica) de los parámetros del modelo considerado y del modelo bajo la hipótesis nula respectivamente.

Las expresiones $\mathrm{LL}(\hat{\theta})$ y $\mathrm{LL}(\hat{\theta}_0)$ del logaritmo de la función de verosimilitud evaluada en las estimaciones $\hat{\theta}$ y $\hat{\theta}_0$ vienen dadas por:

$$\mathrm{LL}(\hat{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij} \log \mathrm{P}(y_{ij} = 1 / x_i)$$

y

$$\mathrm{LL}(\hat{\theta}_0) = \sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij} \log \mathrm{P}(y_{ij} = 1)$$

Sustituyendo las expresiones del logaritmo de la función de verosimilitud, el estadístico LR adopta la forma

$$LR = 2\sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij} \log \frac{\mathrm{P}(y_{ij} = 1 / x_i)}{\mathrm{P}(y_{ij} = 1)} = 2\, N\, I(C, X) \qquad (3\text{-}63)$$

Después de realizar un contraste de significación, Hauser plantea calcular la utilidad y exactitud del modelo utilizando las medidas de información antes definidas.

Según la definición de "utilidad" del modelo es fácil ver que la forma de obtenerla será calculando la reducción de incertidumbre debida al modelo.

Se puede definir el estadístico $U^2 = \dfrac{I(C, X)}{H(C)}$ que es el porcentaje de incertidumbre explicada por el modelo con respecto a la incertidumbre total. Este estadístico U^2 será una medida de utilidad del modelo.

Por otra parte puede verse fácilmente que la información esperada es exactamente igual a la diferencia entre la incertidumbre total y la incertidumbre que queda después de observar x_i, es decir, $\mathrm{E}\, I(C, X) = H(C) - H(C / X)$, de donde se puede definir otra medida de la utilidad del modelo como:

$$\mathrm{E}U^2 = \frac{\mathrm{E}I(C,X)}{H(C)} \tag{3-64}$$

Evidentemente $\mathrm{E}U^2$ es el valor esperado del estadístico U^2 definido antes y puede utilizarse para contrastar la exactitud del modelo, viendo si U^2 dista mucho o no de su valor esperado $\mathrm{E}U^2$.

Para ver la exactitud de un modelo habrá que comparar la información empírica de dicho modelo con su información esperada, ya que la definición de exactitud implica comparar las observaciones con el modelo propuesto.

Intuitivamente la forma de aceptar el modelo como correcto es calculando la información empírica $I(C,X)$ y comprobar que está cerca de su valor esperado $\mathrm{E}I(C,X)$.

Para muestras grandes y bajo las usuales condiciones de regularidad, Hauser demuestra que el estadístico $I(C,X)$ sigue una distribución Normal con media dada por $\mathrm{E}I(C,X)$ y varianza dada por la siguiente expresión:

$$\mathrm{V}(C,X) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{J}\mathrm{P}(y_{ij}=1/x_i)\left[\log\frac{\mathrm{P}(y_{ij}=1/x_i)}{\mathrm{P}(y_{ij}=1)}\right]^2 - \left[\sum_{j=1}^{J}\mathrm{P}(y_{ij}=1/x_i)\log\frac{\mathrm{P}(y_{ij}=1/x_i)}{\mathrm{P}(y_{ij}=1)}\right]^2\right] \tag{3-65}$$

Un test de dos colas se puede aplicar para determinar si la información empírica es una observación razonable del modelo. Si la información empírica, $I(C,X)$, está estadísticamente lejos de la información esperada del modelo, $\mathrm{E}I(C,X)$, se rechaza el modelo como incapaz de explicar las observaciones empíricas.

Nótese que puesto que el estadístico de la razón de verosimilitudes LR verifica la relación $LR = 2\,N\,I(C,X)$ se tiene que este estadístico $2\,N\,I(C,X)$ sigue una distribución χ^2, mientras que Hauser demuestra que $I(C,X)$ sigue una distribución Normal. Esto no es ninguna contradicción, ya que el estadístico $I(C,X)$ utilizado en el test de exactitud sigue una distribución

Normal porque sólo las variables respuesta y_{ij} son variables aleatorias bajo la hipótesis nula de que el modelo es correcto, mientras que $2\ N\ I(C, X)$ en el test de significación sigue una distribución χ^2 porque tanto las variables respuesta y_{ij} como las probabilidades de respuesta $P(y_{ij} = 1 / x_i)$ son variables aleatorias bajo la hipótesis de que el modelo de la hipótesis nula es correcto.

3.5.6. Test sobre la potencia predictiva de modelos binomiales

En el apartado 3.5.5 se ha desarrollado un test para contrastar la adecuación del modelo, es decir, la adecuación de las observaciones al modelo propuesto. También se ha proporcionado una medida de la utilidad del modelo, entendida como el porcentaje de incertidumbre explicada por el modelo con respecto a la incertidumbre total. En este epígrafe se propone un nuevo test para completar los ya conocidos. Se pretende contrastar la adecuación de las predicciones siguiendo la idea de los tests de predicción de series temporales.

Aunque los tests de predicción son más adecuados en series temporales, dada la naturaleza de los datos, que en modelos de variable dependiente limitada, también pueden utilizarse en estos últimos, ya que los cambios estructurales pueden deberse no solo al tiempo, sino que pueden estar relacionados con el espacio de las variables del modelo.

Anderson (1987) desarrolló un test para la potencia predictiva de los modelos de respuesta cualitativa utilizando variables ficticias para el período de predicción, igual que en series temporales.

En este apartado se va a presentar el test de Anderson que es muy sencillo de calcular y asintóticamente válido.

La idea del autor es considerar dos funciones de verosimilitud, una de ellas construida con la muestra completa de observaciones $i = 1,2,.....,N$ y otra omitiendo algunas observaciones $i = 1,2,.....,N_1$, con $N_1 < N$, sin pérdida de generalidad se puede asumir que las últimas observaciones son las omitidas. Y a partir de esas funciones de verosimilitud utiliza un test clásico de la razón de verosimilitudes.

Se define un modelo de variable latente como:

$$y_i = \begin{cases} 1 & y_i^* = x_i'\beta + \varepsilon_i \geq 0 \\ 0 & y_i^* = x_i'\beta + \varepsilon_i < 0 \end{cases}$$

y otro modelo de variable latente que utiliza en la definición de la variable no observada, variables ficticias, d_i, para las observaciones omitidas $i = N_1 + 1, \ldots\ldots, N$ como:

$$y_i = \begin{cases} 1 & y_i^* = x_i'\beta + zd_i + \varepsilon_i \geq 0 \\ 0 & y_i^* = x_i'\beta + zd_i + \varepsilon_i < 0 \end{cases}$$

Sea $\mathrm{LL}_N(\beta)$ el logaritmo de la función de verosimilitud para el primer modelo calculada con toda la muestra, y $\hat{\beta}$ la estimación máximo-verosimil del vector de parámetros de ese modelo. Sea $\mathrm{LL}_{N_1}(\beta)$ el logaritmo de la función de verosimilitud calculada para la muestra reducida y utilizando variables ficticias para las observaciones omitidas y $\hat{\hat{\beta}}$ la estimación máximo-verosímil del vector de parámetros del segundo modelo.

La estimación $\hat{\hat{\beta}}$ puede calcularse sobre el primer modelo considerando únicamente las N_1 primeras observaciones, puesto que la función de verosimilitud del modelo con variables ficticias es igual a la que aparece al considerar en el primer modelo las observaciones $i = 1,2,\ldots\ldots, N_1$.

El test de la razón de verosimilitudes se define como (3-55):

$$LR = -2\left[\mathrm{LL}_N\left(\hat{\beta}\right) - \mathrm{LL}_{N_1}\left(\hat{\hat{\beta}}\right)\right]$$

que asintóticamente sigue una distribución χ^2 con $N - N_1$ grados de libertad.

Con este estadístico se puede contrastar la potencia predictiva del modelo sobre las observaciones $i = N_1 + 1, \ldots\ldots, N$, ya que para el contraste de hipótesis nula H_0: *predicción correcta* frente a la hipótesis alternativa definida por H_1: *predicción incorrecta*, rechaza la hipótesis nula cuando $LR > \chi^2_{N-N_1,\alpha}$, siendo α el nivel de significación exigido.

Si el modelo predice correctamente se tendrá que $\mathrm{LL}_N \approx \mathrm{LL}_{N_1}$ y por lo tanto $LR \approx 0$. Si el modelo predice incorrectamente la relación entre las funciones de

verosimilitud implicará que $LL_N < LL_{N_1}$ y por lo tanto LR será un valor que tenderá a ser grande a medida que aumente la diferencia por lo que llevará a rechazar la hipótesis nula.

3.5.7. Medidas de Bondad de Ajuste

En la literatura de los modelos de respuesta cualitativa se pueden encontrar algunas medidas de bondad de ajuste para este tipo de modelos. Algunas de ellas están basadas en la función de verosimilitud y primariamente tienen un propósito de contraste. Otras se basan en las diferencias entre el valor observado y el predicho.

A partir de la función de verosimilitud McFadden (1974) define un coeficiente semejante al R^2 de un modelo de regresión lineal

$$R_M^2 = 1 - \frac{LL(\tilde{\beta})}{LL_0} \tag{3-66}$$

donde $LL(\tilde{\beta})$ es el logaritmo de la función de verosimilitud evaluada en una estimación consistente del vector de parámetros β y LL_0 es el valor del logaritmo de la función de verosimilitud de un modelo nulo, es decir, que incluye únicamente el término constante.

Por la propia definición, este coeficiente está acotado entre 0 y 1, pero no puede interpretarse como el R^2 de una regresión porque en su definición intervienen todas las características de la distribución y no representa el porcentaje de varianza explicada.

Por otro lado, para modelos binomiales Efron (1978) define una medida basada en las predicciones:

$$R_E^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - F(x_i'\tilde{\beta})\right)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3-67}$$

pero presenta el inconveniente de no estar acotado en el intervalo unidad siendo difícil su interpretación.

Como solución alternativa, Laitilia (1993) definió un coeficiente pseudo-R^2 a partir del modelo de variable latente $y_i^* = x_i'\beta - \varepsilon_i$ como:

$$R_L^2 = \frac{\tilde{\beta}'\,\tilde{\Sigma}_X\,\tilde{\beta}}{\left(\tilde{\sigma}^2 + \tilde{\beta}'\,\tilde{\Sigma}_X\,\tilde{\beta}\right)} \tag{3-68}$$

siendo $\tilde{\beta}$ una estimación consistente de β, $\tilde{\sigma}^2$ una estimación consistente de la varianza σ^2 de la variable aleatoria ε_i y $\tilde{\Sigma}_X = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})'\,(x_i - \bar{x})$ la matriz de varianzas-covarianzas muestral.

Bajo condiciones de regularidad $\underset{N\to\infty}{plim}\, R_L^2 = R^2$ siendo R^2 el coeficiente de determinación del modelo de regresión lineal $y_i^* = x_i'\beta - \varepsilon_i$ estimado por mínimos cuadrados ordinarios.

Otra posibilidad para medir la bondad de ajuste de modelos binomiales es utilizar el porcentaje de incertidumbre explicada por el modelo:

$$\frac{I(y, X)}{H(y)} \tag{3-69}$$

donde:

$H(y) = -\left[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})\right]$ es la entropía a priori o incertidumbre total del modelo e

$$I(y, X) = \frac{1}{N}\sum_{i=1}^{N}\left\{y_i \log\frac{\hat{P}_i}{\bar{y}} + (1-\bar{y})\log\frac{1-\hat{P}_i}{1-\bar{y}}\right\}$$

es la información empírica, siendo

$$\bar{y} = \sum_{i=1}^{N}\frac{y_i}{N} \quad , \quad \hat{P}_i = F(x_i'\beta).$$

Es fácil ver que $R_M^2 = \frac{I(y, X)}{H(y)}$.

Otra posibilidad es utilizar coeficientes que consideran los residuos obtenidos con el modelo, como el coeficiente propuesto por Amemiya (1988)

$$W = \sum_{i=1}^{N} \frac{\left(y_i - \hat{\mathrm{P}}_i\right)^2}{\hat{\mathrm{P}}_i\left(1 - \hat{\mathrm{P}}_i\right)} \tag{3-70}$$

que es la suma ponderada de cuadrados de los residuos.

En este caso se elegirá aquel modelo que presente el menor valor de este coeficiente.

Aunque estas medidas están planteadas para los modelos de elección discreta, algunas de ellas servirán para los modelos de variable dependiente limitada. Por ejemplo el coeficiente R_M^2 ha sido utilizado en modelos de este tipo (Laitila, 1993).

APÉNDICE A: Propiedades de los estimadores

Los epígrafes 3.3 y 3.4 se han dedicado a la obtención de estimadores para los parámetros desconocidos en los modelos de respuesta cualitativa y en los modelos de variable dependiente limitada respectivamente.

En el caso de los modelos de respuesta cualitativa se han encontrado estimadores del vector de parámetros desconocido θ según la información disponible acerca de las distribuciones marginales $P(x)$ y $P(j)$ y según el tipo de muestreo realizado.

Algunos de los estimadores propuestos se reducen al estimador máximo-verosímil, y como tal gozan de buenas propiedades.

Otros estimadores se han obtenido de forma alternativa ya que las condiciones del problema no permitían obtener el estimador máximo-verosímil directamente.

Para los modelos de variable dependiente limitada se ha propuesto un método en dos etapas que combina la estimación máximo-verosímil para el modelo binomial y la estimación por mínimos cuadrados ordinarios para las variables continuas.

En cualquier caso, la elección de un estimador u otro viene condicionada a las particularidades del problema concreto y en caso de existir varias soluciones alternativas, una posibilidad es elegir el que tenga mejores propiedades o el que resulte más fácil computacionalmente.

Se van a analizar las propiedades estadísticas que poseen los estimadores propuestos en este trabajo.

En primer lugar se presentan las definiciones de las propiedades que se van a considerar.

Consistencia

Se dice que un estimador $\hat{\theta}_N$ es consistente para el parámetro θ si $\hat{\theta}_N$ converge casi seguramente al verdadero valor de θ.

Eficiencia

Un estimador $\hat{\theta}_N$ es asintóticamente eficiente cuando su matriz de varianzas-covarianzas asintótica coincide con la inversa de la matriz de información de Fisher.

Normalidad Asintótica

Un estimador $\hat{\theta}_N$ se dice que es asintóticamente Normal cuando se verifica que la variable aleatoria $\sqrt{N}\left(\hat{\theta}_N - \theta^*\right)$ converge a una variable con distribución de probabilidad Normal, siendo θ^* el verdadero valor del parámetro desconocido.

A.1. Propiedades de los estimadores en los modelos de respuesta cualitativa

La demostración de la consistencia y Normalidad Asintótica del estimador máximo-verosímil debida a Rao (1973) no puede ser utilizada directamente en este caso por las razones que se comentan a continuación. Si se consideran como observaciones muestrales los valores que toma la variable respuesta y_i sobre los individuos, no se cumple la condición exigida por Rao de tener variables idénticamente distribuidas, ya que la distribución de la variable respuesta y_i depende de las características x_i que varían con cada individuo.

Por el contrario, al tomar como observaciones los pares (j_i, x_i) que es la propuesta de muestreo cuando las variables x_i se consideran aleatorias, se soluciona el problema anterior, pero de nuevo es necesario modificar los lemas de Rao, ya que éste únicamente tiene en cuenta un modelo sin variables explicativas. No es este el caso, ya que en los modelos de respuesta cualitativa, las variables x_i son variables explicativas de la variable respuesta y_i.

A continuación se enuncian los lemas que, junto con las hipótesis de regularidad enunciadas en el epígrafe correspondiente a la estimación, se van a utilizar para la demostración de las correspondientes propiedades antes comentadas.

Los tres primeros lemas serán los que permitirán comprobar la consistencia del estimador y los lemas 4 y 5 son los que demuestran la Normalidad Asintótica.

Lemas para la demostración de la consistencia de los estimadores

Lema 1 (Amemiya, 1973)

Sea $f_N(s,\delta)$, $N=1,2,.....,\infty$, una sucesión de funciones medibles en un espacio medible S y para cada $s \in S$ una función continua en $\delta \in D$, siendo el espacio D compacto.

Entonces existe una sucesión de funciones medibles $\delta_N(s)$, $N=1,2,.....,\infty$ tal que

$$f_N(s,\delta_N(s)) = \sup_{\delta \in D} f_N(s,\delta), \quad \forall s \in S, \quad N=1,2,.....,\infty$$

Además si para casi todo $s \in S$, $f_N(s,\delta)$ converge uniformente a $f(s)$ para todo elemento $\delta \in D$ y si $f(s)$ tiene un único máximo en $\delta^* \in D$, entonces la sucesión $\delta_N(s)$ converge a δ^* para casi todo $s \in S$.

Lema 2 (Jennrich, 1969)

Sea μ una medida de probabilidad sobre un espacio Euclídeo T, sea D un subconjunto compacto de un espacio Euclídeo, y sea $g_1(t,\delta)$ una función de δ continua para $t \in T$ y una función de t medible para $\delta \in D$.

Se asume que $|g_1(t,\delta)| \leq \alpha(t)$ para todo t, δ y alguna función α que verifique ser μ-integrable. Para cualquier sucesión $s = t_1, t_2, ...$, sea

$$f_N(s,\delta) = \sum_{i=1}^{N} g_1(t_i,\delta)\frac{1}{N}$$

y sea S el conjunto de todas las sucesiones. Si las sucesiones s son sacadas como muestras aleatorias de S, entonces para casi toda realización de tal secuencia, cuando $N \to \infty$ se verifica que $f_N(s,\delta)$ converge uniformente para todo δ a

$$E[g_1(t,\delta)] = f(\delta)$$

Lema 3 (Rao, 1973)

Sea $g_2(t,\delta)$ una función con valores reales sobre un espacio $T \times D$ tal que g es integrable con respecto a una medida μ sobre T y $g_2(t,\delta) \geq 0$, para todo $t \in T$ y $\delta \in D$. Sea δ^* un elemento de D tal que $g_2(t,\delta^*) > 0$ para casi todo $t \in T$ y

$$\int_T \left(g_2(t,\delta^*) - g_2(t,\delta)\right) d\mu \geq 0$$

para todo $\delta \in D$.

Entonces la expresión

$$f(\delta) = \int_T g_2(t,\delta^*) \ln g_2(t,\delta)\, d\mu$$

alcanza su máximo en $\delta = \delta^*$.

El máximo es único si para $\delta \in D$ tal que $\delta \neq \delta^*$, existe un $T_\delta \subset T$ tal que

$$\int_{T_\delta} g_2(t,\delta)\, d\mu \neq \int_{T_\delta} g_2(t,\delta^*)\, d\mu$$

Con los tres primeros lemas se puede demostrar la consistencia de los estimadores propuestos en el trabajo.

Nótese que en el lema 2 aparece la definición de las funciones $f_N(s,\delta)$ del lema 1 como la media muestral de unas funciones $g_1(t,\delta)$, que se pueden interpretar como

$$g_1(t,\delta) = \ln P\left((j,x),\theta\right), \quad t = (j,x), \quad \delta = \theta$$

La condición $|g_1(t,\delta)| \leq \alpha(t)$ se verifica porque $P\left((j,x),\theta\right)$ es estrictamente mayor que cero y por tanto existe una cota inferior estrictamente positiva P_0:

$$P_0 \leq P\left((j,x),\theta\right)$$

y al tomar logaritmos:

$$\ln P_0 \le \ln P((j,x),\theta) \Rightarrow |\ln P_0| \ge |\ln P((j,x),\theta)| \Rightarrow$$

$$|g_1(t,\delta)| \le \alpha(t) = |\ln P_0|$$

Este lema 2 únicamente demuestra la convergencia uniforme de la media muestral a la media de la población.

La primera parte del lema 3 demuestra la existencia del máximo de la función dada por $f(\delta) = E[\ln P((j,x),\theta)]$, ya que considera una función positiva para cualquier valor del vector de parámetros, $g_2(t,\delta) = P((j,x),\theta)$, y además se verifica que $\exists\theta^* \in \Theta$ tal que $P((j,x),\theta^*) > 0$ para casi todo (j,x) y

$$\int_{C\times Z} \left(P((j,x),\theta^*) - P((j,x),\theta)\right) d\mu \ge 0$$

para todo $\theta \in \Theta$.

Así

$$f(\delta) = f(\theta) = \int_{C\times Z} P((j,x),\theta^*) \ln P((j,x),\theta)\, d\mu = E[\ln P((j,x),\theta^*)]$$

tiene un máximo en $\theta = \theta^*$.

Para ver la unicidad basta considerar la hipótesis 2 de identificabilidad que nos asegura que para $\theta \in \Theta, \theta \ne \theta^*$ existe $A \subset C \times Z$ tal que

$$\int_A P((j,x),\theta^*)\, d\mu \ne \int_A P((j,x),\theta)\, d\mu.$$

La secuencia de funciones que plantea el lema 1, serán los logaritmos de la función de verosimilitud para cada muestra:

$$f_N(s,\delta) = \sum_{i=1}^{N} \frac{1}{N} \ln P((j_i,x_i),\theta)$$

siendo $s = t_1, t_2, \ldots.$; $t_i = (j_i, x_i)$, $\delta = \theta$.

En primer lugar el lema 1 demuestra la existencia de los estimadores máximo-verosímiles, ya que buscar el supremo de $f_N(s,\delta)$ para $\delta \in D$ equivale a maximizar el logaritmo de la función de verosimilitud, luego $\exists\ \theta_N(j,x)$, que es el estimador máximo-verosímil calculado sobre una muestra de tamaño N:

$$\exists\ \theta_N(j,x)\ /\ \sum_{i=1}^{N}\frac{1}{N}\ln P\left((j_i,x_i),\theta_N(j_i,x_i)\right)=\sup_{\theta\in\Theta}\sum_{i=1}^{N}\frac{1}{N}\ln P\left((j_i,x_i),\theta\right)$$

Para la segunda parte del lema se puede seguir el lema 2 para ver que las funciones $f_N(s,\delta)$ convergen uniformemente a $f(\delta)$, ya que

$$f_N(s,\delta)=\sum_{i=1}^{N}\frac{1}{N}\ln P\left((j_i,x_i),\theta\right)$$

no es más que la media muestral de observaciones $\ln P\left((j,x),\theta\right)$, y la media muestral converge a la esperanza de la población, luego $f_N(s,\delta)$ converge a $f(\delta)=E\left[\ln P\left((j,x),\theta\right)\right]$ que es una función de θ, y que se representará como $f(\theta)$.

Con el lema 3 se puede demostrar que la función $f(\theta)$ alcanza un único máximo en $\theta=\theta^*$, de donde $f(\delta)$ tiene un único máximo en $\delta=\delta^*$ y el lema 1 asegura que la sucesión $\theta_N(j,x)$ de estimadores máximo-verosímiles converge uniformemente al verdadero valor del parámetro desconocido $\theta=\theta^*$, por lo tanto el estimador es consistente.

En la demostración de esta propiedad de consistencia, para cada estimador se considerará $\ln P\left((j_i,x_i),\theta\right)$ o la transformación adecuada para conseguir que la función f_N sea el logaritmo de la función de verosimilitud, identificando en cada momento si el parámetro δ coincide con θ, y por lo tanto $D=\Theta$, o $D=\Theta_0$ en el caso de las restricciones, o si el parámetro δ es (θ,P) siendo en este caso $D=\Theta\times\Pi$.

Lemas para la demostración de la Normalidad Asintótica de los estimadores

Lema 4

Dadas las hipótesis del lema 1 y además $f_N(s,\delta) \in C^2(D)$ para casi todo $s \in S$ y $f(\cdot) \in C^2(D)$. Sea

$$(r: D \to R^M) \in C^2(D)$$

con $r(\delta^*) = 0$ y sea la matriz

$$R = \frac{d}{d\delta} r(\delta^*)$$

de rango completo.

Sea $\hat{\delta}_N(s)$, $N = 1,2,.....,\infty$ una secuencia de soluciones al problema de maximizar $f_N(s,\delta)$ sujeto a las restricciones $r(\delta) = 0$. Se supone $\delta^* \in \text{int}(D)$, que $F = \frac{d^2}{d\delta\, d\delta'} f(\delta^*)$ es no singular y que $\sqrt{N}\left(\frac{d}{d\delta} f_N(s,\delta^*)\right)$ converge en ley a una variable aleatoria con distribución $N[0,\Delta]$.

Entonces se verifica que la variable $\sqrt{N}\left(\hat{\delta}_N - \delta^*\right)$ converge en ley a una

$$N[0, \Omega^{-1}\Delta\,\Omega^{-1}]$$

donde $\Omega^{-1} = F^{-1} - F^{-1} R\left(R' F^{-1} R\right)^{-1} R' F^{-1}$.

Lema 5

Dadas las hipótesis de lema 4 y las hipótesis del lema 2 considerando

$$f_N(s,\delta) = \sum_{i=1}^{N} g(t_i, \delta, h_N(\delta)) \frac{1}{N}$$

donde

$$h_N(\delta) = \sum_{i=1}^{N} e(t_i, \delta)\frac{1}{N}, \quad e \in C^2(D, R^L) \quad y \quad g \in C^2(D)$$

Se supone que:

$$h(\delta^*) = \int e(t, \delta^*)\, d\mu = E[e(t, \delta^*)]$$

$$V_g = \int \frac{d}{d\delta} g(t, \delta^*, h(\delta^*)) \frac{d}{d\delta'} g(t, \delta^*, h(\delta))\, d\mu = E\left[\left(\frac{d}{d\delta} g(t, \delta^*, h(\delta^*))\right)^2\right]$$

$$V_e = \int e(t, \delta^*) e(t, \delta^*)'\, d\mu - h(\delta^*) h(\delta^*)' = V[e(t, \delta^*)]$$

$$W = \int \frac{d^2}{d\delta\, d\,h'} g(t, \delta^*, h(\delta^*))\, d\mu$$

todas existen y son finitas.

Sea $V_{eg} = \int e(t, \delta^*) \frac{d}{d\delta'} g(t, \delta^*, h(\delta^*))\, d\mu$

Entonces

$$\sqrt{N}\left(\frac{d}{d\delta} f_N(s, \delta^*)\right)$$

converge en ley a una $N[0, \Delta]$

donde $\Delta = V_g + W V_{eg} + V_{eg}' W' + W V_e W'$.

La demostración de la Normalidad Asintótica de los estimadores se sigue de los lemas 4 y 5 como se detalla a continuación.

El lema 4 demuestra que los estimadores-máximo verosímiles siguen una distribución asintóticamente Normal, especificando además cual es su matriz de varianzas-covarianzas (su media es el verdadero valor del parámetro, ya que por el lema 1, son consistentes).

Las únicas condiciones exigidas en este lema 4 hacen referencia a la continuidad de las funciones $f_N(s,\cdot)$ y sus derivadas. La expresión $\frac{d^2}{d\delta\, d\delta'} f(\delta^*)$ significa que hay que calcular primero la segunda derivada $\frac{d^2}{d\delta\, d\delta'} f(\delta)$ y después su valor en $\delta = \delta^*$.

En el lema 5 se demuestra la convergencia de $\sqrt{N}\left(\frac{d}{d\delta} f_N(s,\delta^*)\right)$ a una variable Normal especificando cual es su matriz de varianzas-covarianzas (su media vale cero). Únicamente hay que tener en cuenta que las funciones g que se utilizan en la definición de la función f_N presentan una pequeña diferencia con respecto a la definición hecha en el lema 2, a partir de la función g_1, ya que ahora se considera:

$$g(t_i, \delta, h_N(\delta)) \qquad \text{siendo} \qquad h_N(\delta) = \sum_{i=1}^{N} e(t_i, \delta)\frac{1}{N}$$

Para cada estimador se buscará la función h_N adecuada de forma que se puedan aplicar los lemas anteriores.

Por la propia definición, la matriz V_g coincide con la matriz de información asociada al muestreo considerado:

$$V_g = E\left[\left(\frac{d}{d\theta} LL(\theta)\right)^2\right] = J, J_e, J_c$$

donde $L(\theta)$ representa a la función de verosimilitud que corresponda al muestreo utilizado y $LL(\theta)$ es el logaritmo de esta función.

Por otro lado, la matriz F del lema 4 está definida como la esperanza de una segunda derivada, ya que $F = \frac{d^2}{d\theta\, d\theta'} f(\theta)$, donde $f(\theta) = E[LL(\theta)]$ y por lo tanto se tendrá que:

$$F = \frac{d^2}{d\theta\, d\theta'} f(\theta) = \frac{d^2}{d\theta\, d\theta'} E[LL(\theta)] = E\left[\frac{d}{d\theta\, d\theta'} LL(\theta)\right]$$

Puesto que

$$\mathrm{E}\left[\frac{d}{d\theta\, d\theta'}\mathrm{LL}(\theta)\right]=-\mathrm{E}\left[\left(\frac{d}{d\theta}\mathrm{LL}(\theta)\right)^2\right]$$

siempre que L sea la función de verosimilitud asociada al muestreo considerado, se puede comprobar que $\mathrm{F}=-V_g$.

En casi todos los estimadores que se han propuesto se ha obtenido la estimación desde el muestreo considerado, de forma que se verificará, por la propia definición, que la matriz F y la matriz V_g verifican la relación $\mathrm{F}=-V_g=-J$ donde J es la matriz de información asociada al muestreo utilizado.

Como se verá posteriormente, únicamente hay dos estimadores para los cuales la relación anterior no será cierta: el estimador propuesto por Manski y Lerman para $\mathrm{P}(x)$ desconocida y $\mathrm{P}(j)$ conocida, que está definido sobre muestras exógenas, pero que se utiliza en muestreo basado en la elección, y el estimador definido de forma equivalente al anterior, pero para $\mathrm{P}(j)$ también desconocida.

Se comentan a continuación las demostraciones de las propiedades de consistencia y Normalidad asintótica que verifica cada uno de los estimadores propuestos, viendo si cumplen o no las hipótesis exigidas en los lemas anteriores.

Además se compara la matriz de varianzas-covarianzas asintótica con la inversa de la matriz de información de Fisher para probar la eficiencia asintótica.

Las expresiones de la matriz de información para los distintos tipos de muestreo vienen dadas por:

$$J=\sum_{Z}\sum_{C}\mathrm{P}(x)\,\mathrm{P}(j/x,\theta^*)\left[\frac{d}{d\theta}\ln\mathrm{P}(j/x,\theta^*)\right]\left[\frac{d}{d\theta'}\ln\mathrm{P}(j/x,\theta^*)\right]$$

para una muestra aleatoria simple,

$$J_e=\sum_{b\in B}\sum_{Z_b}\mathrm{g}(x/b)\,H(b)\sum_{C}\mathrm{P}(j/x,\theta^*)\left[\frac{d}{d\theta}\ln\mathrm{P}(j/x,\theta^*)\right]\left[\frac{d}{d\theta'}\ln\mathrm{P}(j/x,\theta^*)\right]$$

para muestreo exógeno, y

$$J_c = \sum_C H(j) \sum_z \frac{P(j/x,\theta^*)P(x)}{P(j)} \left[\frac{d}{d\theta} \ln \frac{P(j/x,\theta^*)P(x)}{\sum_z P(j/z,\theta^*)P(z)} \right]$$

$$\left[\frac{d}{d\theta'} \ln \frac{P(j/x,\theta^*)P(x)}{\sum_z P(j/z,\theta^*)P(z)} \right]$$

para muestreo basado en la elección.

Consistencia y Normalidad Asintótica de los estimadores propuestos

Como los modelos binomiales no son más que una particularización de un modelo multinomial en el que el número de alternativas es 2, los estimadores que se utilizan en los modelos binomiales coincidirán con los de los modelos multinomiales, y gozarán de las mismas propiedades. Por lo tanto únicamente se planteará la demostración de las propiedades para los estimadores asociados a modelos multinomiales.

El primero de los estimadores propuestos en el apartado 3.3.1. es el que surgía como estimador máximo-verosímil considerando que las variables explicativas x son fijas. Amemiya (1973) demuestra que este estimador es consistente y asintóticamente Normal, aunque las observaciones y_i/x_i no son idénticamente distribuidas, se modifica la demostración de Rao adecuándola a esta situación. Así, este estimador verifica las propiedades de ser consistente, asintóticamente Normal y eficiente por el hecho de ser el estimador máximo-verosímil.

Cuando las variables explicativas x_i son aleatorias (apartado 3.3.2.), es necesario utilizar los lemas enunciados anteriormente. Para cada estimador se buscarán las funciones adecuadas que verifiquen las hipótesis de los lemas.

Dadas las particularidades de cada uno de los estimadores propuestos se van a analizar por separado. McFadden (1981) indica las lineas generales de las demostraciones de las propiedades de algunos de los estimadores. A continuación se presentan los desarrollos completos de la demostración correspondiente a cada uno de los estimadores presentados en este trabajo.

1. En el caso de considerar una muestra aleatoria o una muestra exógena pero siendo las distribuciones marginales $P(x)$ y $P(j)$ una o ambas desconocidas, el estimador máximo-verosímil viene dado como la solución al problema (3-35):

$$\max_{\theta\in\Theta} \sum_{i=1}^{N} \ln P(j_i / x_i, \theta)$$

Considerando como función $g_1(t, \delta) = \ln P(j/x, \theta)$ *y* $D = \Theta$ se cumplen las condiciones necesarias para la consistencia del estimador, ya que con $g_2(s, \delta) = P(j/x, \theta)$ se encuentra la condición de un único máximo de la función $f(\delta)$ en el verdadero valor del parámetro.

Cuando el muestreo es aleatorio simple:

$$f(\delta) = E[g_1(t, \delta)] = E[\ln P(j/x, \theta)] = \sum_{Z}\sum_{C} \ln P(j/x, \theta)\, P(j/x, \theta^*)\, P(x) =$$

$$\sum_{Z} P(x) \sum_{C} P(j/x, \theta^*)\, \ln P(j/x, \theta)$$

Si el muestreo es exógeno:

$$f(\delta) = E[g_1(t, \delta)] = E[\ln P(j/x, \theta)] =$$

$$\sum_{h\in B}\sum_{Z_h}\sum_{C} \ln P(j/x, \theta)\, P(j/x, \theta^*)\, g(x/b)\, H(b) =$$

$$\sum_{h\in B}\sum_{Z_h} g(x/b)\, H(b) \sum_{C} \ln P(j/x, \theta)\, P(j/x, \theta^*)$$

Para comprobar la Normalidad asintótica basta considerar la función

$$g(t, \delta, h_N(\delta)) = g_1(t, \delta) = \ln P(j/x, \theta)$$

y

$$f_N(s, \delta) = \sum_{i=1}^{N} \frac{1}{N} \ln P(j_i / x_i, \theta)$$

Con estas funciones los lemas 4 y 5 demuestran que $\sqrt{N}\left(\hat{\theta}_N - \theta^*\right)$ converge en ley a una variable aleatoria con distribución $N\left[0\,,\,\Omega^{-1}\Delta\,\Omega^{-1}\right]$ siendo $\Omega^{-1} = \mathrm{F}^{-1}$ y $\Delta = V_g$.

Para la muestra aleatoria simple

$$\mathrm{F} = \frac{d^2}{d\delta\, d\delta'}\mathrm{f}(\delta^*) = -J = -V_{\mathrm{g}}$$

y para el muestreo exógeno $\mathrm{F} = -J_e = -V_{\mathrm{g}}$, donde J y J_e son respectivamente las matrices asintóticas de información para muestras aleatorias simples y muestreo exógeno.

En cualquier caso se tendrá que $\Omega^{-1}\Delta\,\Omega^{-1} = \mathrm{F}^{-1}$, es decir, las matrices de varianzas-covarianzas asintóticas coinciden con la inversa de la matriz de información para ambos estimadores, lo que demuestra que los dos son asintóticamente eficientes.

2. El segundo estimador que aparecía en el desarrollo del epígrafe de estimación anterior viene definido como la solución al problema (3-36)

$$\max_{\theta\in\Theta_0}\sum_{i=1}^{N}\ln\mathrm{P}(j_i / x_i, \theta)$$

Este estimador corresponde a un muestreo exógeno con las dos distribuciones marginales $\mathrm{P}(x)$ y $\mathrm{P}(j)$ conocidas, y coincide también con el estimador correspondiente a un muestreo basado en la elección bajo la restricción de conocer ambas distribuciones marginales.

Se puede observar que este estimador no es más que un caso particular del analizado en el punto anterior y por lo tanto la consistencia ya demostrada para el primero garantiza la consistencia del estimador propuesto ahora.

Análogamente la demostración de la Normalidad asintótica del estimador anterior sirve para demostrar la Normalidad asintótica de este estimador que es una versión restringida del primero.

No obstante, a diferencia del anterior este estimador no es eficiente bajo ninguno de los muestreos, ya que al existir una restricción en el problema a

maximizar, la matriz de varianzas-covarianzas asintótica que tiene este estimador será:

$$\Omega^{-1}\Delta\Omega^{-1} = J_e^{-1} - J_e^{-1}R\left(R'J_e^{-1}R\right)^{-1}R'J_e^{-1}$$

puesto que $\mathrm{F} = -J_e = -V_{\mathrm{g}}$, $\Delta = V_{\mathrm{g}}$ y

$$R = \frac{d}{d\delta}r(\delta^*) = \frac{d}{d\theta}\left[\sum_z \mathrm{P}(j/z,\theta^*)\,\mathrm{P}(z) - \mathrm{P}(j)\right] = \sum_z \frac{d}{d\theta}\mathrm{P}(j/z,\theta^*)\mathrm{P}(z), \quad j \in C$$

y

$$\Omega^{-1} = -J_e^{-1} + J_e^{-1}R\left(R'J_e^{-1}R\right)^{-1}R'J_e^{-1}$$

bajo muestreo exógeno y cambiando J_e por J_c, se obtiene que

$$\Omega^{-1}\Delta\Omega^{-1} = J_c^{-1} - J_c^{-1}R\left(R'J_c^{-1}R\right)^{-1}R'J_c^{-1}$$

bajo muestreo basado en la elección.

Como puede apreciarse estas matrices no coinciden con la inversa de la matriz de información en ningún caso.

3. El estimador que se va a analizar ahora es el correspondiente a un muestreo basado en la elección cuando la distribución marginal $\mathrm{P}(x)$ es conocida y la distribución $\mathrm{P}(j)$ desconocida.

Este estimador viene dado como la solución al problema (3-37)

$$\max_{\theta\in\Theta}\left[\sum_{i=1}^{N}\ln\mathrm{P}(j_i/x_i,\theta) - \sum_{i=1}^{N}\ln\sum_z \mathrm{P}(j_i/z,\theta)\,\mathrm{P}(z)\right]$$

Para demostrar la consistencia de este estimador se considera la función

$$\mathrm{g}_1(t,\delta) = \ln\frac{\mathrm{P}(j/x,\theta)}{\sum_z \mathrm{P}(j/z,\theta)\,\mathrm{P}(z)} \quad y \quad D = \Theta$$

Por el lema 2, cuando $N \to \infty$, $f_N(s, \delta)$ converge uniformemente para casi todo $s \in S$ sobre Θ a $f(\delta)$, donde

$$f(\delta) = E[g_1(t, \delta)] = \sum_C H(j) \sum_z \frac{P(j/x, \theta^*)P(x)}{P(j)} \ln \frac{P(j/x, \theta)P(x)}{\sum_z P(j/z, \theta)P(z)} + K$$

Considerando

$$g_2(s, \delta) = \frac{P(j/x, \theta^*)\, P(x)}{P(j)}$$

por el lema 3 la función $f(\delta)$ alcanza un único máximo en $\delta = \theta^*$ y por el lema 1 el estimador es consistente para el parámetro θ^*.

En la demostración de la Normalidad asintótica basta considerar la siguiente función

$$g(t, \delta, h_N(\delta)) = g_1(t, \delta) = \ln \frac{P(j/x, \theta)}{\sum_z P(j/z, \theta)P(z)}$$

y

$$f_N(s, \delta) = \sum_{i=1}^{N} \frac{1}{N} \ln \frac{P(j_i/x_i, \theta)}{\sum_z P(j_i/z, \theta)P(z)}$$

Los lemas 4 y 5 demuestran que $\sqrt{N}\left(\hat{\theta}_N - \theta^*\right)$ converge en ley a una $N[0, \Omega^{-1}\Delta\,\Omega^{-1}]$ donde $\Omega^{-1} = F^{-1}$, $\Delta = V_g$ siendo $F = -J_c = -V_g$ de donde $\Omega^{-1}\Delta\,\Omega^{-1} = J_c^{-1}$.

Puesto que se verifica que la matriz de varianzas-covarianzas de este estimador coincide con la inversa de la matriz de información se puede afirmar que este estimador es eficiente.

4. A continuación se presentan las demostraciones de las propiedades para los estimadores correspondientes a un muestreo basado en la elección siendo la distribución marginal $P(x)$ desconocida y $P(j)$ conocida.

4.1 El primero de los estimadores obtenidos para esta situación es el propuesto por Manski y Lerman (1977) y viene definido por el problema de maximización dado por (3-38):

$$\max_{\theta \in \Theta} \sum_{i=1}^{N} w(j_i) \ln P(j_i / x_i, \theta)$$

con

$$w(j_i) = \frac{P(j_i)}{H(j_i)}$$

En la demostración de la consistencia basta considerar la función

$$g_1(t, \delta) = \frac{P(j)}{H(j)} \ln P(j/x, \theta) \quad y \quad D = \Theta$$

Por el lema 2, cuando $N \to \infty$, $f_N(s, \delta)$ converge uniformemente para casi todo $s \in S$ sobre Θ a $f(\delta)$ donde

$$f(\delta) = \sum_z \sum_C \frac{P(j)}{H(j)} \ln P(j/x, \theta) \frac{P(j/x, \theta^*) P(x)}{P(j)} H(j) =$$

$$\sum_z P(x) \sum_C P(j/x, \theta^*) \ln P(j/x, \theta)$$

Considerando $g_2(s, \delta) = P(j/x, \theta)$ el lema 3 asegura que la función $f(\delta)$ alcanza un único máximo en el verdadero valor del parámetro $\delta = \theta^*$ y por el lema 1 se puede afirmar que el estimador es consistente para θ^*.

Con la función

$$g(t, \delta, h_N(\delta)) = g_1(t, \delta) = \frac{P(j)}{H(j)} \ln P(j/x, \theta)$$

y

$$f_N(s,\delta)=\sum_{i=1}^{N}\frac{1}{N}\frac{P(j_i)}{H(j_i)}\ln P(j_i/x_i,\theta)$$

los lemas 4 y 5 aseguran la Normalidad asintótica del estimador, ya que se tendrá que $\sqrt{N}\left(\hat{\theta}_N-\theta^*\right)$ converge en ley a una $N\left[0,\Omega^{-1}\Delta\,\Omega^{-1}\right]$ con $\Omega^{-1}=F^{-1}$ y $\Delta=V_g$ siendo en este caso

$$F=\sum_{Z}P(z)\sum_{C}P(j/x,\theta^*)\frac{d}{d\theta}\ln P(j/x,\theta^*)\frac{d}{d\theta'}\ln P(j/x,\theta^*)$$

como se considera $\frac{P(j)}{H(j)}\ln P(j/x,\theta)$ que no coincide con el logaritmo de la función de verosimilitud, la matriz F no es igual a $-V_g$:

$$V_g=\sum_{Z}\sum_{C}P(j/x,\theta^*)\frac{P(j)}{H(j)}\frac{d}{d\theta}\ln P(j/x,\theta^*)\frac{d}{d\theta'}\ln P(j/x,\theta^*)$$

Según estos resultados este estimador no sería asintóticamente eficiente, ya que su matriz de varianzas-covarianzas es $\Omega^{-1}\Delta\,\Omega^{-1}=F^{-1}V_g\,F^{-1}$ y no coincide con la inversa de la matriz de información.

4.2 El segundo estimador propuesto para esta situación informativa es el de Cosslett (1981a), y que está definido como la solución al problema (3-39):

$$\max_{\theta\in\Theta}\min_{\lambda\in\Delta_2}\sum_{i=1}^{N}\ln\frac{P(j_i/x_i,\theta)}{\sum_{C}\lambda(k)P(k/x_i,\theta)}$$

siendo

$$\Delta_2=\left\{\lambda\ /\ \sum_{C}\lambda(k)P(k)=1,\ \sum_{C}\lambda(k)P(k/x_i,\theta)>0,\quad i=1,.....N\right\}$$

A diferencia de los estimadores anteriores, ahora nos encontramos con dos parámetros, θ y λ. No pueden considerarse los dos simultáneamente. En primer lugar hay que resolver el problema de minimizar respecto a λ y después se maximizará para θ.

Reescribiendo la función objetivo como

$$\sum_{i=1}^{N} \ln \mathrm{P}(j_i / x_i, \theta) - \sum_{i=1}^{N} \ln \sum_{C} \lambda(k) \mathrm{P}(k / x_i, \theta)$$

minimizar para λ equivale a maximizar para λ la expresión

$$\sum_{i=1}^{N} \ln \sum_{C} \lambda(k) \mathrm{P}(k / x_i, \theta)$$

El estimador para λ es consistente ya que si se toma la función

$$\mathrm{g}_1^{\lambda}(t, \delta) = \ln \sum_{C} \lambda(k) \mathrm{P}(k / x, \theta)$$

y

$$\lambda^*(j) = \frac{H(j)}{\mathrm{P}(j)}$$

restringiendo los valores de λ por la relación $k^{-1}\lambda^*(j) < \lambda(j) < k\lambda^*(j)$ el lema 2 asegura que la sucesión de funciones f_N^{λ} definidas como

$$\mathrm{f}_N^{\lambda}(s, \delta) = \sum_{i=1}^{N} \frac{1}{N} \mathrm{g}_1^{\lambda}(t_i, \delta)$$

converge uniformemente para casi todo $s \in S$ sobre Δ_2 a $\mathrm{f}^{\lambda}(\delta)$ donde

$$\mathrm{f}^{\lambda}(\delta) = \sum_{Z} \sum_{C} \mathrm{P}(x) \frac{\mathrm{P}(j / x, \theta^*) H(j)}{\mathrm{P}(j)} \ln \sum_{C} \lambda(k) \mathrm{P}(k / x, \theta)$$

El lema 3 comprueba la existencia de un único máximo para la función $\mathrm{f}^{\lambda}(\delta)$ y el lema 1 concluye la consistencia:

$$\hat{\lambda}_N \longrightarrow \lambda^*$$

donde $\lambda^*(j) = \dfrac{H(j)}{\mathrm{P}(j)}$

En consecuencia λ^* es un mínimo para λ de la función objetivo.

Tomando este valor de λ únicamente faltará ver la maximización para θ:

$$\max_{\theta\in\Theta}\sum_{i=1}^{N}\ln\frac{\mathrm{P}(j_i/x_i,\theta)}{\sum_{C}\frac{H(k)}{\mathrm{P}(k)}\mathrm{P}(k/x_i,\theta)}$$

Se considera ahora la siguiente función:

$$g_1(t,\delta)=\ln\frac{\mathrm{P}(j/x,\theta)}{\sum_{C}\frac{H(k)}{\mathrm{P}(k)}\mathrm{P}(k/x,\theta)}$$

Aplicando el lema 2 a la función $g_1(t,\delta)$ se obtiene la convergencia uniforme de

$$\mathrm{f}_N(s,\delta)=\sum_{i=1}^{N}\frac{1}{N}g_1(t_i,\delta)$$

a la función

$$\mathrm{f}(\delta)=\sum_{Z}\mathrm{P}(x)\sum_{C}\frac{\mathrm{P}(j/x,\theta^*)\mathrm{P}(x)H(j)}{\mathrm{P}(j)}\ln\frac{\mathrm{P}(j/x,\theta)}{\sum_{C}\frac{H(k)}{\mathrm{P}(k)}\mathrm{P}(k/x,\theta)}$$

Aplicando el lema 3 se demuestra la existencia de un único máximo en $\theta=\theta^*$ y el lema 1 concluye la consistencia.

En todo este desarrollo se ha hecho uso de unas funciones para la aplicación de los lemas correspondientes. Estas funciones deben de cumplir una serie de condiciones, como el estar acotadas. Aunque no se presenta aquí el correspondiente análisis, Cosslett (1981a) demuestra que las funciones utilizadas en su demostración están convenientemente acotadas y cumplen las hipótesis de los lemas.

En la minimización para λ ha sido necesario restringir el conjunto de sus posibles valores. Al considerar por separado la minimización para este parámetro λ y la maximización para θ pueden aparecer problemas si la función objetivo no tiene el máximo en Δ_2. Esto sólo puede ocurrir en el caso de no estar acotado.

Este problema lo comenta Cosslett en su desarrollo y lo soluciona con una extensión suave del lema de Amemiya.

En lo referente a la Normalidad asintótica la demostración es bastante sencilla si se considera la siguiente función:

$$g(t\,,\delta,\mathrm{h}_N(\delta)) = \ln\frac{\mathrm{P}(j\,/\,x,\theta)}{\sum_C \lambda(k)\mathrm{P}(k\,/\,x,\theta)}$$

ya que por el lema 5 se verifica que

$$\sqrt{N}\,\frac{d}{d\delta}\mathrm{f}_N(s\,,\delta) = \sqrt{N}\,\frac{d}{d\delta}\left[\sum_{i=1}^{N}\frac{1}{N}\ln\frac{\mathrm{P}(j_i\,,x_i\,,\theta)}{\sum_C \lambda(k)\mathrm{P}(k\,/\,x_i\,,\theta)}\right]$$

converge en ley a una $N[0\,,\Delta]$, donde $\Delta = V_g$.

Por el lema 4, $\sqrt{N}\left(\hat{\delta}_N - \delta^*\right)$ converge en ley a una $N[0\,,\Omega^{-1}\Delta\,\Omega^{-1}]$ con $\Omega^{-1} = \mathrm{F}^{-1}$ siendo $\delta = (\theta\,,\lambda)$ y $\delta^* = (\theta^*,\lambda^*)$ y $\mathrm{F} = -V_g$ por construcción.

En este caso, el estimador sí es eficiente asintóticamente, ya que según las relaciones anteriores, la matriz de varianzas-covarianzas del estimador cumple que:

$$\Omega^{-1}\Delta\,\Omega^{-1} = \mathrm{F}^{-1}\,\Delta\,\mathrm{F}^{-1} = \left(-V_g\right)^{-1}V_g\left(-V_g\right)^{-1} = V_g^{-1}$$

es decir, coincide con la inversa de la matriz de información.

Notar que en la estimación se ha obtenido un estimador para el parámetro λ y otro para el parámetro que realmente interesaba θ. Si se desea encontrar la matriz de varianzas-covarianzas asociada al último de los dos basta particionar la matriz de información según el vector de parámetros $(\theta\,,\lambda)$ como:

$$\begin{bmatrix}\Delta_{\theta\theta} & \Delta_{\theta\lambda}\\ \Delta_{\lambda\theta} & \Delta_{\lambda\lambda}\end{bmatrix}$$

se obtiene que la matriz de varianzas-covarianzas para el estimador $\hat{\theta}$ es:

$$V_{\hat{\theta}} = \left(\Delta_{\theta\theta} - \Delta_{\theta\lambda}\, \Delta_{\lambda\lambda}^{-1}\, \Delta_{\lambda\theta}\right)^{-1}$$

4.3 El siguiente estimador que se propuso para la situación de muestreo basado en la elección con la distribución $\mathrm{P}(x)$ desconocida y $\mathrm{P}(j)$ conocida, es la solución al problema (3-42):

$$\max_{\theta\in\Theta} \sum_{i=1}^{N} \left[\ln \mathrm{P}(j_i / x_i, \theta) - \ln \left[\sum_{C} \frac{\mathrm{P}(k)}{N_k} \sum_{m\in N(k)} \mathrm{P}(j_i / x_m, \theta) \right] \right]$$

La demostración de la consistencia de este estimador se realiza utilizando la función siguiente:

$$g_1(t, \delta) = \ln \mathrm{P}(j / x, \theta) - \ln B_N(j / \theta)$$

donde

$$B_N(j / \theta) = \sum_{C} \frac{\mathrm{P}(k)}{N_k} \sum_{m\in N(k)} \mathrm{P}(j / x_m, \theta) \quad \text{y} \cdot D = \Theta.$$

Para demostrar la convergencia de la función $\mathrm{f}_N(s, \delta)$ se considerará por separado la convergencia de los dos sumandos anteriores.

Por un lado el lema 2 asegura que cuando $N \to \infty$, $\frac{1}{N} \sum_{i=1}^{N} \ln \mathrm{P}(j_i / x_i, \theta)$ converge uniformemente para casi todo (j, x) sobre Θ a

$$\mathrm{E}\left[\ln \mathrm{P}(j / x, \theta)\right] = \sum_{Z} \sum_{C} \frac{\mathrm{P}(j / x, \theta^*)\mathrm{P}(x)}{\mathrm{P}(j)} H(j) \ln \mathrm{P}(j / x, \theta)$$

Además, por el lema 2, cuando $N \to \infty$, $\frac{1}{N_k} \sum_{m\in N_k} \ln \mathrm{P}(j / x_m, \theta)$ converge uniformemente para casi todo (j, x) sobre Θ a

$$\sum_{Z} \mathrm{P}(j / x, \theta) \frac{\mathrm{P}(k / x, \theta^*)\, \mathrm{P}(x)}{\mathrm{P}(k)}$$

Desde este resultado se tiene que $B_N(j/\theta)$ converge uniformemente para casi todo j sobre Θ a $\sum_z \mathrm{P}(j/x,\theta)\,\mathrm{P}(x)$ y, por lo tanto, se afirma que $\ln B_N(j/\theta)$ converge casi seguramente a $\ln \sum_z \mathrm{P}(j/x,\theta)\,\mathrm{P}(x)$.

Es decir, $\sum_C \frac{N_j}{N} \ln B_N(j/\theta)$ converge casi seguramente a la función:

$$\sum_C H(j) \ln \sum_z \mathrm{P}(j/x,\theta)\,\mathrm{P}(x)$$

Considerando los resultados anteriores se tiene que $\mathrm{f}_N(s,\delta)$ converge uniformemente para casi todo $s \in S$ a $\mathrm{f}(\delta)$, donde

$$\mathrm{f}(\delta) = \sum_C H(j) \sum_z \frac{\mathrm{P}(j/x,\theta^*)\mathrm{P}(x)}{\mathrm{P}(j)} \ln \frac{\mathrm{P}(j/x,\theta)\mathrm{P}(x)}{\sum_z \mathrm{P}(j/z,\theta)\mathrm{P}(z)} + K$$

Por el lema 3, $\mathrm{f}(\delta)$ alcanza un único máximo en θ^* y por el lema 1 el estimador es consistente para θ^*.

Para demostrar la Normalidad asintótica se utiliza la función

$$g(t_i,\delta,\mathrm{h}_N(\delta)) = \ln \mathrm{P}(j_i/x_i,\theta) - \frac{1}{N}\sum_{i=1}^{N} \ln\left[\sum_C \frac{\mathrm{P}(k)}{N_k} \sum_{m\in N_k} \mathrm{P}(j_i/x_m,\theta)\right]$$

donde

$$\mathrm{h}_N(\delta) = \frac{1}{N}\sum_{i=1}^{N} \ln\left[\sum_C \frac{\mathrm{P}(k)}{N_k} \sum_{m\in N_k} \mathrm{P}(j_i/x_m,\theta)\right]$$

y

$$e(t_i,\delta) = \ln\left[\sum_C \frac{\mathrm{P}(k)}{N_k} \sum_{m\in N_k} \mathrm{P}(j_i/x_m,\theta)\right]$$

Los lemas 4 y 5 demuestran que $\sqrt{N}\left(\hat{\theta}_N - \theta^*\right)$ converge en ley a una variable aleatoria con distribución $N\left[0, \Omega^{-1}\Delta\,\Omega^{-1}\right]$ siendo $\Omega^{-1} = \mathrm{F}^{-1}$, $\mathrm{F} = -V_g$ y

$$\Delta = V_g + W\,V_{eg} + V_{eg}'\,W' + W\,V_e\,W'$$

La matriz de varianzas-covarianzas del estimador viene dada por $\mathrm{F}^{-1}\,\Delta\,\mathrm{F}^{-1} \neq J_C^{-1}$ que no coincide con la inversa de la matriz de información y por lo tanto se puede afirmar que este estimador no es asintóticamente eficiente.

4.4 Otro estimador posible para esta situación es el que se obtiene como solución al problema (3-40)

$$\max_{\theta\in\Theta} \sum_{i=1}^{N} \ln \mathrm{P}(j_i / x_i, \theta)$$

sujeto a

$$\mathrm{P}(j) = \sum_{k\in C} \frac{\mathrm{P}(k)}{N_k} \sum_{m\in N(k)} \mathrm{P}(k / x_m, \theta) \quad , j \in C$$

Es fácil comprobar que este estimador es consistente, ya que es una versión restringida del estimador anterior. Para la Normalidad asintótica también se utiliza el mismo resultado.

Así este estimador es consistente y asintóticamente Normal por serlo el estimador anterior. Sin embargo, la matriz de varianzas-covarianzas de este estimador no coincide con la del anterior, ya que en este caso hay una restricción en la definición que hay que tener en cuenta.

Ahora se tiene que $\mathrm{F} = -V_g$, $\Delta = V_g$, $\Omega^{-1} = \mathrm{F}^{-1} - \mathrm{F}^{-1}\,R\left(R'\,\mathrm{F}^{-1}\,R\right)^{-1} R'\,\mathrm{F}^{-1}$, siendo $V_g = -J_C$ la matriz de información y

$$R = \frac{d}{d\delta} r(\delta^*) = \frac{d}{d\theta}\left[\sum_{k\in C} \frac{\mathrm{P}(k)}{N_k} \sum_{m\in N(k)} \mathrm{P}(k / x_m, \theta^*) - \mathrm{P}(j)\right] =$$

$$\sum_{k\in C} \frac{\mathrm{P}(k)}{N_k} \sum_{m\in N(k)} \frac{d}{d\theta} \mathrm{P}(k / x_m, \theta^*)$$

La matriz de varianzas-covarianzas del estimador es:

$$\Omega^{-1}\Delta\,\Omega^{-1} = -\left(\mathrm{F}^{-1} - \mathrm{F}^{-1} R\left(R'\,\mathrm{F}^{-1} R\right)^{-1} R'\,\mathrm{F}^{-1}\right)$$

que no coincide tampoco con la inversa de la matriz de información, y en consecuencia, este estimador no es asintóticamente eficiente.

4.5 El último estimador propuesto bajo muestreo basado en la elección con la distribución marginal $\mathrm{P}(x)$ desconocida y la distribución $\mathrm{P}(j)$ conocida viene dado por (3-43):

$$\max_{\theta\in\Theta} \sum_{i=1}^{N} \ln \frac{\mathrm{P}(j_i / x_i, \theta)^{H(j_i)/\mathrm{P}(j_i)}}{\sum_{C} \mathrm{P}(k / x_i, \theta)^{H(k)/\mathrm{P}(k)}}$$

La demostración de la consistencia se basa en la función

$$\mathrm{g}_1(t, \delta) = \ln \frac{\mathrm{P}(j/x, \theta)^{H(j)/\mathrm{P}(j)}}{\sum_{C} \mathrm{P}(k/x, \theta)^{H(k)/\mathrm{P}(k)}} \quad y \quad D = \Theta$$

Por el lema 2, cuando $N \to \infty$, $\mathrm{f}_N(s, \delta)$ converge uniformemente para casi todo $s \in S$ sobre Θ a $\mathrm{f}(\delta)$, donde

$$\mathrm{f}(\delta) = \sum_{Z} \mathrm{q}(x) \sum_{C} \frac{\mathrm{P}(j/x, \theta^*)^{H(j)/\mathrm{P}(j)}}{\sum_{C} \mathrm{P}(k/x, \theta^*)^{H(k)/\mathrm{P}(k)}} \ln \frac{\mathrm{P}(j/x, \theta)^{H(j)/\mathrm{P}(j)}}{\sum_{C} \mathrm{P}(k/x, \theta)^{H(k)/\mathrm{P}(k)}}$$

siendo

$$\mathrm{q}(x) = \sum_{C} \frac{\mathrm{P}(j/x, \theta^*)\,\mathrm{P}(x)}{\mathrm{P}(j)} H(j)$$

Considerando

$$g_2(t,\delta)=\frac{P(j/x,\theta^*)^{H(j)/P(j)}}{\sum_C P(k/x,\theta^*)^{H(k)/P(k)}}$$

por el lema 3 la función $f(\delta)$ alcanza un único máximo en $\delta=\theta^*$ y por el lema 1 el estimador es consistente para θ^*.

Para demostrar que el estimador es asintóticamente Normal se utiliza la función

$$g(t,\delta,h_N(\delta))=g_1(t,\delta)=\ln\frac{P(j/x,\theta)^{H(j)/P(j)}}{\sum_C P(k/x,\theta)^{H(k)/P(k)}}$$

y por el lema 5 se verifica que

$$\sqrt{N}\,\frac{d}{d\delta}f_N(s,\delta)=\sqrt{N}\,\frac{d}{d\delta}\left[\sum_{i=1}^{N}\frac{1}{N}\ln\frac{P(j_i/x_i,\theta)^{H(j_i)/P(j_i)}}{\sum_C P(k/x_i,\theta)^{H(k)/P(k)}}\right]$$

converge en ley a una $N[0,\Delta]$ donde $\Delta=V_g$.

Por el lema 4 $\sqrt{N}\left(\hat{\theta}_N-\theta^*\right)$ converge en ley a una $N[0,\Omega^{-1}\Delta\,\Omega^{-1}]$ con $\Omega^{-1}=F^{-1}$ y $F=-V_g$.

Así, este estimador es asintóticamente eficiente puesto que se verifica que la matriz de varianzas-covarianzas coincide con la inversa de la matriz de información:

$$\Omega^{-1}\Delta\,\Omega^{-1}=(-J_c)^{-1}J_c(-J_c)^{-1}=J_c^{-1}$$

5. Para la situación de tener un muestreo basado en la elección con ambas distribuciones marginales desconocidas, los estimadores propuestos estaban

basados en los mismos que se utilizaban cuando la distribución marginal $P(j)$ era conocida.

Se va a demostrar que de los cuatro estimadores propuestos únicamente el de Cosslett y el último son consistentes.

5.1 El primero de los estimadores que se han propuesto para esta situación es el (3-44) de Manski y Lerman:

$$\max_{(\theta,\ddot{P})\in\Theta\times\Pi} \frac{1}{N}\sum_{i=1}^{N}\frac{\ddot{P}(j_i)}{H(j_i)}\ln P(j_i / x_i,\theta)$$

con

$$\Pi=\left\{\ddot{P} \quad / \sum_{C}\ddot{P}(j)=1\right\}$$

Este estimador y el que viene dado como la solución al problema (3-46)

$$\max_{(\theta,\ddot{P})\in\Theta\times\Pi}\left\{\frac{1}{N}\sum_{i=1}^{N}\ln P(j_i / x_i,\theta)-\frac{1}{N}\sum_{i=1}^{N}\ln\left[\sum_{C}\frac{\ddot{P}(k)}{N_k}\sum_{m\in N(k)}P(j_i / x_m,\theta)\right]\right\}$$

no son, en general, consistentes. Para verlo basta examinar el comportamiento de la función objetivo en el límite.

Para el primero de los dos se tiene que:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\ddot{P}(j_i)}{H(j_i)}\ln P(j_i / x_i,\theta)\xrightarrow{c.s.}\sum_{j\in C}\ddot{P}(j)\sum_{z}\frac{P(j/x,\theta)}{P(j)}P(x)\ln P(j/x,\theta)$$

y esta función es lineal en $\ddot{P}(j)$, alcanzando el máximo en uno de los vértices del simplex unidad, ya que toda función lineal alcanza el máximo en uno de los extremos del intervalo donde está definida. Además para cada $j\in C$ la suma

$$\sum_{z}\frac{P(j/x,\theta^*)}{P(j)}P(x)\ln P(j/x,\theta)$$

no será maximizada generalmente en $\theta = \theta^*$. Por lo tanto este estimador puede no ser consistente para θ^* o para P.

Para el segundo estimador considerado se verifica que la función objetivo converge a:

$$\sum_Z \sum_C \frac{P(j/x,\theta^*)\,P(x)\,H(j)}{P(j)} \ln P(j/x,\theta) -$$

$$\sum_C H(j) \ln\left[\sum_C \ddot{P}(k) \sum_Z P(j/x,\theta)\, q(x/k)\right]$$

y considerándola para $\theta = \theta^*$ como una función de $\ddot{P}$, se observa que el valor $\ddot{P} \in \Pi$ que minimiza el segundo término, y por lo tanto maximiza la función objetivo, dependerá de las cantidades $H(j)$. Así $\ddot{P} = P$ no tiene porque ser el valor que maximice la función objetivo. En consecuencia, este estimador puede no ser consistente.

Puesto que ninguno de los dos estimadores es consistente, no se puede demostrar tampoco la Normalidad asintótica, ya que los lemas 4 y 5 utilizan en su demostración la consistencia de los estimadores.

5.2 El estimador propuesto por Cosslett cuando $P(x)$ y $P(j)$ son ambas desconocidas bajo muestreo basado en la elección se define como (3-45)

$$\max_{\theta\in\Theta} \max_{\lambda\in\Delta} \sum_{i=1}^{N} \ln \frac{\lambda(j_i)\,P(j_i/x_i,\theta)}{\sum_C \lambda(k)\,P(k/x_i,\theta)}$$

siendo

$$\Delta = \left\{ \lambda \quad / \lambda(j) \geq 0 \quad y \quad \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sum_C \lambda(j) P(j/x_i,\theta)} = 1 \right\}$$

Para demostrar la consistencia de este estimador se considera la función

$$g_1(t,\delta) = \ln \frac{\lambda(j)\,P(j/x,\theta)}{\sum_C \lambda(k)\,P(k/x,\theta)} \quad y \quad D = \Theta \times \Delta$$

Cosslett (1981a) demuestra que restringiendo los valores del parámetro λ por la relación $k^{-1}\,\lambda^*(j) < \lambda(j) < k\,\lambda^*(j)$, siendo k una constante mayor que 1, esta función cumple las condiciones necesarias para poder aplicar los lemas correspondientes, y el lema 2 aplicado a la función $g_1(t,\delta)$ anterior asegura que $f_N(s,\delta) = \sum_{i=1}^{N} \frac{1}{N}\, g_1(t_i,\delta)$ converge uniformemente para casi todo $s \in S$ sobre $\Theta \times \Delta$ a

$$f(\delta) = \sum_C \sum_Z \frac{P(j/x,\theta^*)\,P(x)\,H(j)}{P(j)} \ln \left\{ \frac{\lambda(j)\,P(j/x,\theta)}{\sum_C \lambda(k)\,P(k/x,\theta)} \right\}$$

Y por el lema 3 la función

$$G(s,\delta) = \sum_C \frac{\lambda^*(j)\,P(j/x,\theta^*)}{\sum_C \lambda^*(k)\,P(k/x,\theta^*)} \ln \frac{\lambda(j)\,P(j/x,\theta)}{\sum_C \lambda(k)\,P(k/x,\theta)}$$

alcanza un único máximo en $\theta = \theta^*$ y $\lambda = \lambda^*$.

Por otro lado, la función $f(\delta)$ puede escribirse como

$$f(\delta) = \sum_Z \sum_C P(x) \frac{P(r/x,\theta^*)\,H(r)}{P(r)} G(s,\delta)$$

así que $f(\delta)$ alcanza un único máximo en $\theta = \theta^*$ y $\lambda = \lambda^*$.

El lema 1 concluye la consistencia del estimador para θ^* y λ^*.

En esta demostración, Cosslett utiliza una acotación para el parámetro λ. En el caso de que esta acotación no se verificara no cambia el resultado, ya que es fácil demostrar que la restricción para el parámetro λ no tiene ningún efecto para tamaños de muestra elevados, es decir, cuando N es suficientemente grande.

La demostración de la Normalidad asintótica es inmediata si se considera la función

$$g(t, \delta, h_N(\delta)) = \ln \frac{\lambda(j)\, P(j/x, \theta)}{\sum_{C} \lambda(k)\, P(k/x, \theta)}$$

Por el lema 5 se verifica que

$$\sqrt{N}\, \frac{d}{d\delta} f_N(\delta) = \sqrt{N}\, \frac{d}{d\delta} \left[\sum_{i=1}^{N} \frac{1}{N} \ln \frac{\lambda(j_i)\, P(j_i/x_i, \theta)}{\sum_{C} \lambda(k)\, P(k/x_i, \theta)} \right]$$

converge en ley a una $N[0, \Delta]$ donde $\Delta = V_g$.

Por el lema 4 $\sqrt{N}\left(\hat{\delta}_N - \delta^*\right)$ converge en ley a una variable aleatoria con distribución $N[0, \Omega^{-1}\Delta\,\Omega^{-1}]$, siendo $\delta = (\theta, \lambda)$, $\delta^* = (\theta^*, \lambda^*)$, $\Omega^{-1} = F^{-1}$ y $F = -V_g$.

En este caso se tendrá que $\Omega^{-1}\Delta\,\Omega^{-1} = F^{-1}\Delta\,F^{-1} = \left(-V_g\right)^{-1} V_g \left(-V_g\right)^{-1} = V_g^{-1}$, es decir, coincide con la inversa de la matriz de información y por lo tanto el estimador es asintóticamente eficiente.

Particionando la matriz de información según el vector de parámetros (θ, λ) como

$$\begin{bmatrix} \Delta_{\theta\theta} & \Delta_{\theta\lambda} \\ \Delta_{\lambda\theta} & \Delta_{\lambda\lambda} \end{bmatrix}$$

se obtiene que la matriz de varianzas-covarianzas para el estimador $\hat{\theta}$ viene dada por

$$V_{\hat{\theta}} = \left(\Delta_{\theta\theta} - \Delta_{\theta\lambda}\, \Delta_{\lambda\lambda}^{-1}\, \Delta_{\lambda\theta}\right)^{-1}$$

5.3. El otro estimador propuesto para esta situación viene dado como la solución al problema (3-47)

$$\max_{(\theta,\ddot{P})\in\Theta\times\Pi} \frac{1}{N}\sum_{i=1}^{N}\ln\frac{P(j_i/x_i,\theta)^{H(j_i)/\ddot{P}(j_i)}}{\sum_C P(k/x_i,\theta)^{H(k)/\ddot{P}(k)}}$$

Como se ha comentado, este estimador, junto con el de Cosslett que ya ha sido analizado, es el único que posee la propiedad de la consistencia. Para la demostración se considerará la función

$$g_1(t,\delta)=\ln\frac{P(j/x,\theta)^{H(j)/\ddot{P}(j)}}{\sum_C P(k/x,\theta)^{H(k)/\ddot{P}(k)}} \quad y \quad D=\Theta\times\Pi$$

Aplicando el lema 2 se tiene que cuando $N\to\infty$, $f_N(s,\delta)$ converge uniformemente para casi todo $s\in S$ sobre $\Theta\times\Pi$ a $f(\delta)$ donde

$$f(\delta)=\sum_z q(x)\sum_C \frac{P(j/x,\theta^*)^{H(j)/P(j)}}{\sum_C P(k/x,\theta^*)^{H(k)/P(k)}}\ln\frac{P(j/x,\theta)^{H(j)/\ddot{P}(j)}}{\sum_C P(k/x,\theta)^{H(k)/\ddot{P}(k)}}$$

siendo

$$q(x)=\sum_C \frac{P(j/x,\theta^*)\,P(x)}{P(j)}H(j)$$

Considerando

$$g_2(s,\delta)=\frac{P(j/x,\theta)^{H(j)/\ddot{P}(j)}}{\sum_C P(k/x,\theta)^{H(k)/\ddot{P}(k)}}$$

por el lema 3 la función $f(\delta)$ alcanza un máximo en $\delta^* = (\theta^*, P)$. Sin embargo ahora no se puede garantizar la unicidad del máximo directamente. Será necesario modificar las hipótesis de regularidad para conseguirlo:

Para cada $\delta = (\theta, \ddot{P}) \in \Theta \times \Pi$ tal que $(\theta, \ddot{P}) \neq (\theta^*, P)$ existe $A \subset C \times Z$ de forma que

$$\sum_A q(x) \frac{P(j/x,\theta)^{H(j)/\ddot{P}(j)}}{\sum_C P(k/x,\theta)^{H(k)/\ddot{P}(k)}} \neq \sum_A q(x) \frac{P(j/x,\theta^*)^{H(j)/P(j)}}{\sum_C P(k/x,\theta^*)^{H(k)/P(k)}}$$

Además el proceso de muestreo estratificado satisface

$$\sum_{A_b} P(j/x,\theta) P(x) > 0$$

entonces $H(b) > 0$ para cada $b \in B$ y $\bigcup_{b \in B} A_b = C \times Z$.

Por el lema 1 el estimador es consistente para $\delta^* = (\theta^*, P)$.

En lo referente a la Normalidad asintótica, basta considerar la función

$$g(t, \delta, h_N(\delta)) = g_1(t, \delta) = \ln \frac{P(j/x,\theta)^{H(j)/\ddot{P}(j)}}{\sum_C P(k/x,\theta)^{H(k)/\ddot{P}(k)}}$$

y

$$f_N(s, \delta) = \sum_{i=1}^{N} \frac{1}{N} \ln \frac{P(j_i/x_i,\theta)^{H(j_i)/\ddot{P}(j_i)}}{\sum_C P(k/x_i,\theta)^{H(k)/\ddot{P}(k)}}$$

Por el lema 5 se verifica que $\sqrt{N} \frac{d}{d\theta} f_N(s, \delta)$ converge en ley a una $N[0, \Delta]$ con $\Delta = V_g$.

Por el lema 4 $\sqrt{N}\left(\hat{\delta}_N - \delta^*\right)$ converge en ley a una variable aleatoria con distribución $N\left[0, \Omega^{-1}\Delta\,\Omega^{-1}\right]$ siendo F = $-V_g$ por la definición de la misma.

De nuevo en este caso se verifica que la matriz de varianzas-covarianzas del estimador coincide con la inversa de la matriz de información, y por lo tanto el estimador será asintóticamente eficiente.

Separando la matriz de varianzas-covarianzas asociada al estimador $\hat{\theta}$ se obtiene:

$$V_{\hat{\theta}} = \left(\Delta_{\theta\theta} - \Delta_{\theta m}\,\Delta_{mm}^{-1}\,\Delta_{m\theta}\right)^{-1}$$

donde

$$\Delta_{\alpha\beta} = -\sum_C \sum_Z \frac{\mathrm{P}(j/x,\theta^*)\,\mathrm{P}(x)}{\mathrm{P}(j)}\,H(j)\,\frac{d^2}{d\alpha\,d\beta'}\,\frac{\mathrm{P}(j/x,\theta^*)\,m(j)}{\sum_C \mathrm{P}(k/x,\theta^*)\,m(k)}$$

con α y β iguales a θ o $m(j) = \dfrac{H(j)}{\mathrm{P}(j)}$.

A.2. <u>Propiedades de los estimadores en los modelos de variable dependiente limitada</u>

Para la estimación del vector de parámetros θ en un modelo tobit, o en cualquiera de las variantes del mismo presentadas en el trabajo, Amemiya (1984) ha propuesto varios procesos de estimación.

Uno de ellos consiste en la estimación del parámetro β/σ del modelo tobit modificando la expresión de la función de verosimilitud de este modelo:

$$\mathrm{L}(\theta) = \prod_{i/y_i^* \le 0}\left[1 - \Phi(x_i'\beta/\sigma)\right]\prod_{i/y_i^*>0}\sigma^{-1}\,\phi\left((y_i - x_i'\beta)/\sigma\right) =$$

$$\prod_{i/y_i^* \le 0}\left[1 - \Phi(x_i'\beta/\sigma)\right]\prod_{i/y_i^*>0}\Phi(x_i'\beta/\sigma)\prod_{i/y_i^*>0}\Phi^{-1}(x_i'\beta/\sigma)\,\sigma^{-1}\,\phi\left((y_i - x_i'\beta)/\sigma\right)$$

Considerando esta última expresión se estima $\alpha = \beta / \sigma$ con los dos primeros factores que constituyen la función de verosimilitud de un modelo probit binomial.

El estimador $\hat{\alpha}$ es consistente, pero al ignorar una parte de la función de verosimilitud no es eficiente. La razón de la pérdida de eficiencia es evidente, ya que con este procedimiento únicamente se considera el signo de la variable dependiente y_i^* ignorando su valor numérico incluso cuando es observado.

La segunda alternativa para estimar los parámetros θ del modelo tobit o del modelo continuo con separación muestral es utilizar un proceso en dos etapas.

Amemiya (1984) demuestra la consistencia y Normalidad asintótica de estos estimadores argumentando que el propio mecanismo de obtención garantiza estas propiedades:

1. En primer lugar se realiza la estimación por máxima-verosimilitud del modelo de respuesta cualitativa asociado al modelo de variable dependiente limitada.

Por ser el estimador máximo-verosímil posee las propiedades de consistencia, Normalidad y eficiencia asintóticas. La demostración de estas propiedades tampoco puede hacerse utilizando la demostración de Rao por las mismas razones espuestas en el caso de la estimación de los modelos de respuesta cualitativa. También ahora es necesario hacer uso de los lemas comentados antes.

2. El siguiente paso es la estimación por mínimos cuadrados ordinarios del modelo de variable dependiente limitada corregida para eliminar el sesgo de selección muestral.

De nuevo se verifican las propiedades anteriores por el propio método de obtención de los estimadores, ya que los estimadores por mínimos cuadrados ordinarios son consistentes y asintóticamente Normales.

No obstante, y a título ilustrativo, en Lee, Maddala y Trost (1980) se puede encontrar una demostración detallada de estas propiedades para el modelo continuo con selección muestral en el caso de asumir la hipótesis de Normalidad conjunta para el vector aleatorio $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})'$. Estos autores utilizan los mismos argumentos que Amemiya considera en su artículo.

En la obtención de los estimadores se ha propuesto una tercera alternativa: utilizar el proceso de estimación en dos etapas como un paso previo que proporciona valores iniciales para la estimación por máxima-verosimilitud conjunta. De nuevo se puede comprobar que los estimadores obtenidos siguen cumpliendo las hipótesis deseadas para demostrar la consistencia y Normalidad Asintótica. De hecho, Amemiya (1984) y Lee y Trost (1978) demuestran la eficiencia de estos estimadores con el usual razonamiento sobre las propiedades de los estimadores máximo-verosímiles.

4. ANALISIS DE LOS DATOS

En este capítulo de la tesis se comenta la fuente que ha proporcionado los datos que se han utilizado en el estudio; así como las muestras seleccionadas (epígrafes 4.1 y 4.2 respectivamente), también se exponen las variables dependientes y explicativas que aparecen en los análisis realizados (epígrafe 4.3.) y un análisis descriptivo de los datos (epígrafe 4.4.).

4.1. Fuente de datos. Encuesta de Presupuestos Familiares 1990/91

La información empírica utilizada para este trabajo ha sido extraída de la Encuesta de Presupuestos Familiares (EPF) de 1990/91, elaborada por el Instituto Nacional de Estadística (INE). Esta encuesta tiene como objetivo suministrar información sobre los gastos de consumo de los hogares según características geográficas, económicas, familiares y sociales.

Los resultados de la EPF están estructurados en cinco bloques, referidos a diferentes áreas temáticas:

1) Registros de tipo 1. Datos del hogar.

La información relevante disponible es:

— Composición del hogar.

— Datos del sustentador principal.

— Datos de la vivienda principal.

— Instalaciones de la vivienda.

— Gastos en inversión en vivienda.

— Préstamos y amortizaciones de préstamos.

— Situación económica del hogar.

— Ingresos ordinarios monetarios y no monetarios del hogar.

2) Registros de tipo 2. Gastos de consumo del hogar.

En este registro se recogen todos los gastos que ha realizado el hogar, elevados al año.

3) Registros de tipo 3. Datos de los miembros del hogar.

La información recogida es:

— Descripción de los miembros que constituyen el hogar.

— Ingresos de los miembros del hogar.

4) Registros de tipo 4. Datos de los bienes de equipamiento.

Ofrece información general sobre todos los bienes de equipamiento que disfruta el hogar en su vivienda.

5) Registros de tipo 5. Datos de las viviendas secundarias.

En este registro la información relevante es la siguiente:

— Datos de la vivienda secundaria.

— Gastos en la vivienda secundaria.

— Instalaciones en la vivienda secundaria.

En la EPF 1990/91 se dispone de 21.155 encuestas efectuadas a hogares privados que residen en viviendas familiares principales. El ámbito temporal del estudio corresponde al período abril 1990-marzo 1991 y el ámbito geográfico de la investigación es todo el territorio español, incluidas Ceuta y Melilla.

Los hogares seleccionados para suministrar la información se han obtenido mediante un muestreo bietápico con estratificación de las unidades de primera etapa, diseñándose una muestra independiente para cada provincia. La estratificación se ha realizado según un doble criterio, el criterio geográfico y el criterio socioeconómico.

A continuación se expondrán escuetamente los conceptos básicos que tienen mayor relevancia para el trabajo que se ha realizado. Estos conceptos permitirán un mejor conocimiento de la base de datos.

Se define el "hogar" como la persona o conjunto de personas que ocupan en común una vivienda familiar o parte de ella y consumen o comparten alimentos u otros bienes con cargo al mismo presupuesto.

La "vivienda familiar" es toda habitación o conjunto de habitaciones y sus dependencias, que ocupan un edificio o una parte estructuralmente separada del mismo y que están destinadas a ser habitadas por uno o varios hogares.

Se considera "vivienda familiar principal" a toda vivienda familiar que es utilizada como residencia habitual de uno o más hogares.

Y se considera "vivienda familiar secundaria" a toda vivienda familiar no principal, disponible por el hogar durante todo un año, cuyo fin primordial es ser utilizada para esparcimiento de los miembros del hogar de forma estacional, periódica o esporádica.

Según la definición de las viviendas principal y secundaria dadas en la EPF, una vivienda familiar no puede ser principal para un hogar y secundaria para otros hogares o para ese mismo hogar.

El "sustentador principal" es aquel miembro del hogar, cuya aportación periódica al presupuesto común se destina a atender los gastos del hogar en mayor grado que las aportaciones de cada uno de los restantes miembros.

4.2. Muestras utilizadas en los diferentes análisis

Todas las muestras utilizadas en los diferentes análisis han sido obtenidas del conjunto de familias recogidas en la EPF 1990/91. De las 21.155 familias consideradas en esta encuesta únicamente se han eliminado dos por falta de información sobre ellas, ya que presentaban una renta disponible nula. Asumiendo que estos datos son erróneos o, en caso de no serlo corresponden a dos hogares no representativos, el número de hogares analizados es de 21.153.

Disponibilidad o no de vivienda secundaria

Para este análisis se ha utilizado la muestra completa de los 21.153 hogares. El análisis particularizado al ámbito rural se ha realizado con una muestra de 10.183 hogares, y en el ámbito urbano la muestra disponible es de 10.970 hogares.

Si el análisis sobre la disponibilidad o no de vivienda secundaria se particulariza al ámbito de la Comunidad Valenciana, el tamaño muestral se reduce a 1.706 hogares. En esta Comunidad para el análisis desagregado según el ámbito donde el hogar tiene fijada su residencia, las muestras son de 843 hogares para el ámbito rural y de 863 hogares para el ámbito urbano.

Análisis conjunto del régimen de tenencia de la vivienda principal y disponer o no de vivienda secundaria

Para el modelo que analiza primero el régimen de tenencia de la vivienda principal y a continuación si la familia dispone o no de vivienda secundaria, la muestra consta de 19.596 hogares. En este análisis se han eliminado de la muestra todos aquellos hogares cuya vivienda principal se disfruta en un régimen de cesión gratuita o semigratuita.

El motivo que ha llevado a eliminar de la muestra aquellos hogares cuya vivienda principal se disfruta en régimen de cesión gratuita o semigratuita, es que estos hogares no pueden considerar la elección de vivienda secundaria igual que los propietarios o inquilinos de la vivienda principal, ya que éstos están realizando una inversión (gasto) en vivienda que no hacen los hogares cuya vivienda principal es en cesión gratuita o semigratuita.

El objetivo de este modelo es analizar si el comportamiento de los hogares que disfrutan de la vivienda principal en propiedad es distinto de los que son inquilinos de la misma cuando se plantean el disponer o no de una vivienda secundaria.

También se ha realizado este análisis conjunto desagregando de acuerdo al ámbito rural o urbano en el que se ubica la vivienda principal. La muestra disponible para el ámbito rural consta de 9.324 hogares y de 10.272 hogares para el ámbito urbano.

La muestra de la Comunidad Valenciana que permitirá estudiar este modelo consta de 1.578 hogares.

Número de viviendas secundarias

La muestra disponible en el análisis del modelo sobre el número de viviendas secundarias que disfruta el hogar consta de 2.130 hogares, de los cuales 2.036 disfrutan únicamente de una vivienda secundaria y 94 hogares tienen a su disposición un número de viviendas secundarias superior o igual a dos. El número máximo de viviendas secundarias que posee un hogar en la EPF es de 4 viviendas. De los 94 hogares que tienen más de una vivienda secundaria son 3 hogares los que disfrutan del número máximo de viviendas secundarias.

El pequeño número de hogares que se encuentran en las categorías superiores (2 ó más) ha llevado a plantear el análisis en términos binarios (una ó más de una).

Para el número de viviendas secundarias que disfruta un hogar se han considerado todos los hogares que tienen a su disposición al menos una vivienda secundaria, sea cual sea su régimen de tenencia (se han incluido aquellos hogares que disfrutan la vivienda secundaria en régimen de cesión gratuita o semigratuita).

Régimen de tenencia de la vivienda secundaria

El siguiente análisis realizado es la elección del régimen de tenencia de la vivienda secundaria. En total hay 2.130 hogares que disfrutan de vivienda secundaria, pero la muestra utilizada en este análisis consta únicamente de 1.598 hogares. Se han eliminado todos aquellos que la disfrutan en régimen de cesión

gratuita o semigratuita (34 hogares) y los que han heredado la vivienda secundaria (498 hogares).

Se ha considerado para esta situación que los hogares que han heredado la vivienda secundaria disfrutan de un régimen de tenencia distinto al de propietarios; aunque en el momento actual dichos hogares disfrutan de esa vivienda en las mismas condiciones que un hogar que la ha comprado. La razón que ha llevado a ello es que a priori parece razonable pensar que estos hogares tengan características diferentes.

Tipo de vivienda secundaria

En el análisis del tipo de vivienda secundaria que disfruta el hogar (vivienda de tipo unifamiliar o vivienda de tipo colectivo, no unifamiliar) la muestra consta de 2.096 hogares, de los cuales 1.306 disponen de la vivienda secundaria de tipo unifamiliar y 790 de tipo colectivo.

De los 2.130 hogares que disfrutan de una vivienda secundaria se han eliminado los 34 hogares que disfrutan de la vivienda secundaria en un régimen de tenencia en cesión gratuita o semigratuita por considerar que ellos no han realizado en ningún momento la elección de su tipo de vivienda.

A diferencia del análisis realizado sobre el régimen de tenencia de la vivienda secundaria, ahora los hogares cuya vivienda secundaria ha sido heredada sí que se han incluido en la muestra. Cuando se pretende determinar las características que discriminan entre un tipo u otro de vivienda, el colectivo de hogares que han heredado su vivienda secundaria se considera incluido en los propietarios. La razón por la que se han mantenido los hogares cuya vivienda secundaria es heredada es porque se ha considerado que éstos están satisfechos con el tipo de vivienda que disfrutan, en otro caso se plantearían cambiar esta vivienda que han heredado por otra vivienda cuyas características, incluido el hecho de ser unifamiliar o no, se ajusten más adecuadamente a sus necesidades y a sus gustos.

Para este análisis con muestras desagregadas se tiene: en el ámbito rural una muestra de 721 hogares (421 hogares con vivienda unifamiliar y 300 hogares con una vivienda no unifamiliar); en el ámbito urbano se dispone de 1.375 hogares (885 y 490 hogares con vivienda unifamiliar y no unifamiliar respectivamente).

Para la Comunidad Valenciana se dispone de un total de 286 hogares en la muestra, de los cuales 192 disfrutan de una vivienda secundaria de tipo unifamiliar y 94 disfrutan de la vivienda secundaria de tipo colectivo.

En el ámbito rural de la Comunidad Valenciana hay un total de 115 hogares con vivienda secundaria, de los cuales 74 son unifamiliares y 41 no unifamiliares. En el ámbito urbano de esta Comunidad se dispone de una muestra de 171 hogares que disfrutan de vivienda secundaria, 120 son unifamiliares y 51 no unifamiliares.

Tamaño de la vivienda secundaria

Cuando se realiza el análisis sobre el tamaño de la vivienda secundaria los tamaños de muestra son: 2.096 hogares para la muestra total (809 hogares con una vivienda de tamaño pequeño, 822 hogares con una vivienda de tamaño medio y 465 hogares con una vivienda grande); 721 hogares para el análisis en el ámbito rural (282, 306 y 133 hogares para las viviendas pequeña, media y grande respectivamente) y 1.375 hogares en el ámbito urbano (con un tamaño pequeño 527 hogares, con un tamaño medio 516 y con un tamaño grande 332).

Particularizando a la Comunidad Valenciana, la muestra total consta igualmente de los 286 hogares anteriores, de los cuales 119 disfrutan de vivienda secundaria de tamaño pequeño, 102 de tamaño medio y 65 de tamaño grande.

Al desagregar en ámbito rural (115 hogares) y urbano (171 hogares) dentro de esta Comunidad se han obtenido en la categoría de viviendas pequeñas 53 hogares en el ámbito rural y 66 hogares en el ámbito urbano, en la categoría de tamaño medio se tienen 42 y 60 hogares en el ámbito rural y urbano respectivamente y para la categoría de viviendas grandes se han encontrado 20 hogares en zona rural y 45 hogares en zona urbana.

Aunque no se detallan aquí todas las muestras correspondientes, todos los análisis anteriores se han realizado también desagregando por la Comunidad Autónoma en la que reside habitualmente el hogar.

4.3. Variables

4.3.1. Variables explicativas

El conjunto de variables independientes que recogen las características sociodemográficas del hogar y sus factores económicos son las mismas para todos los análisis que se proponen en este trabajo.

A continuación se describen todas las variables explicativas que se han utilizado para alguno de los modelos que se han estudiado a lo largo del trabajo. Estas variables se presentan en tres grupos, el primero recoge las características del sustentador principal del hogar, el segundo las características económicas del hogar y el último grupo recoge aquellas características sociodemográficas del propio hogar.

1. Características del sustentador principal

SEXO: indica el sexo del sustentador principal. Es una variable categórica que toma el valor 1 cuando el sustentador principal es varón y el valor 0 cuando es mujer, siendo esta última la categoría de referencia.

ESTUDIOS: indica el nivel de estudios del sustentador principal. Es una variable categórica con tres valores: nivel de estudios primarios, nivel de estudios secundarios y nivel de estudios universitarios. Esta variable se ha introducido en los diversos modelos del trabajo a través de las siguientes variables ficticias.

ESTUDIO1: toma el valor 1 si los estudios que tiene el sustentador principal son primarios y el valor 0 si posee otro tipo de estudios.

ESTUDIO2: toma los valores 1 y 0 para representar un nivel de estudios secundarios y otro tipo de estudios, respectivamente.

La categoría de estudios universitarios viene identificada por una variable ficticia, ESTUDIO3, pero no se introduce en el modelo por considerarla como la categoría de referencia.

EDAD: edad (en años) del sustentador principal. En todos los modelos esta variable se ha introducido en forma cuadrática; es decir, en el modelo se

especifica la variable edad y además esta misma variable elevada al cuadrado (EDAD2).

2. Características económicas

RENTA: renta disponible del hogar.

Esta variable se obtiene como el conjunto de ingresos monetarios y no monetarios percibidos por los miembros del hogar perceptores de ingresos, cualquiera que sea su naturaleza. La renta disponible es el valor que resulta una vez deducidas las cantidades satisfechas en concepto de impuestos, cotizaciones a la Seguridad Social y otros pagos asimilados.

En definitiva, se trata de computar los ingresos en especies y los ingresos monetarios netos que el hogar tiene disponibles para hacer frente a sus gastos inmediatos, a sus necesidades futuras o para incrementar su patrimonio, es decir, con independencia del destino final que el hogar dé a estas cantidades.

La variable RENTA se introduce en todos los análisis en forma logarítmica, LRENTA, puesto que en cualquiera de los modelos propuestos se cree que un aumento en la probabilidad que se modeliza (compra, unifamiliar, etc.....) no se corresponde con un aumento proporcional de la renta. El aumento que se produciría sería en términos logarítmicos.

3. Características sociales y demográficas del hogar

MURB: ámbito donde está ubicado el domicilio habitual del hogar. Esta variable toma el valor 1 si el domicilio se encuentra en un ámbito urbano y el valor 0 si se encuentra en un ámbito rural. La categoría de referencia es la correspondiente al ámbito rural.

La EPF proporciona otras variables que reflejan características de la ubicación del hogar. Se dispone del tamaño del municipio de residencia (TMUNI). En análisis preliminares se introdujo como variable geográfica el tamaño del municipio y después de observar los resultados obtenidos se llegó a la conclusión que la variable MURB refleja mejor la característica geográfica para cualquiera de los modelos en proyecto.

MIEMHOG: número de miembros del hogar (incluyendo al sustentador principal).

TENENCIA: régimen de tenencia de la vivienda principal del hogar. Los valores de esta variable son 1 y 0 que corresponden al régimen de propiedad de la vivienda principal y al régimen de alquiler de la misma, respectivamente.

La variable TENENCIA se utiliza como variable explicativa en el modelo binomial que analiza la disponibilidad o no de una vivienda secundaria y en el modelo que analiza el régimen de tenencia de la vivienda secundaria. En ambos casos se toma como categoría de referencia la correspondiente al régimen de alquiler. Sin embargo, en el modelo que analiza conjuntamente si la vivienda principal la disfruta el hogar en régimen de propiedad o en régimen de alquiler y la disponibilidad o no de vivienda secundaria, esta variable es la variable dependiente utilizada en el primer paso del modelo.

Además de estas variables, la EPF presenta otras que recogen otro tipo de características del hogar, tal como TIPOHOG o TIPOHOGA que reflejan el tipo de hogar (1 adulto, 1 pareja, etc...) o CA que indica la Comunidad Autónoma en la que reside.

4.3.2. Variables dependientes

SECUND: es una variable que toma dos valores, el valor 1 indica que el hogar disfruta de una vivienda secundaria (independientemente del régimen en que la disfruta) y el valor 0 indica que no se dispone de vivienda secundaria.

Esta variable permite plantear los modelos binomiales que discriminan entre poseer o no una vivienda secundaria; tanto para la muestra global como para las muestras obtenidas al desagregar según el ámbito de residencia (ámbito rural y ámbito urbano) y para las diferentes Comunidades Autónomas.

TENENCIA: indica el régimen de tenencia de la vivienda principal. Únicamente se han considerado el régimen de alquiler (valor 0) y el régimen de propiedad (valor 1). En este último valor se incluyen los hogares que han heredado la vivienda.

VASECU: indica si el hogar dispone de una única vivienda secundaria o si dispone de un número mayor de viviendas secundarias. Con el valor 0 se representa la primera situación y con el valor 1 la segunda.

TENENVS: indica el régimen de tenencia de la vivienda secundaria. Los valores que toma son 0 y 1 indicando el régimen de alquiler y el régimen de

propiedad, respectivamente. En este caso, en el valor 1 no se han incluido los hogares que han heredado su vivienda secundaria, por las razones que se comentaban en el epígrafe anterior.

UNIFAM: representa el tipo de vivienda secundaria que disfruta el hogar. Los valores para dicha variable son el 0 y el 1. El primero (valor 0) representa a los hogares cuya vivienda secundaria está en un edificio colectivo, es decir, hay más de una vivienda en el edificio, y el segundo valor indica que el hogar tiene la vivienda secundaria unifamiliar (el edificio donde está ubicada la vivienda se encuentra integramente a disposición del hogar).

TAMAÑO: es una variable categórica con tres valores que representan viviendas de tamaño pequeño, medio y grande.

Para la definición de esta variable categórica se ha considerado la variable VSM2UT que recoge los metros cuadrados útiles que tiene la vivienda secundaria. Se entiende por superficie útil la que en planta queda comprendida dentro de los muros exteriores de la vivienda, más la destinada a terrazas.

Atendiendo al valor de la variable VSM2UT se ha definido una vivienda pequeña como aquella cuya superficie no supera los 75 m^2 útiles. La segunda categoría de tamaño está constituida por las viviendas con un número de m^2 útiles superior a 75, pero inferior a 110. Son las viviendas de tamaño medio. En último lugar están las viviendas grandes cuya dimensión es no inferior a 110 m^2.

Para analizar el tamaño de la vivienda secundaria se planteará, en primer lugar un modelo de *regresión lineal* con la variable VSM2UT y después un modelo de respuesta cualitativa cuya variable dependiente toma tres posibles valores. Cualquier modelo multinomial servirá para este propósito. En el presente trabajo se ha elegido un modelo *logit multinomial* para analizar la elección del tamaño de la vivienda secundaria. Con este planteamiento, un hogar elegirá una de entre las tres categorías, independientemente de las otras.

Otro planteamiento es contemplar el tamaño de la vivienda como una variable que presenta una ordenación natural: pequeña, mediana y grande.

Según cual sea el modelo elegido, las diferentes categorías del tamaño se identificarán con un valor numérico o con otro.

En el modelo *logit multinomial*, la categoría de tamaño medio se ha identificado con el valor 0, la de tamaño pequeño con el valor 1 y la de tamaño grande con el valor 2, puesto que con fines comparativos y de intepretación es

más razonable comparar las viviendas de tamaño pequeño y grande con respecto a las de tamaño medio.

Para analizar el tamaño de la vivienda como una ordenación, los valores que se asignan a cada una de las categorías son: 0 para el tamaño pequeño, 1 para el tamaño medio y 2 para el tamaño grande.

Todos los análisis que se pretenden realizar en este trabajo, salvo la regresión lineal para el tamaño, tienen en común el hecho de que la variable dependiente es de naturaleza cualitativa. Esto hace que los modelos propuestos se encuadren en el marco de los modelos de respuesta discreta.

En cada uno de los análisis se elegirá el modelo que se considere más adecuado al objetivo perseguido, o bien se compararán diversas alternativas que pudieran resultar igualmente aceptables.

Todos los modelos del trabajo se van a estimar con el programa LIMDEP. Este programa informático está desarrollado bajo la dirección del profesor W. H. Greene y permite analizar los modelos econométricos más frecuentes con datos de corte transversal, de panel y longitudinales.

Para el análisis de los datos el programa dispone de procedimientos básicos (cálculo de estadísticos descriptivos, histogramas, etc...), de modelos de regresión clásicos, así como de técnicas econométricas avanzadas (modelos logit anidados, modelos ARIMA y ARMAX, etc...). También proporciona numerosas herramientas de programación, incluye álgebra matricial y una función de optimización.

El LIMDEP es muy conocido por sus numerosas posibilidades en la estimación de modelos de respuesta cualitativa o modelos de variable dependiente limitada de donde procede su nombre (LIMited DEPendent variables).

4.4. Análisis descriptivo sobre la vivienda secundaria

Antes de proceder al análisis de los resultados obtenidos con los diferentes modelos estadísticos planteados en este trabajo sobre la demanda de vivienda secundaria se ha realizado un análisis descriptivo de los datos utilizados en el trabajo, cuyos resultados se comentan a continuación.

4.4.1. Análisis descriptivo sobre la disponibilidad o no de vivienda secundaria

El primer planteamiento es averiguar si hay muchas familias que disponen de viviendas secundarias, así en la tabla 1 están clasificados los 21.153 hogares según si disponen o no de vivienda secundaria sin considerar el régimen de tenencia de la misma, es decir, incluyendo tanto los hogares que disfrutan de la vivienda secundaria en propiedad como los que son inquilinos de la misma o la disfrutan gratuita o semigratuitamente

TABLA 1

Clasificación de los hogares según si disponen o no de vivienda secundaria

	Muestra global	Muestra rural	Muestra urbano
Dispone secund.	2130	734	1.396
No dispone secund.	19.023	9.449	9.574
Total	21.153	10.183	10.970

Con estos resultados aparece ya un dato muy significativo, únicamente el 10,07% de los hogares entrevistados en la EPF disponen de vivienda secundaria. Esto indicará muy poca predisposición hacia las viviendas secundarias por parte de los hogares españoles.

No obstante, no hay que dejar de lado que la definición de vivienda secundaria considerada en la EPF implica que no se consideren como tales aquellas viviendas que un hogar puede alquilar por períodos cortos de tiempo.

Bajo este concepto de vivienda secundaria, no tienen cabida las viviendas que un hogar alquila para pasar sus vacaciones o un período de tiempo bastante inferior a un año, ya que esta vivienda no cumple los requisitos de segunda vivienda que exige la EPF, aunque el uso que el hogar da a esta vivienda es el de una vivienda secundaria.

Teniendo en cuenta estas consideraciones no debe resultar extraño encontrar un número pequeño de hogares que disfruten de vivienda secundaria.

En la tabla 1 se encuentran también las muestras desagregadas según el ámbito de residencia del hogar: ámbito rural y ámbito urbano y se puede apreciar que en el ámbito rural únicamente el 7,21% de los hogares disfrutan de vivienda secundaria, mientras que parece que los hogares que residen habitualmente en ciudades, o núcleos urbanos, tienen una mayor predisposición hacia las viviendas secundarias, ya que el 12,73% de ellos disponen de vivienda secundaria.

Este fenómeno será comentado posteriormente con más detalle, ya que en el análisis mediante los modelos de respuesta cualitativa sobre la disponibilidad o no de vivienda secundaria, se analiza la influencia del ámbito de residencia del hogar en la decisión de disponer o no de vivienda secundaria.

La distribución muestral se ha particularizado con detalle para la Comunidad Valenciana, ya que va a ser un punto concreto del análisis de este trabajo.

En la tabla 2 se encuentra la correspondiente clasificación de los 1.706 hogares de la EPF con residencia habitual en un municipio de la Comunidad Valenciana.

TABLA 2

Clasificación de los hogares de la Comunidad Valenciana según si disponen o no de vivienda secundaria

	Muestra Comunidad Valenciana	Muestra rural Comun. Valenciana	Muestra urbano Comun. Valenciana
Dispone secund.	293	118	175
No dispone secund.	1.413	725	688
Total	1.706	843	863

El primer dato destacable es que en la Comunidad Valenciana hay un porcentaje de hogares que disponen de vivienda secundaria superior al porcentaje

nacional (el 17,18% frente al 10,07% nacional). Tal vez la situación geográfica de esta Comunidad sea un factor determinante.

En segundo lugar destaca la gran diferencia que, dentro de la Comunidad, se encuentra entre los hogares que residen en un ámbito urbano y un ámbito rural. Para los primeros, ámbito urbano, nos encontramos con que un 20,28% de hogares disfrutan de vivienda secundaria, mientras que en el ámbito rural esta proporción disminuye al 13,99%.

Este fenómeno es fácilmente explicable en la Comunidad Valenciana en la que las tres capitales de provincia son costeras y cuentan con municipios turísticos a muy pocos kilómetros. Muchas familias de las capitales disponen de un apartamento, piso o chalet en estos municipios, generalmente costeros, disfrutando de una segunda vivienda situada a muy poca distancia de la ciudad que les permite trasladarse en períodos de tiempo relativamente cortos, lo cual puede hacer a diario sin perjuicio de su trabajo.

En la decisión sobre la disponibilidad o no de vivienda secundaria, sería razonable pensar que el régimen de tenencia de la vivienda principal puede influir en esta decisión.

En las tablas 3a y 3b se encuentra la clasificación de los hogares frente a la disponibilidad de vivienda secundaria según el régimen de tenencia de la vivienda principal.

El número de hogares clasificados en estas tablas es de 19.596, que son aquellos que disfrutan de la vivienda principal en propiedad o en alquiler. Se han eliminando de la muestra los hogares que disfrutan de su vivienda principal en régimen de cesión gratuita o semigratuita.

TABLA 3a

Clasificación de los hogares propietarios de la vivienda principal según si disponen o no de vivienda secundaria

	Muestra propietarios	Muestra rural propietarios	Muestra urbano propietarios
Dispone secund.	1.778	592	1.186
No dispone secund.	14.845	7.762	7.083
Total	16.623	8.354	8.269

TABLA 3b

Clasificación de los hogares inquilinos de la vivienda principal según si disponen o no de vivienda secundaria

	Muestra inquilinos	Muestra rural inquilinos	Muestra urbano inquilinos
Dispone secund.	216	67	149
No dispone secund.	2.757	903	1.854
Total	2.973	970	2.003

Al comparar los resultados de las tablas 3a y 3b con las tablas 1 y 2 se pueden apreciar algunas diferencias. El porcentaje de hogares que disponen de vivienda secundaria se mantiene (10,17%). De estos hogares se puede matizar que son los propietarios de la vivienda principal los que más tendencia presentan a disfrutar de una segunda vivienda: el 10,7% de los hogares propietarios de la vivienda principal disponen de vivienda secundaria, es decir, el 89,2% de los hogares con vivienda secundaria son propietarios de la vivienda principal.

Por el contrario, con los inquilinos se ha obtenido un porcentaje inferior, ya que únicamente el 7,2% de ellos disponen de vivienda secundaria, es decir, el 10.8% de los hogares con vivienda secundaria son inquilinos de la vivienda principal.

También se ha desagregado cada muestra en dos muestras según el ámbito rural o urbano en el que reside habitualmente el hogar. Puede observarse que en cualquier caso el porcentaje de hogares que disfrutan de vivienda secundaria es claramente superior en las zonas urbanas. En el ámbito urbano se tiene en total que el 12,96% de los hogares disfrutan de vivienda secundaria frente al 7,07% de la zona rural.

Observando los resultados separados en propietarios e inquilinos de nuevo se aprecia un mayor porcentaje de hogares con vivienda secundaria entre los propietarios de la vivienda principal que entre los inquilinos. Destaca la gran diferencia entre estos hogares en las ciudades, ya que para los propietarios se tiene 14,34% frente al 7,24% de los inquilinos (el 89,11% de los hogares que disfrutan de vivienda secundaria son propietarios de la vivienda principal).

En zonas rurales los porcentajes, aunque inferiores para ambos tipos de hogares (propietarios e inquilinos) son prácticamente iguales: el 7,09% y el

6,91% de los propietarios y de los inquilinos respectivamente de la vivienda principal disfrutan de vivienda secundaria.

Análogamente a como se ha hecho en el análisis con toda la muestra, en la Comunidad Valenciana se han eliminado aquellos hogares que disfrutan de la vivienda principal en régimen de cesión gratuita o semigratuita y se han clasificado los 1.578 hogares restantes según el régimen de tenencia de la principal y según si disponen o no de vivienda secundaria. En la tabla 4 están los resultados.

TABLA 4

Clasificación de los hogares de la Comunidad Valenciana según si disponen o no de vivienda secundaria y según el régimen de tenencia de la vivienda principal

	Muestra propietarios Comun. Valenciana	Muestra inquilinos Comun. Valenciana
Dispone secund.	267	16
No dispone secund.	1.114	181
Total	1.381	197

Se sigue manteniendo la relación anterior: los propietarios de la vivienda principal con un 19,33% tienen mayor tendencia a disfrutar de vivienda secundaria que los inquilinos (únicamente el 8,12% de los mismos disponen de vivienda secundaria).

En este caso del total de hogares que disponen de vivienda secundaria la proporción de hogares propietarios de la vivienda principal es el 94,35%, es decir, en la Comunidad Valenciana hay muy pocos hogares que disponiendo de vivienda secundaria son inquilinos de la vivienda principal (únicamente el 5,65%).

Para finalizar con el análisis descriptivo sobre la disponibilidad o no de vivienda secundaria se presenta la clasificación de los hogares desagregando la muestra total por Comunidades Autónomas.

En la tabla 5 se pueden apreciar las grandes diferencias en la disponibilidad o no de vivienda secundaria entre las diversas Comunidades Autónomas.

TABLA 5

Clasificación de los hogares según si disponen o no de vivienda secundaria y según la Comunidad Autónoma de residencia habitual

Comunidad Autónoma	Dispone de secundaria	No dispone de secundaria	Total	Porcentaje
Andalucia	260	3.414	3.674	7,08%
Aragón	130	975	1.105	11,76%
Asturias	42	401	443	9,48%
Baleares	74	355	429	17,25%
Canarias	42	729	771	5,45%
Cantabria	30	331	361	8,31%
Castilla-La Mancha	171	1.523	1.694	10,09%
Castilla-León	265	2.897	3.162	8,38%
Cataluña	206	1.438	1.644	12,53%
Com.Valenciana	293	1.413	1.706	17,17%
Extremadura	73	757	830	8,79%
Galicia	117	1.622	1.739	6,73%
Madrid	101	663	764	13,22%
Murcia	88	438	526	16,73%
Navarra	38	329	367	10,73%
País Vasco	163	1.197	1.360	11,98%
La Rioja	30	327	357	8,40%
Ceuta-Melilla	7	214	221	3,17%
Total	2.130	19.023	21.153	10,07%

El porcentaje extremadamente pequeño de Ceuta y Melilla (3,17%) puede no tenerse en cuenta dadas las peculiaridades de estos municipios.

Dejando de lado esta Comunidad, destaca el diferente comportamiento encontrado en las dos Comunidades isleñas. Mientras Baleares presenta el mayor porcentaje, 17,25%, de hogares con vivienda secundaria a su disposición, Canarias presenta el menor porcentaje, 5,44%.

En las otras Comunidades se puede observar que la Comunidad Valenciana y Murcia son las que presentan la mayor tendencia a disponer de vivienda secundaria con un 17,17% y un 16,73% respectivamente.

Por el contrario, las Comunidades que tienen los porcentajes más bajos, son Galicia (6,73%) y Andalucia (7,08%).

Navarra y Castilla-La Mancha con un 10,35% y un 10,09% respectivamente, se encuentran alrededor de la media nacional, situada en el 10,07%.

4.4.2. Análisis descriptivo del régimen de tenencia de la vivienda secundaria

Con el fin de analizar el comportamiento de los hogares que disponen de vivienda secundaria frente a la elección del régimen de tenencia de la misma se han eliminado, de la muestra total de 2.130 hogares, aquellos que disfrutan de la vivienda secundaria en régimen de cesión gratuita o semigratuita. A su vez tampoco se han considerado los hogares que son propietarios de la vivienda secundaria por haberla heredado.

La muestra final en los análisis es de 1.598 hogares, que clasificados según el régimen de tenencia de la vivienda secundaria junto con los que la han heredado dan lugar a la tabla 6.

En esta tabla se puede apreciar la poca tendencia que tienen los hogares españoles a alquilar la vivienda secundaria. Del total de hogares que disponen de viviendas secundarias únicamente el 3,07% son inquilinos. Probablemente los motivos sean debidos al concepto de vivienda secundaria que se encuentra en la EPF y que ya ha sido comentado.

TABLA 6

Clasificación de los hogares según el régimen de tenencia de la vivienda secundaria

	Muestra global	Muestra rural	Muestra urbano
Propiedad	1.549	539	1.010
Alquiler	49	10	39
Herencia	498	172	326
Total	2.096	721	1.375

El hecho de que haya tan pocos hogares que disfrutan de la vivienda secundaria en alquiler, lleva a que los análisis de la elección del régimen de tenencia de la vivienda secundaria no puedan hacerse desagregando las muestras

según el ámbito de residencia habitual del hogar. En el ámbito urbano únicamente hay 39 hogares que son inquilinos de la vivienda secundaria y en el ámbito rural este número es de 10 hogares. En la muestra de 1.598 hogares estas cantidades representan únicamente el 3,72% y el 1,82% respectivamente.

Un dato particularmente significativo de los hogares que disfrutan de vivienda secundaria es que la mayoría disponen de la misma en régimen de propiedad y hay un elevado porcentaje de los mismos que son propietarios por haberla heredado.

Aunque no se ha realizado el análisis correspondiente, a continuación se presenta desagregando por Comunidades Autónomas la clasificación de los hogares según el régimen de tenencia de su vivienda secundaria.

TABLA 7

Clasificación de los hogares según el régimen de tenencia de la vivienda secundaria y según la Comunidad Autónoma de residencia habitual

Comunidad Autónoma	Propiedad	Alquiler	Herencia	Total
Andalucia	207	6	43	256
Aragón	82	2	45	129
Asturias	31	0	11	42
Baleares	50	5	15	70
Canarias	33	2	7	42
Cantabria	28	0	2	30
Castilla-La Mancha	111	2	56	169
Castilla-León	178	3	78	259
Cataluña	152	9	42	203
Com.Valenciana	230	3	53	286
Extremadura	46	2	24	72
Galicia	75	4	37	116
Madrid	78	1	20	101
Murcia	68	1	18	87
Navarra	22	1	15	38
País Vasco	131	5	24	160
La Rioja	22	1	6	29
Ceuta-Melilla	5	0	2	7
Total	1.549	49	498	2096

En la tabla anterior se puede apreciar de nuevo el bajo porcentaje de hogares que disponen de vivienda secundaria en alquiler. Salvo en la Comunidad de Baleares que este grupo de hogares representa el 7,14% del total de hogares que disponen de vivienda secundaria en ninguna Comunidad se pasa del 4%.

Incluso hay alguna Comunidad en la que no hay ningún hogar que disfrute de su vivienda secundaria en régimen de alquiler (Asturias, Cantabria, Ceuta y Melilla) y otras con un único hogar en esta categoría.

De hecho, para la Comunidad Valenciana se ha encontrado que únicamente el 1,05% de las viviendas secundarias son en alquiler.

En la tabla 7 se puede observar el gran número de viviendas que están catalogadas como secundarias y que son propiedad del hogar por haberlas heredado en algún momento. Exceptuando la Comunidad de Cantabria con un 6,67%, en todas las Comunidades se tienen porcentajes superiores al 15%.

4.4.3. Análisis descriptivo del tipo de vivienda secundaria: unifamiliar o no unifamiliar

Un aspecto del mercado de la vivienda que tiene mucha importancia es el tipo de vivienda, unifamiliar o no unifamiliar, que demandan los hogares.

El análisis descriptivo realizado para las diferentes muestras utilizadas en este estudio indica que en general el número de hogares que disfruta de una vivienda secundaria de tipo unifamiliar es superior al número de hogares con vivienda secundaria de tipo colectivo.

Considerando la muestra de 2.096 hogares que disponen de vivienda secundaria (eliminando los que la disfrutan en cesión gratuita o semigratuita) en la tabla 8 se presenta la distribución de los hogares según si la vivienda secundaria es de tipo unifamiliar o no unifamiliar.

A primera vista ya se detecta que en cualquier situación el número de viviendas secundarias de tipo unifamiliar es superior al número de viviendas no unifamiliares, más de la mitad de los hogares tienen una vivienda secundaria unifamiliar.

TABLA 8

Clasificación de los hogares según el tipo de vivienda secundaria

	Muestra global	Muestra rural	Muestra urbano
Unifamiliar	1.306	421	885
No unifamiliar	790	300	490
Total	2.096	721	1.375

En la muestra global el 62,31% de las viviendas son unifamiliares, en el ámbito rural la proporción es algo inferior, el 58,39% y en el ámbito urbano la proporción vuelve a aumentar al 64,36%.

Para la Comunidad Valenciana la clasificación de los hogares según el tipo de vivienda se presenta en la tabla 9. Se puede ver que en esta Comunidad el porcentaje de viviendas secundarias de tipo unifamiliar es superior a la media nacional, destacando la submuestra del ámbito urbano que presenta un 70,18%.

TABLA 9

Clasificación de los hogares de la Comunidad Valenciana según el tipo de vivienda secundaria

	Muestra Comun. Valenciana	Muestra rural Comun. Valenciana	Muestra urbano Comun. Valenciana
Unifamiliar	194	74	120
No unifamiliar	92	41	51
Total	286	115	171

La clasificación de los hogares según la Comunidad Autónoma y el tipo de vivienda se presenta en la tabla 10.

Como se podía prever, en la mayoría de las Comunidades Autónomas el número de hogares que disfrutan de una vivienda unifamiliar es superior a los que disfrutan de una vivienda no unifamiliar, con la excepción de Cantabria y el País Vasco.

Se puede apreciar que el porcentaje de viviendas secundarias de tipo unifamiliar no es el mismo para todas las Comunidades Autónomas. Mientras en

algunas no llega al 60%, como en Andalucía o La Rioja, en otras este porcentaje sobrepasa el 80% (este es el caso de la Comunidad de Extremadura).

TABLA 10

Clasificación de los hogares según el tipo de vivienda secundaria y según la Comunidad Autónoma de residencia habitual

Comunidad Autónoma	Unifamiliar	No unifamiliar	Total
Andalucia	150	106	256
Aragón	87	42	150
Asturias	26	16	42
Baleares	45	25	70
Canarias	26	16	42
Cantabria	14	16	30
Castilla-La Mancha	114	55	169
Castilla-León	169	90	259
Cataluña	119	84	203
Com.Valenciana	194	92	286
Extremadura	58	14	72
Galicia	69	47	116
Madrid	67	34	101
Murcia	52	35	87
Navarra	28	10	38
País Vasco	65	95	160
La Rioja	17	12	29
Ceuta-Melilla	6	1	7
Total	1.306	790	2.096

4.4.4. Análisis descriptivo del tamaño de la vivienda secundaria

Otra característica de la vivienda cuyo análisis presenta un gran interés es el tamaño de la misma. A continuación se presenta la distribución de los hogares según el tamaño de la vivienda secundaria que tienen a su disposición. Se han distinguido tres tamaños según los metros cuadrados útiles de las mismas: pequeño (no superior a 75 m^2), mediano (entre 75 y 110 m^2) y grande (no inferior a 110 m^2).

TABLA 11

Clasificación de los hogares según el tamaño de la vivienda secundaria

	Muestra global	Muestra rural	Muestra urbano
Pequeño	809	282	521
Medio	822	306	516
Grande	465	133	332
Total	2.096	721	1.375

Observando la tabla 11 se puede decir que la tendencia general de los hogares es a disfrutar de vivienda secundaria de tamaño pequeño o medio, hay pocas viviendas secundarias de gran tamaño. Los porcentajes de viviendas secundarias pequeñas o medias están ambos alrededor del 40% en la muestra total y en las dos submuestras para el ámbito rural y urbano.

La muestra de hogares del ámbito rural es la que presenta el menor porcentaje de viviendas secundarias clasificadas como grandes, el 18,44%, y en la muestra de hogares del ámbito urbano aparece el mayor porcentaje, el 24,14%.

A continuación, en la tabla 12, se encuentra la distribución correspondiente para los hogares de la Comunidad Valenciana.

TABLA 12

Clasificación de los hogares de la Comunidad Valenciana según el tamaño de la vivienda secundaria

	Muestra Comun. Valenciana	Muestra rural Comun. Valenciana	Muestra urbano Comun. Valenciana
Pequeña	119	53	66
Media	102	42	60
Grande	65	20	45
Total	286	115	171

De nuevo se repite la misma situación que a nivel nacional, ya que el menor porcentaje se encuentra en las viviendas secundarias de tamaño grande. El 22,73% de las viviendas secundarias de la Comunidad Valenciana son de tamaño grande, mientras que en los ámbitos rural y urbano este porcentaje es del 17,35%

y 26,31% respectivamente, lo que confirma la semejanza con la muestra global. También aquí son los hogares que habitualmente residen en ciudades los que presentan el mayor porcentaje de viviendas secundarias grandes.

Análogamente a lo que ocurría con la muestra completa, el número de hogares que disfrutan de una vivienda secundaria de tamaño pequeño es prácticamente igual, aunque algo superior, al número de hogares clasificados en la categoría que corresponde al tamaño medio.

En la tabla 13 se proporciona la distribución muestral de los hogares teniendo en consideración la Comunidad Autónoma donde está situada la vivienda principal y el tamaño de la vivienda secundaria.

TABLA 13

Clasificación de los hogares según el tamaño de la vivienda secundaria y según la Comunidad Autónoma de residencia habitual

Comunidad Autónoma	Pequeña	Media	Grande	Total
Andalucia	99	104	53	256
Aragón	52	55	22	129
Asturias	25	13	4	42
Baleares	29	27	14	70
Canarias	16	18	8	42
Cantabria	6	17	7	30
Castilla-La Mancha	46	77	46	169
Castilla-León	94	100	65	259
Cataluña	90	71	42	203
Com.Valenciana	119	102	65	286
Extremadura	25	25	22	72
Galicia	43	52	21	116
Madrid	30	41	30	101
Murcia	38	33	16	87
Navarra	8	8	22	38
País Vasco	82	59	19	160
La Rioja	6	16	7	29
Ceuta-Melilla	1	4	2	7
Total	809	822	465	2.096

Observando esta tabla de resultados no se aprecian grandes diferencias con respecto a los resultados globales, únicamente merece la pena destacar los resultados obtenidos para las Comunidades de Cantabria, Navarra, La Rioja y Ceuta y Melilla. En todas ellas, para la categoría de viviendas de tamaño pequeño es donde la muestra registra el menor número de hogares y en Navarra, al contrario que en el resto, el mayor número de hogares está en las viviendas de tamaño grande.

4.4.5. Análisis descriptivo de la correlación entre las variables explicativas

Este apartado está dedicado al análisis descriptivo de las variables explicativas que se incluyen en los diferentes modelos. Este conjunto de variables, que reflejan las características sociodemográficas y económicas del hogar, está formado por las variables MURB, SEXO, ESTUDIO1, ESTUDIO2, ESTUDIO3, MIEMHOG, RENTA y EDAD.

Para conocer estas variables en primer lugar se han calculado los estadísticos descriptivos media y varianza. En la tabla 14 están los valores para la muestra de los 21.153 hogares y en la tabla 15 los de la muestra de 1.706 hogares de la Comunidad Valenciana. En las tablas 16 y 17 se presentan los resultados correspondientes a las muestras de los hogares que disponen de vivienda secundaria en toda España y en la Comunidad Valenciana (sin incluir los que la disponen gratuita o semigratuitamente).

TABLA 14

Media y desviación típica de las variables explicativas para la muestra global

Variables	Media	Desviación típica
Sexo	0,82277	0,38187
Estudio1	0,80939	0,39279
Estudio2	0,09748	0,29662
Estudio3	0,09313	0,29062
Edad	52,86600	15,64000
Renta	2.157.200	1.625.100
Murb	0,51860	0,49967
Miemhog	3,40930	1,59680

TABLA 15

Media y desviación típica de las variables explicativas para la muestra de la Comunidad Valenciana

Variables	Media	Desviación típica
Sexo	0,82298	0,38180
Estudio1	0,84291	0,36399
Estudio2	0,08206	0,27454
Estudio3	0,07503	0,26353
Edad	51,26800	16,00600
Renta	2.060.300	1.265.900
Murb	0,50586	0,50011
Miemhog	3,29070	1,49250

TABLA 16

Media y desviación típica de las variables explicativas para la muestra global de hogares que disponen de vivienda secundaria

Variables	Media	Desviación típica
Sexo	0,88073	0,32419
Estudio1	0,69609	0,46005
Estudio2	0,12023	0,32531
Estudio3	0,18368	0,38732
Edad	53,95800	11,92400
Renta	3.292.800	2.188.900
Murb	0,65601	0,47515
Miemhog	3,72900	1,51370

TABLA 17

Media y desviación típica de las variables explicativas para la muestra de hogares de la Comunidad Valenciana que disponen de vivienda secundaria

Variables	Media	Desviación típica
Sexo	0,89161	0,31142
Estudio1	0,80769	0,39480
Estudio2	0,09441	0,29290
Estudio3	0,09790	0,29770
Edad	52,64300	12,19100
Renta	2.825.000	1.329.800
Murb	0,59790	0,49118
Miemhog	3,67830	1,32240

Observando estas tablas se puede decir que en los cuatro casos el número de hogares muestreados es mayor en las ciudades que en los pueblos, que hay un mayor número de hogares cuyo sustentador principal es hombre y que los hogares cuyo sustentador principal tiene estudios primarios son más numerosos que los hogares cuyo sustentador principal tiene otro nivel de estudios (secundarios o universitarios).

El número de miembros del hogar, tanto para toda España como para la Comunidad Valenciana, por término medio está entre 3 ó 4 miembros y la renta media se sitúa entre 2 y 3 millones, aunque en la Comunidad Valenciana los hogares muestreados tienen por término medio una renta inferior a la de los hogares de la muestra completa. En cualquier caso se aprecia que los hogares que disfrutan de una vivienda secundaria son los que tienen la renta más alta.

La edad media del sustentador principal, para todas las muestras, está entre 50 y 54 años, pero los hogares que tienen a su disposición una vivienda secundaria tienen un sustentador principal con una edad media entre 52 y 54 años.

Se ha calculado para las cuatro muestras la matriz de correlación entre estas variables. Las tablas 18 y 19 presentan la matriz de correlación para los hogares de toda la muestra y los de la Comunidad Valenciana y las tablas 20 y 21 son únicamente para aquellos hogares de toda España y de la Comunidad Valenciana que disfrutan de una vivienda secundaria.

TABLA 18

Matriz de correlación entre las variables explicativas para la muestra global

Variables	Sexo	Estudio1	Estudio2	Estudio3	Edad	Renta	Murb	Miemhog
Sexo	1							
Estudio1	-0,0383	1						
Estudio2	0,0469	-0,6772	1					
Estudio3	0,0039	-0,6604	-0,1053	1				
Edad	-0,2232	0,2534	-0,2105	-0,1277	1			
Renta	0,1439	-0,3150	0,1179	0,3054	-0,1702	1		
Murb	-0,0418	-0,2114	0,1208	0,1626	-0,0814	0,1553	1	
Miemhog	0,3429	-0,0323	0,0235	0,0197	-0,3629	0,3137	0,0147	1

TABLA 19

Matriz de correlación entre las variables explicativas para la muestra de hogares de la Comunidad Valenciana

Variables	Sexo	Estudio1	Estudio2	Estudio3	Edad	Renta	Murb	Miemhog
Sexo	1							
Estudio1	-0,0483	1						
Estudio2	0,0491	-0,6926	1					
Estudio3	0,0155	-0,6597	-0,0851	1				
Edad	-0,1798	0,2435	-0,2000	-0,1279	1			
Renta	0,1635	-0,3335	0,1552	0,2990	-0,2057	1		
Murb	-0,0591	-0,1367	0,0648	0,1213	-0,0558	0,1143	1	
Miemhog	0,3188	-0,0260	0,0147	0,0205	-0,3793	0,3958	0,0134	1

TABLA 20

Matriz de correlación entre las variables explicativas para la muestra global de hogares que disponen de vivienda secundaria

Variables	Sexo	Estudio1	Estudio2	Estudio3	Edad	Renta	Murb	Miemhog
Sexo	1							
Estudio1	-0,0159	1						
Estudio2	0,0002	-0,5595	1					
Estudio3	0,0187	-0,7179	-0,1754	1				
Edad	-0,0901	0,1874	-0,1701	-0,0798	1			
Renta	0,0572	-0,3569	0,1003	0,3396	-0,1326	1		
Murb	-0,0278	-0,1989	0,1040	0,1489	-0,0154	0,1611	1	
Miemhog	0,2852	-0,0518	0,0148	0,0491	-0,3505	0,2477	0,0017	1

TABLA 21

Matriz de correlación entre las variables explicativas para la muestra de hogares de la Comunidad Valenciana que disponen de vivienda secundaria

Variables	Sexo	Estudio1	Estudio2	Estudio3	Edad	Renta	Murb	Miemhog
Sexo	1							
Estudio1	0,0296	1						
Estudio2	-0,0028	-0,6617	1					
Estudio3	-0,0365	-0,6751	-0,1064	1				
Edad	-0,0619	0,1884	-0,1595	-0,0928	1			
Renta	-0,0974	-0,3762	0,2433	0,2595	-0,1665	1		
Murb	-0,0565	-0,1830	0,0453	0,1982	-0,0686	0,2514	1	
Miemhog	0,2984	0,0289	-0,0209	-0,0178	-0,4113	0,2676	0,0000	1

La correlación entre las variables, como puede verse en las tablas anteriores, únicamente supera el valor 0,5 (en términos absolutos) para los pares de variables ESTUDIO1 con ESTUDIO2 y ESTUDIO1 con ESTUDIO3.

Aunque exista esta alta correlación entre las variables ESTUDIO no se eliminan del conjunto de variables explicativas de los siguientes modelos, puesto que se pretende discriminar sobre todas las características del sustentador principal, entre ellas, el nivel de ESTUDIOS que éste posee. Esta es la razón que lleva a que todas las variables planteadas inicialmente como variables explicativas se tengan en cuenta en los modelos siguientes.

Si se comparan las tablas 18 y 19, no se observa ningún cambio en el sentido de la correlación. Sin embargo, para los hogares que disponen de vivienda secundaria en toda España o en la Comunidad Valenciana (tablas 20 y 21) sí que existen cambios. Las variables asociadas al nivel de estudios del sustentador principal (ESTUDIO1, ESTUDIO2, ESTUDIO3) tienen una correlación no nula con las variables SEXO y MIEMHOG con signos diferentes en las dos muestras.

Aunque a priori no exista una explicación determinante de este fenómeno, es razonable pensar que el hecho de seleccionar los hogares que sí disponen de vivienda secundaria puede introducir cambios en las relaciones entre las variables. En los análisis correspondientes se buscará la relación concreta que liga cada variable con el hecho de disponer o no de vivienda secundaria.

5. ANÁLISIS DE RESULTADOS

5.1. Análisis sobre la disponibilidad de vivienda secundaria

En el primer análisis de este trabajo se estudia si el hogar dispone o no de vivienda secundaria, y se plantea un modelo *logit binomial* para determinar aquellas variables que discriminan entre los dos grupos de hogares.

Antes de comentar el modelo, se hace notar que la elección del modelo *logit binomial* frente a otro modelo de elección discreta binomial no es significativa, en el sentido que las conclusiones obtenidas con uno u otro serán equivalentes. De hecho el modelo *logit binomial* y el modelo *probit binomial* proporcionan las mismas conclusiones debido a que las distribuciones de probabilidad asociadas a ambos modelos son casi iguales, únicamente tienen pequeñas variaciones en las colas (apartado 3.1.4).

En análisis previos se ha analizado la disponibilidad o no de una segunda vivienda con un modelo *probit* y con un modelo *logit* y evidentemente los resultados no experimentan ningún cambio. En consecuencia se ha optado por un único modelo, el *logit* en este caso.

El conjunto de variables explicativas introducidas en el análisis son: MURB, SEXO, ESTUDIO1, ESTUDIO2, MIEMHOG, LRENTA, EDAD y EDAD2. En un primer modelo se incluyó como variable explicativa el régimen de tenencia de la vivienda principal (propietarios o inquilinos) y en los resultados del análisis se obtuvo que no influía en el modelo. Por ello, el modelo que finalmente se va a comentar no incluye a dicha variable.

El modelo se ha estimado con el programa LIMDEP. Se ha modelizado la probabilidad del valor 1 de la variable SECUND (disponer de vivienda secundaria) y por lo tanto la modalidad de no disponer de vivienda secundaria lleva asignado el valor cero a todos sus coeficientes. La interpretación del modelo se basa siempre en la probabilidad del valor 1 de la variable dependiente.

5.1.1. Análisis de la muestra global

Los resultados obtenidos en la estimación de este modelo con la muestra completa de los 21.153 hogares se encuentran en la tabla 22. Puede observarse que todos los coeficientes estimados son altamente significativos, a excepción del asociado a la variable que recoge la categoría de estudios secundarios del sustentador principal. Este resultado nos indicaría que entre los hogares cuyo sustentador principal tiene estudios secundarios y aquellos cuyo sustentador principal tiene estudios universitarios no existe una diferencia determinante en la elección entre disponer o no de segunda vivienda. Sin embargo, en el modelo se mantienen todas las categorías de estudios (primarios, secundarios) puesto que se considera interesante la comparación entre todas las categorías para ver el efecto global de los estudios del sustentador principal.

TABLA 22

Análisis sobre la disponibilidad de vivienda secundaria con la muestra global

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-30,54800	0,8496	-35,955	0,00000
Sexo	0,23167	0,0754	3,072	0,00213
Estudio1	-0,12916	0,0747	-1,729	0,08378
Estudio2	-0,05333	0,0939	-0,568	0,57022
Edad	0,20668	0,0148	13,935	0,00000
Edad2	-0,00173	0,0001	-12,575	0,00000
Lrenta	1,55900	0,0542	28,733	0,00000
Murb	0,30251	0,0518	5,831	0,00000
Miemhog	-0,12176	0,0187	-6,489	0,00000
Nº observaciones	21.153			
Log-verosimilitud	-5.981,349			
Log-veros. restrin.	-6.908,935			
Chi-cuadrado (8)	1.855,171			
Nivel signific.	0,000000			
% predic. correc.	89,94%			

Con respecto a la variable SEXO, el signo positivo de su coeficiente permite decir que los hogares cuyo sustentador principal es un hombre son los que presentan mayor probabilidad de disponer de vivienda secundaria.

Los coeficientes estimados para las variables que recogen los estudios del sustentador principal (ESTUDIOS) son negativos y crecientes con respecto al nivel de estudios. Se podría concluir que cuanto mayor es el nivel de estudios del sustentador principal, mayor es la probabilidad hacia la disposición de vivienda secundaria. Así, el grupo de universitarios sería el primero mientras que el grupo de hogares cuyo sustentador principal tiene un nivel de estudios primarios son los que menos probabilidad tienen de disponer de segunda vivienda.

Los coeficientes estimados para la EDAD dan una forma cuadrática negativa, pudiendo concluir que la probabilidad de disponer de segunda vivienda aumenta con la edad. Al principio el aumento es más rapido, pero llega un momento en el que el crecimiento es más suave hasta que cambia la tendencia y a partir de una edad determinada decrece la probabilidad.

Para la variable LRENTA, se ha obtenido un coeficiente positivo. Este resultado indica que, como se esperaba, el disponer de segunda vivienda va relacionado directamente con la renta. Son los hogares con mayor renta disponible los que disfrutan de vivienda secundaria.

Disponer de una segunda vivienda siempre implica tener unos gastos añadidos a los habituales del hogar y es muy razonable encontrar que son los hogares de mayores rentas son los que acceden a disponer de una segunda vivienda.

Para la variable MURB se ha obtenido un coeficiente positivo que indica que los hogares cuya vivienda principal está en una ciudad tienen mayor probabilidad de disponer de una segunda vivienda que aquellos hogares que viven en ámbitos no urbanos.

Esta conclusión parecía bastante lógica a priori, puesto que los hogares que viven en una ciudad tienen más preferencia por salir de la misma cuando llegan los fines de semana o las vacaciones. En determinadas fechas del año las ciudades se encuentran prácticamente desiertas. En las zonas rurales estos fenómenos no son tan abundantes.

Por último, al observar el número de miembros del hogar, MIEMHOG, se comprueba que el signo del coeficiente estimado es negativo lo cual indicará que al aumentar el número de miembros del hogar disminuye la probabilidad de tener vivienda secundaria.

Desde el valor del estadístico de la razón de verosimilitudes que aparece en la tabla 22 se concluye que en este caso debe aceptarse el modelo propuesto como correcto frente al modelo constante para cualquier nivel de significación.

Este resultado era de prever, puesto que todas las variables incluidas en el modelo resultaban ser altamente significativas.

La información proporcionada por el número de predicciones correctas no resulta contradictoria con las conclusiones obtenidas mediante el test de la razón de verosimilitudes. Como consecuencia, el modelo propuesto resulta muy razonable para explicar la situación de disponer o no de segunda vivienda a través de las características del hogar.

5.1.2. Análisis de las muestras desagregadas según el ámbito rural y urbano

El modelo anterior sobre la disponibilidad o no de una vivienda secundaria, se ha estimado con las muestras correspondientes al ámbito rural y al ámbito urbano. Los valores de los coeficientes estimados y del nivel crítico se encuentran en la tabla 23.

Comparando estos valores con los obtenidos para la muestra global (tabla 22), se puede observar que en la significatividad de las variables hay pocos cambios para las dos submuestras. Para la submuestra del ámbito rural la variable SEXO pasa a ser no significativa; mientras que la variable ESTUDIO2 es ahora significativa. En la submuestra del ámbito urbano únicamente hay un cambio para los ESTUDIOS del sustentador principal, la variable ficticia que representa el nivel de estudios primarios (ESTUDIO1) pasa a ser no significativa.

Por lo tanto, en el ámbito rural no influye el sexo del sustentador principal para disfrutar o no de una vivienda secundaria; y que en un ámbito urbano no se pueden establecer diferencias de comportamiento del hogar debidas al nivel de estudios del sustentador principal.

Dado que los signos de los coeficientes significativos se mantienen iguales que los de la tabla 22 se puede decir que la influencia de las variables explicativas en la disponibilidad o no de una vivienda secundaria prácticamente no varía si se considera la muestra global o las muestras del ámbito rural y urbano.

El contraste sobre la validez del modelo, también en ambos casos lleva a la aceptación del modelo propuesto para cualquier nivel de significación.

Los resultados sobre las predicciones correctas confirman que para las dos muestras el modelo propuesto debe ser aceptado como correcto.

TABLA 23

Análisis sobre la disponibilidad de vivienda secundaria con las muestras del ámbito rural y del ámbito urbano

Variable dependiente: SECUND

	RURAL			URBANO		
Variable	Coef	Est t	Nivel	Coef	Est. t	Nivel
Constante	-27,94500	-21,176	0,00000	-31,91800	-28,419	0,00000
Sexo	0,15563	1,200	0,23024	0,26868	2,895	0,00379
Estudio1	-0,45623	-3,080	0,00207	-0,01741	-0,200	0,84183
Estudio2	-0,33659	-1,690	0,09110	0,02407	0,225	0,82201
Edad	0,17764	7,672	0,00000	0,22417	11,607	0,00000
Edad2	-0,00151	-7,039	0,00000	-0,00187	-10,395	0,00000
Lrenta	1,46910	17,367	0,00000	1,62700	22,844	0,00000
Miemhog	-0,13345	-4,369	0,00001	-0,11466	-4,789	0,00000
N° observaciones	10.183			10.970		
Log-verosimilitud	-2.366,209			-3.606,864		
Log-veros. restrin.	-2.637,282			-4.181,076		
Chi-cuadrad (7)	542,1457			1.148,424		
Nivel signific.	0,0000000			0,0000000		
% predic. corre.	92,68%			87,24%		

5.1.3. Análisis de la muestra global de la Comunidad Valenciana

En este apartado se pretende analizar el mismo modelo anterior pero limitando el ámbito geográfico del estudio a la Comunidad Valenciana. Ello permitirá llevar a cabo un análisis comparativo de gran interés entre los hogares del territorio nacional y los de la Comunidad Valenciana.

Los resultados del modelo *logit binomial* sobre disponer o no de una vivienda secundaria se encuentran en la tabla 24. Comparándolos con los resultados a nivel nacional (tabla 22) existen ciertas diferencias en los coeficientes estimados y los niveles críticos.

Los niveles críticos correspondientes a los coeficientes de las variables ESTUDIOS del sustentador principal y número de miembros del hogar, MIEMHOG, indican que estas variables no son significativas. En la Comunidad Valenciana es indistinto el nivel de estudios que tiene el sustentador principal para disponer o no de una segunda vivienda, y tampoco influye en la elección el

número de miembros que componen el hogar. Como en el apartado anterior tampoco ahora se eliminará ninguna de las variables explicativas del modelo.

TABLA 24

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de la Comunidad Valenciana

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-31,01600	2,5300	-12,261	0,00000
Sexo	0,39001	0,2198	1,774	0,07604
Estudio1	0,28955	0,2611	1,109	0,26738
Estudio2	0,18319	0,3253	0,563	0,57334
Edad	0,19569	0,0401	4,886	0,00000
Edad2	-0,00164	0,0004	-4,285	0,00002
Lrenta	1,62160	0,1679	9,660	0,00000
Murb	0,31716	0,1418	2,237	0,02530
Miemhog	-0,07767	0,0571	-1,361	0,17350
N° observaciones	1.706			
Log-verosimilitud	-673,7154			
Log-veros. restrin.	-782,4487			
Chi-cuadrado (8)	217,4665			
Nivel signific.	0,000000			
% predic. correc.	83,18%			

También para la variable SEXO las conclusiones obtenidas con los análisis para los dos ámbitos geográficos (nacional y Comunidad Valenciana) coinciden. Cuando el sustentador principal del hogar es un hombre aumenta la probabilidad de disfrutar de una vivienda secundaria.

La variable EDAD tiene una forma parabólica negativa como en el mismo modelo llevado a cabo para todo el territorio español.

El logaritmo de la RENTA disponible por el hogar tiene un coeficiente estimado con signo positivo, como en el análisis de toda España, lo cual indica que los hogares con niveles de renta más altos tienen una probabilidad más alta de disponer de una vivienda secundaria.

Finalmente, el coeficiente de la variable MURB indica que en las zonas urbanas se tiene mayor tendencia a poseer una vivienda secundaria que en las

zonas rurales. Este resultado coincide con el obtenido para toda España (tabla 22), que como ya se ha comentado, resulta bastante lógico.

Respecto a la bondad del ajuste, si contrastamos este modelo frente a un modelo constante, el valor del estadístico de la razón de verosimilitudes permite comprobar que para cualquier nivel de significación debe aceptarse el modelo propuesto.

El número de predicciones correctas es bastante elevado, por lo que puede considerarse coherente con la conclusión del test de la razón de verosimilitudes.

5.1.4. Análisis de las muestras desagregadas según el ámbito rural y urbano en la Comunidad Valenciana

Los resultados de la estimación del modelo anterior para la muestra del ámbito rural y la muestra del ámbito urbano en la Comunidad Valenciana, se encuentran en la tabla 25. En la significatividad de las variables explicativas incluidas en el modelo hay pocos cambios con respecto a la muestra completa de la Comunidad Valenciana. Con las muestras del ámbito rural y urbano el coeficiente de la variable SEXO tampoco es una variable significativa. En esta Comunidad se mantienen las preferencias por disponer de una vivienda secundaria tanto si el sustentador principal es un hombre como una mujer independientemente del ámbito de residencia habitual del hogar.

Los signos de los coeficientes estimados coinciden en los tres análisis realizados sobre la Comunidad Valenciana (tablas 24 y 25) para todas las variables explicativas que son significativas.

Después de comentar los resultados para la Comunidad Valenciana, se puede apreciar que en todos los análisis existen diferencias con respecto a la muestra de toda España en la variable ESTUDIOS, lo cual lleva a pensar que la influencia de las categorías de los estudios del sustentador principal en la disponibilidad de vivienda secundaria varía según la ubicación geográfica.

Para la bondad del ajuste, si se contrasta el modelo propuesto frente a un modelo constante, en ambas muestras se obtiene la aceptación del modelo propuesto para cualquier nivel de significación.

TABLA 25

Análisis sobre la disponibilidad de vivienda secundaria con las muestras del ámbito rural y urbano de la Comunidad Valenciana

Variable dependiente: SECUND

	RURAL			URBANO		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-24,74000	-6,807	0,00000	-37,29000	-10,149	0,00000
Sexo	0,50617	1,383	0,16880	0,31393	1,116	0,26459
Estudio1	1,04090	1,571	0,11612	0,23384	0,765	0,44419
Estudio2	1,08550	1,468	0,14208	0,02324	0,060	0,95229
Edad	0,16737	2,970	0,00298	0,20896	3,644	0,00027
Edad2	-0,00141	-2,647	0,00813	-0,00173	-3,115	0,00184
Lrenta	1,17770	4,923	0,00000	2,05410	8,426	0,00000
Miemhog	-0,03481	-0,398	0,69037	-0,09608	-1,254	0,20983
N° observaciones	843			863		
Log-verosimilitud	-310,9568			-356,9689		
Log-veros. restrin.	-341,3479			-435,1536		
Chi-cuadrado (7)	60,78222			156,3695		
Nivel signific.	0,0000001			0,0000000		
% predic. correc.	85,76%			82,27%		

5.1.5. Análisis de las muestras desagregadas según la Comunidad Autónoma

Como ya se ha comentado anteriormente, un análisis comparativo entre toda España y cualquiera de las Comunidades Autónomas puede resultar muy ilustrativo, tanto en aspectos económicos como en aspectos sociales (distribución de la riqueza, etc...). Este apartado se dedica íntegramente a comentar las diferencias en los resultados del modelo disponer o no de una vivienda secundaria para las Comunidades Autonómas en que se encuentra dividido el territorio español, salvo para la Comunidad Valenciana que ya se ha analizado por separado en apartados anteriores (5.1.3 y 5.1.4).

La gran desagregación que supone dividir por Comunidades Autónomas puede conducir a que algunas muestras obtenidas sean excesivamente pequeñas para realizar los correspondientes análisis. Para evitar este problema se han realizado agrupaciones de Comunidades teniendo en cuenta su proximidad geográfica y rasgos comunes (sociales e históricos). Las Comunidades agrupadas han sido:

- Andalucia con Extremadura

- Asturias con Cantabria

- Castilla-La Mancha con Castilla-León

- Navarra con el País Vasco y La Rioja

No se han considerado ni Canarias ni Ceuta y Melilla debido a los pocos hogares que disponen de vivienda secundaria en las muestras correspondientes.

El modelo utilizado, como en los análisis anteriores, es el modelo *logit binomial* y se mantiene el mismo conjunto de variables explicativas. Los coeficientes estimados y los niveles críticos obtenidos para las muestras de las diferentes Comunidades Autónomas, o agrupaciones de Comunidades, se presentan en las tablas 26a-26j que se recogen en el apéndice B al final del capítulo 5.

Observando el nivel crítico de los coeficientes estimados se comprueba que la única variable altamente significativa en todas las Comunidades Autónomas es la RENTA, así en todos los casos la decisión de disponer o no de segunda vivienda está muy influenciada por la renta disponible del hogar. El valor del coeficiente estimado en todos los casos tiene signo positivo, lo que indica que en cualquier Comunidad son los hogares con un mayor nivel de renta los que tiene una mayor probabilidad de disfrutar de vivienda secundaria.

La variable EDAD es significativa en todas las Comunidades, excepto en la Comunidad de Murcia, donde el coeficiente estimado tiene un nivel de significación del 11% para el término lineal y del 19,8% para el término cuadrático.

Dicha variable presenta una forma cuadrática negativa en todas las Comunidades analizadas, la tendencia a disfrutar de una vivienda secundaria aumenta cuando aumenta la edad del sustentador principal, en los primeros tramos de la edad y la relación se invierte en los últimos tramos, es decir si el sustentador principal es de edad madura su tendencia hacia las viviendas secundarias disminuye. Parece lógico el resultado que las familias de mediana edad sean las que tienen mayores posibilidades de poseer una vivienda secundaria.

La significatividad de la variable MIEMHOG, como puede observarse en las tablas correspondientes, depende de la Comunidad analizada. El número de

miembros del hogar tiene un coeficiente estimado de signo negativo en todos los casos. Los hogares formados por un elevado número de miembros tienen una probabilidad menor de poseer vivienda secundaria.

La variable SEXO del sustentador principal sólo es significativa para Baleares. El coeficiente de esta variable es positivo en todas las Comunidades, a excepción de Castilla-La Mancha/Castilla-León. En general, si el sustentador principal es varón aumenta la tendencia a disfrutar de segunda vivienda.

Si se observan los niveles críticos, como ocurría con el sexo, la significatividad de la variable MURB varía según la Comunidad.

La variable MURB tiene un coeficiente negativo en Baleares y en Asturias y Cantabria, en el resto de las Comunidades el signo es positivo. En la mayoría de las Comunidades se puede concluir que en las ciudades se tiene mayor tendencia a disfrutar de segunda vivienda.

En cuanto a la significatividad de las variables ESTUDIOS hay que matizar que en ninguna Comunidad (o agrupación) hay diferencias entre los hogares cuyo sustentador principal tenga estudios secundarios y los hogares de la categoría de estudios universitarios. Sólamente se pueden establecer distinciones en el comportamiento de los hogares de la categoría de estudios primarios con respecto a los universitarios, y no en todas las Comunidades consideradas.

Los ESTUDIOS del sustentador principal es la variable que presenta más variaciones en el signo del coeficiente entre las Comunidades. Se puede ver que mientras en Castilla-La Mancha/Castilla-León la mayor probabilidad la tienen los hogares cuyo sustentador principal posee estudios universitarios, en la Comunidad de Madrid son estos hogares los que tienen menor probabilidad.

Si se compara el signo de los coeficientes del modelo estimado para todas las Comunidades con los obtenidos para toda España (tabla 22) se puede establecer que, en términos generales, no hay prácticamente ninguna variación.

La contrastación de la validez del modelo con todo el conjunto de variables explicativas frente a un modelo nulo, en todas las Comunidades Autónomas analizadas lleva a la aceptación del modelo completo como el correcto para cualquier nivel de significación.

5.2. Análisis conjunto de la elección del régimen de tenencia de la vivienda principal y la elección entre disponer o no de una vivienda secundaria

En este epígrafe se va a proceder a realizar el estudio de la decisión de ser propietario o inquilino de la vivienda principal y disponer o no de vivienda secundaria conjuntamente.

El modelo que se plantea ahora pretende modelizar la elección entre las siguientes cuatro alternativas de elección:

1.propietario de la vivienda principal y disponer de vivienda secundaria

2.propietario de la vivienda principal y no disponer de vivienda secundaria

3.inquilino de la vivienda principal y disponer de vivienda secundaria

4.inquilino de la vivienda principal y no disponer de vivienda secundaria

Para modelizar esta situación de elección será necesario utilizar un modelo multinomial que tenga en cuenta la estructura de relaciones entre las alternativas de elección. No será adecuado un modelo *logit multinomial* ni un modelo *probit independiente*, ya que por la propia construcción del conjunto de elección no puede aceptarse la independencia entre las alternativas.

Una posibilidad es considerar que el hogar decide en primer lugar cuál va a ser el régimen de tenencia de la vivienda principal, y una vez realizada esta elección se plantea la posibilidad de disponer de una vivienda secundaria.

Se supone que el comportamiento del hogar frente a la segunda elección está condicionado por la alternativa elegida en el primer paso. En este caso puede utilizarse un modelo de eliminación jerárquica que considera la elección de una alternativa como el resultado de ir pasando por subconjuntos de alternativas hasta llegar a la alternativa elegida.

Así, la probabilidad de elegir una alternativa cualquiera se calculará como el producto de todas las probabilidades intermedias (o probabilidades de transición). Por ejemplo la probabilidad de elegir la alternativa 1 será el producto de la probabilidad de que el hogar decida comprar su vivienda principal frente a

alquilarla y la probabilidad de disponer de vivienda secundaria condicionada a la situación de ser propietario de la vivienda principal:

$$P(alternativa1) = P(comprar)\, P(si\ disponer\ /\ propietario)$$

Para realizar este análisis se considerará que las probabilidades de transición, o probabilidades intermedias, son de la forma *logit*. Puesto que en cada paso del proceso de eliminación se plantean decisiones binarias, se tendrán modelos *logit binomiales*.

La estimación del modelo conjunto se realizará en dos etapas. En primer lugar se estimará el modelo binomial para la elección del régimen de tenencia de la vivienda principal y en un segundo paso se estimará el modelo binomial para la elección entre disponer o no de vivienda secundaria, considerando las dos opciones posibles (ser propietario y ser inquilino de la vivienda principal). La verosimilitud del modelo global se calculará como el producto de las verosimilitudes de todos los modelos intermedios estimados.

En este modelo, tanto en el primer paso cuando se discrimina entre el régimen de tenencia de la vivienda principal (propietarios/inquilinos) como en el segundo paso donde se intenta determinar si el hogar disfruta o no de vivienda secundaria, las variables explicativas utilizadas son las mismas que en el análisis realizado en el epígrafe 5.1 anterior.

En el análisis de este modelo jerárquico (el modelo de elección discreta usado en los dos pasos es el modelo *logit binomial*), el programa informático LIMDEP modeliza la probabilidad de que cada una de las variables dependientes tome el valor 1.

La variable dependiente del primer modelo es la variable TENENCIA que toma el valor 0 si el hogar tiene la vivienda principal en alquiler y el valor 1 si la disfruta en propiedad. Esta última categoría recoge tanto a los hogares que compraron la vivienda principal como aquellos que la han heredado. Para el segundo modelo, la variable dependiente es SECUND que toma los valores 1 y 0 para representar la disponibilidad o no de vivienda secundaria, respectivamente.

5.2.1. Análisis de la muestra global

Los resultados de la estimación de este modelo con la muestra global pueden verse en las tablas 27a y 27b. En la tabla 27a están los valores de los coeficientes y los niveles críticos del modelo que discrimina entre propietarios e inquilinos de la vivienda principal y en la tabla siguiente, tabla 27b, están los valores correspondientes al modelo que discrimina, dentro de cada uno de los régimenes de tenencia de la vivienda principal, entre disfrutar o no de una vivienda secundaria

En el modelo *logit binomial* estimado para analizar la primera etapa del proceso de eliminación (régimen de tenencia de la vivienda principal), se ha obtenido que todas las variables incluidas resultan ser altamente significativas.

Respecto a los coeficientes estimados se puede decir que la variable SEXO tiene un coeficiente positivo, así cuando el sustentador principal es hombre los hogares prefieren disponer de la vivienda principal en propiedad.

TABLA 27a

Análisis del régimen de tenencia de la vivienda principal con la muestra global

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-19,72100	0,65640	-30,044	0,00000
Sexo	0,22742	0,05707	3,985	0,00007
Estudio1	0,99638	0,07358	13,541	0,00000
Estudio2	0,45834	0,08644	5,303	0,00000
Edad	0,12833	0,00854	15,022	0,00000
Edad2	-0,00094	0,00008	-11,579	0,00000
Lrenta	1,21160	0,04440	27,285	0,00000
Murb	-0,85975	0,04580	-18,772	0,00000
Miemhog	-0,08444	0,01657	-5,097	0,00000
N° observaciones	19.596			
Log-verosimilitud	-7.386,904			
Log-veros. restrin.	-8.341,465			
Chi-cuadrado (8)	1.909,121			
Nivel signific.	0,000000			
% predic. correc.	85,55%			

TABLA 27b

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-30,79800	-31,552	0,00000	-30,14600	-12,941	0,00000
Sexo	0,19438	2,338	0,01941	0,25169	1,157	0,24704
Estudio1	-0,20493	-2,464	0,01372	0,07515	0,348	0,72797
Estudio2	-0,13954	-1,321	0,18657	0,41865	1,708	0,08761
Edad	0,21535	12,419	0,00000	0,19157	4,849	0,00000
Edad2	-0,00180	-11,306	0,00000	-0,00164	-4,258	0,00002
Lrenta	1,55770	25,458	0,00000	1,58630	10,435	0,00000
Murb	0,40625	7,069	0,00000	-0,20608	-1,242	0,21424
Miemhog	-0,11856	-5,685	0,00000	-0,17170	-3,021	0,00252
N° observaciones	16.623			2.973		
Log-verosimilitud	-4.901,701			-663,1295		
Log-veros. restrin.	-5.653,688			-774,3186		
Chi-cuadrado (8)	1.503,975			222,3781		
Nivel signific.	0,0000000			0,0000000		
% predic. correc.	89,23%			92,73%		

Los coeficientes estimados para las variables ESTUDIO1 y ESTUDIO2 son positivos y crecientes con respecto al nivel de estudios del sustentador principal. Este resultado permite afirmar que los hogares con mayor probabilidad de disfrutar de la vivienda principal en propiedad son los de la categoría de estudios primarios.

A medida que aumenta el nivel de estudios la tendencia con respecto al régimen de tenencia de la vivienda principal va cambiando. Al pasar de la categoría de estudios primarios a la de secundarios se sigue manteniendo, aunque en menor grado, la propensión a comprar la vivienda principal, y los hogares cuyo sustentador principal posee estudios universitarios son los que menos probabilidad presentan hacia la compra de su vivienda principal.

La variable EDAD del sustentador principal se comporta de forma parabólica y negativa; es decir, en los primeros tramos, un aumento de la edad lleva a un aumento en la probabilidad de ser el propietario de la vivienda principal; mientras que en los tramos finales de la edad, si ésta aumenta, disminuye la probabilidad de ser propietario.

El logaritmo de la RENTA tiene un coeficiente estimado positivo, lo que indicará que al aumentar la renta disponible del hogar aumentan las probabilidades de que la vivienda principal sea propiedad del hogar.

Este resultado es el que se esperaba, ya que ser propietario de una vivienda siempre implica haber realizado una gran inversión en la compra de la misma y en consecuencia se necesitará disponer de una cierta cantidad de dinero que en el caso de alquilar no hace falta.

La variable MURB tiene signo negativo, lo que indicaría que los hogares ubicados en un ámbito rural tienen una probabilidad mayor a tener la vivienda principal en propiedad que los hogares del ámbito urbano.

Una posible razón podría ser que en una ciudad hay mayor oferta de viviendas en alquiler que en una zona no urbana y otro motivo razonable es que el precio de compra de las viviendas en una ciudad es posiblemente más elevado que en una zona rural.

Para el número de miembros del hogar, MIEMHOG, se concluye que los hogares con muchos miembros son los que menos probabilidad tienen a vivir en propiedad, puesto que el coeficiente asociado con esa variable es de signo negativo.

En la segunda etapa del proceso de eliminación, el hogar debe decidir entre disponer o no de una vivienda secundaria. Como ya se ha comentado el modelo binomial planteado es el *logit* y se ha realizado la estimación del mismo para los subgrupos de hogares que quedan con la primera decisión.

De nuevo se utilizan las mismas variables explicativas que en el modelo anterior. Observando la correspondiente tabla de resultados se puede apreciar que para el subgrupo de los propietarios prácticamente todas las variables explicativas consideradas resultan significativas. Para el subgrupo de los inquilinos, las variables MURB y SEXO no son significativas, tienen niveles críticos de 0,21424 y 0,24708, respectivamente. La variable ESTUDIO1 tampoco es significativa, aunque se ha mantenido en el análisis para discriminar con los tres niveles de estudio del sustentador principal.

Se han estimado modelos alternativos para el subgrupo de los inquilinos eliminando las variables MURB y SEXO y los resultados obtenidos no modifican las conclusiones proporcionadas por el modelo anterior. Por ello, se mantiene el modelo con el conjunto de variables explicativas completo.

Para la variable SEXO se ha obtenido un coeficiente estimado con signo positivo en ambos subgrupos, en consecuencia son los hogares cuyo sustentador principal es varón los que tienen mayores preferencias por una vivienda secundaria, tanto si el hogar tiene la vivienda principal en propiedad como si la tiene en alquiler.

En los niveles de estudio (ESTUDIO) los coeficientes estimados son negativos y crecientes para los propietarios de la vivienda principal y para los inquilinos los coeficientes asociados son positivos y crecientes. Por lo tanto, en el subgrupo de los propietarios de la vivienda principal la probabilidad mayor de disfrutar de una vivienda secundaria la tienen los hogares cuyo sustentador principal tiene un nivel de estudios universitarios y en el subgrupo de los inquilinos dicha probabilidad la tienen los hogares cuyo sustentador principal posee un nivel de estudios medios.

La EDAD del sustentador principal actúa de forma parabólica y negativa tanto en los propietarios como en los inquilinos de la vivienda principal.

Para la RENTA disponible del hogar, el coeficiente estimado es positivo, en ambos casos, lo cual nos lleva a concluir que cuanto mayor es el nivel de renta del hogar mayores son sus posibilidades de poseer una vivienda secundaria.

El coeficiente estimado para la variable MURB tiene signo diferente para cada subgrupo, en el subgrupo de los propietarios el signo es positivo, mientras que en el subgrupo de los inquilinos es negativo. Ello indicará que en las ciudades, cuando se posee la vivienda principal en propiedad, la probabilidad de disfrutar de una segunda vivienda es mayor que en los pueblos. Sin embargo, cuando la vivienda principal es disfrutada por el hogar en régimen de alquiler la probabilidad de poseer una vivienda secundaria es menor en las zonas urbanas que en las rurales.

El número de miembros del hogar, MIEMHOG, influye negativamente en la disponibilidad o no de vivienda secundaria. Tanto para los propietarios como para los inquilinos de la vivienda principal, si aumenta el número de miembros del hogar disminuye la probabilidad de disfrutar de una segunda vivienda.

Como conclusión se podría decir que al mirar si un hogar dispone o no de vivienda secundaria, los hogares que son propietarios de la vivienda principal no tienen las mismas características que los hogares que disfrutan de su vivienda principal en régimen de alquiler.

Sin embargo, las conclusiones obtenidas para los propietarios de la vivienda principal coinciden con las obtenidas para la muestra completa de hogares (tabla 22). Tal vez la razón sea el hecho de que el número de hogares inquilinos de su vivienda principal es muy pequeño comparado con la muestra de propietarios, de forma que al considerar la muestra completa principalmente se reflejan las características de los hogares propietarios de la vivienda principal, quedando encubiertas las de los hogares inquilinos.

En lo que respecta a la bondad de ajuste se puede ver que los tres modelos estimados por separado deben ser aceptados como correctos frente a un modelo nulo, para cualquier nivel de significación.

Al considerar el modelo conjunto se tiene que de nuevo se acepta como correcto para cualquier nivel de significación.

5.2.2. Análisis de las muestras desagregadas según el ámbito rural y urbano

Igual que en el modelo analizado en el epígrafe 5.1, en este apartado se realizará la estimación del modelo de eliminación jerárquica considerando la muestra de los hogares españoles que proporciona la EPF desagregada según el ámbito donde se encuentra la residencia habitual del hogar. Los resultados de la estimación de este modelo se encuentran en las tablas 28 y 29.

En las tablas 28a y 29a están los coeficientes estimados y los niveles críticos del modelo de elección del régimen de tenencia de la vivienda principal para el ámbito rural y urbano, respectivamente y en la tabla 28b se encuentran los resultados obtenidos del análisis del modelo sobre la disposición o no de vivienda secundaria para el subgrupo de los propietarios y los inquilinos en el ámbito rural y en la tabla 29b están los correspondientes resultados obtenidos con la muestra del ámbito urbano.

Al estimar el primer modelo de este análisis, en donde se discrimina a los hogares por el régimen de tenencia de su vivienda principal, el signo de los coeficientes obtenidos para ambas muestras (rural y urbano) coinciden con los de la muestra global (apartado 5.2.1.). El nivel de significatividad de los mismos es también casi igual que en el modelo global para todas las variables explicativas, la única diferencia se encuentra en la significatividad de la variable MIEMHOG, que en el ámbito rural es no significativa.

TABLA 28a

Análisis del régimen de tenencia de la vivienda principal con la muestra del ámbito rural

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-15,93600	1,0860	-14,674	0,00000
Sexo	0,23897	0,1009	2,369	0,01784
Estudio1	0,95740	0,1528	6,266	0,00000
Estudio2	0,70748	0,1881	3,761	0,00017
Edad	0,14229	0,0138	10,315	0,00000
Edad2	-0,00094	0,0001	-7,050	0,00000
Lrenta	0,87749	0,0724	12,115	0,00000
Miemhog	-0,01747	0,0286	-0,611	0,54113
N° observaciones	9.324			
Log-verosimilitud	-2.816,985			
Log-veros. restrin.	-3.112,857			
Chi-cuadrado (7)	591,7444			
Nivel signific.	0,00000			
% predic. correc.	89,68%			

TABLA 28b

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal del ámbito rural

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-29,38900	-19,208	0,00000	-26,93300	-6,709	0,00000
Sexo	0,18184	1,236	0,21662	0,10137	0,255	0,79866
Estudio1	-0,51335	-2,932	0,00337	-0,41111	-1,043	0,29690
Estudio2	-0,31428	-1,350	0,17707	-0,41336	-0,753	0,45134
Edad	0,21336	7,443	0,00000	0,14789	2,352	0,01868
Edad2	-0,00179	-6,889	0,00000	-0,00134	-2,144	0,03205
Lrenta	1,48620	15,686	0,00000	1,53970	5,768	0,00000
Miemhog	-0,11904	-3,500	0,00047	-0,34758	-3,163	0,00156
N° observaciones	8.354			970		
Log-verosimilitud	-1.914,647			-212,2503		
Log-veros. restrin.	-2.137,528			-243,6953		
Chi-cuadrado (7)	445,762			62,89006		
Nivel signific.	0,0000000			0,1E-06		
% predic. correc.	92,79%			92,89%		

TABLA 29a

Análisis del régimen de tenencia de la vivienda principal con la muestra del ámbito urbano

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-23,21600	0,8484	-27,366	0,00000
Sexo	0,20709	0,0704	2,943	0,00325
Estudio1	1,14440	0,0853	13,411	0,00000
Estudio2	0,43366	0,0974	4,451	0,00001
Edad	0,11570	0,0111	10,415	0,00000
Edad2	-0,00091	0,0001	-8,569	0,00000
Lrenta	1,43270	0,0574	24,952	0,00000
Miemhog	-0,11136	0,0206	-5,410	0,00000
N° observaciones	10.272			
Log-verosimilitud	-4.504,978			
Log-veros. restrin.	-5.086,069			
Chi-cuadrado (7)	1.126,182			
Nivel signific.	0,00000			
% predic. correc.	91,99%			

TABLA 29b

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal del ámbito urbano

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-31,34100	-24,384	0,00000	-32,56700	-11,132	0,00000
Sexo	0,19889	1,968	0,04907	0,30984	1,179	0,23847
Estudio1	-0,10831	-1,132	0,25783	0,29330	1,129	0,25873
Estudio2	-0,10090	-0,848	0,39621	0,69048	2,446	0,01446
Edad	0,21692	9,896	0,00000	0,22416	4,346	0,00001
Edad2	-0,00181	-8,928	0,00000	-0,00189	-3,805	0,00014
Lrenta	1,61280	20,012	0,00000	1,63720	8,702	0,00000
Miemhog	-0,11777	-4,441	0,00001	-0,09452	-1,437	0,15079
N° observaciones	8269			2003		
Log-verosimilitud	-2.984,486			-442,5618		
Log-veros. restrin.	-3.399,688			-530,4851		
Chi-cuadrado (7)	830,4039			175,8466		
Nivel signific.	0,0000000			0,0000000		
% predic. correc.	85,58%			92,61%		

Para el modelo binomial que discrimina entre los hogares que disfrutan de vivienda secundaria y los hogares que no disfrutan de ella, en el ámbito rural las variables explicativas que no son significativas en el subgrupo de los propietarios son SEXO y ESTUDIO2. Y en el subgrupo de los inquilinos SEXO y ESTUDIOS.

En el ámbito urbano no son significativas las variables ESTUDIOS del sustentador principal para los propietarios y SEXO y ESTUDIO1 para los inquilinos.

Comparando la significatividad de las variables explicativas para estas muestras con la muestra global si que se observan algunas diferencias.

Se han llevado a cabo en todos los casos las estimaciones de modelos más reducidos, eliminando del conjunto total de variables explicativas las no significativas. En los resultados obtenidos no hay ningún cambio en los signos de los coeficientes estimados o en la significatividad de los mismos con respecto a los obtenidos para el modelo completo. Puesto que no se mejora el modelo total y no hay cambios significativos, se ha mantenido el modelo completo.

En cuanto al signo de los coeficientes estimados en el ámbito urbano, se ha obtenido que no hay ningún cambio respecto de la muestra global. En el ámbito rural, se observa un cambio en el signo de los coeficientes estimados del nivel de ESTUDIOS del sustentador principal para el subgrupo de los inquilinos, ya que ahora éstos son negativos y decrecientes. En este caso se tiene que los hogares cuyo sustentador principal posee estudios universitarios son los que prefieren disfrutar de una vivienda secundaria cuando la vivienda principal la tienen en régimen de alquiler.

Con respecto a la bondad de ajuste de nuevo se deberán aceptar los modelos propuestos para las decisiones intermedias como válidos, tanto con la muestra del ámbito rural como con la muestra del ámbito urbano.

Lo mismo ocurre con el modelo conjunto para la modelización de la situación de elección planteada en este epígrafe: el modelo propuesto se acepta como válido frente a un modelo nulo para cualquier nivel de significación, tanto en el ámbito rural como urbano.

5.2.3. Análisis de la muestra global de la Comunidad Valenciana

En las tablas 30a y 30b se encuentran los resultados de la estimación de los modelos correspondientes con la muestra de hogares de la Comunidad Valenciana.

El modelo de eliminación propuesto en este epígrafe no se estimará para las muestras desagregadas según el ámbito de residencia del hogar. La razón es el pequeño número de hogares que disponen de vivienda secundaria al considerar el régimen de tenencia de la vivienda principal dentro del ámbito rural y urbano por separado.

Para el primero de los análisis casi todas las variables explicativas resultan significativas. La variable ESTUDIO2 tiene un nivel crítico del 17%, aunque se mantiene en el modelo para realizar comparaciones entre las tres categorias de la variable ESTUDIOS del sustentador principal.

TABLA 30a

Análisis del régimen de tenencia de la vivienda principal con la muestra de la Comunidad Valenciana

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-16,95400	2,4270	-6,985	0,00000
Sexo	0,53122	0,2016	2,635	0,00841
Estudio1	0,78608	0,2951	2,664	0,00772
Estudio2	0,53097	0,3874	1,370	0,17054
Edad	0,15170	0,0296	5,119	0,00000
Edad2	-0,00135	0,0003	-4,772	0,00000
Lrenta	1,04530	0,1676	6,237	0,00000
Murb	-0,50992	0,1650	-3,091	0,00200
Miemhog	-0,21648	0,0633	-3,419	0,00063
N° observaciones	1.578			
Log-verosimilitud	-540,3796			
Log-veros. restrin.	-594,0567			
Chi-cuadrado (8)	107,3541			
Nivel signific.	0,000000			
% predic. correc.	87,52%			

TABLA 30b

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de la Comunidad Valenciana

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-29,38800	-10,861	0,00000	-38,59400	-3,811	0,00014
Sexo	0,23783	1,028	0,30407	1,83610	1,632	0,10276
Estudio1	0,11859	0,432	0,66603	0,80771	0,781	0,43478
Estudio2	0,13869	0,401	0,68825	0,39680	0,321	0,74830
Edad	0,17198	3,966	0,00007	0,29243	1,793	0,07303
Edad2	-0,00140	-3,393	0,00069	-0,00274	-1,733	0,08314
Lrenta	1,55810	8,721	0,00000	1,95380	2,863	0,00420
Murb	0,36754	2,446	0,01444	0,03214	0,050	0,95981
Miemhog	-0,02935	-0,484	0,62821	-0,39608	-1,636	0,10182
Nº observaciones	1.381			197		
Log-verosimilitud	-593,2964			-42,82935		
Log-veros. restrin.	-678,1087			-55,50175		
Chi-cuadrado (8)	169,6247			25,34481		
Nivel signific.	0,0000000			0,001358		
% predic. correc.	81,32%			91,88%		

En cuanto al signo de los coeficientes, la variable SEXO también tiene asociado un signo positivo a su coeficiente estimado, por lo tanto si el sustentador principal es un hombre la probabilidad de ser propietario de la vivienda principal es mayor que cuando es mujer.

Los coeficientes de la variable ESTUDIO son positivos y decrecientes indicando con ello que, los hogares cuyo sustentador principal tiene estudios primarios tiene mayor probabilidad de ser propietario de la vivienda principal, y esta probabilidad decrece al aumentar el nivel de estudios. Observando la tabla 27a se puede ver que los valores de los coeficientes coinciden en el signo, aunque ahora la magnitud de los mismos es menor en las dos categorías de los estudios.

Para la variable EDAD se ha obtenido una relación parabólica negativa, como en el análisis global.

El coeficiente estimado del logaritmo de la RENTA es positivo, lo que indica que cuando aumenta la renta del hogar aumentan las preferencias por disponer de la vivienda principal en propiedad.

La variable MURB tiene signo negativo, como en la muestra completa (tabla 27a), lo que indica que si una vivienda está ubicada en un ámbito urbano disminuye la tendencia a disponer de la vivienda principal en propiedad.

La estimación del coeficiente asociado a la variable que representa el número de miembros del hogar, MIEMHOG, tiene signo negativo y tampoco presenta variación, en consecuencia si aumenta el número de miembros del hogar disminuye la propensión a ser propietario de la vivienda.

Para el modelo binomial que discrimina entre la posesión o no de una vivienda secundaria dentro de cada uno de los subgrupos obtenidos en el primer análisis, la significatividad de las variables explicativas ha disminuido con respecto al mismo análisis con la muestra global.

Ahora las únicas variables significativas en ambos grupos analizados son el nivel de renta del hogar (LRENTA) y la edad del sustentador principal (EDAD y EDAD2). Para la muestra de propietarios además es significativo el ámbito de residencia habitual del hogar (MURB).

Comparando los signos de los coeficientes estimados ahora con los obtenidos para la muestra global del territorio nacional se pueden observar algunas diferencias.

Análogamente a los modelos anteriores la variable EDAD se comporta de una forma parabólica y negativa, es decir, para los primeros tramos de la edad un aumento de la misma indicaría un aumento en la probabilidad de disponer de una vivienda secundaria y para los últimos tramos de la edad, un aumento indicaría una disminución de esta probabilidad.

El logaritmo de la RENTA tiene un signo positivo, así las familias con mayor nivel de renta tienen mayor probabilidad de poseer vivienda secundaria, independientemente del régimen de tenencia de su vivienda principal.

La variable MURB tiene un coeficiente estimado positivo, lo que indicaría que las posibilidades de poseer una vivienda secundaria es mayor en los núcleos urbanos tanto si los hogares son propietarios como inquilinos de su vivienda principal.

El test de la razón de verosimilitudes que proporciona el programa informático LIMDEP, lleva a considerar el modelo conjunto propuesto como correcto frente a un modelo constante, para cualquier nivel de significación.

5.2.4. Análisis de las muestras desagregadas según la Comunidad Autónoma

A continuación se presentan todos los resultados de la estimación del modelo de eliminación jerárquica para las diferentes muestras separadas por Comunidades Autónomas.

De nuevo se han eliminado las Comunidades de Canarias y Ceuta y Melilla por el pequeño número de hogares, recogidos en las correspondientes muestras, que disponen de vivienda secundaria.

Para cada Comunidad Autónoma, o agrupación de Comunidades, se presenta en primer lugar el resultado de la estimación de un modelo *logit binomial* para el régimen de tenencia de la vivienda principal y en segundo lugar aparecen los resultados de la estimación del modelo binomial para la elección entre disponer o no de vivienda secundaria, separando la muestra según el régimen de tenencia de la vivienda principal.

Todos estos resultados se presentan en las tablas $31a_1$ -$31j_1$, para el régimen de tenencia de la vivienda principal, y en las tablas $31a_2$ -$31j_2$, para la disponibilidad o no de vivienda secundaria, que están recogidas en el apéndice B de este capítulo.

Comparando los resultados con los de las tablas 27a y 27b que recogen la estimación del mismo modelo para la muestra global de los 19596 hogares propietarios o inquilinos de su vivienda principal, se puede observar que hay algunas diferencias sobre todo en el segundo paso del modelo.

En primer lugar destaca el hecho que en el modelo *logit binomial* estimado para el régimen de tenencia de la vivienda principal, se han obtenido prácticamente los mismos resultados en cuanto al signo de los coeficientes y a su significatividad se refiere, para todas las Comunidades.

Se concluiría que el comportamiento de los hogares ante la decisión de comprar o alquilar su vivienda principal es el mismo en todas las Comunidades Autónomas analizadas, y por supuesto coincide con el comportamiento a nivel nacional. También ahora se acepta el modelo propuesto como correcto para cualquier nivel de significación en cualquier Comunidad.

Al analizar la disponibilidad de vivienda secundaria se puede concluir que el comportamiento de los hogares propietarios de su vivienda principal es prácticamente el mismo en todas las Comunidades y se mantienen los resultados obtenidos a nivel nacional.

Sólo hay unos cambios de signos y destaca la disminución de la significatividad de algunos de ellos.

En Navarra/País Vasco/La Rioja cambian de signo, con respecto a la muestra global, los coeficientes asociados a las variables que recogen los ESTUDIOS del sustentador principal. En Aragón sólo cambia el referente a ESTUDIO1 y en Andalucia/Extremadura el de ESTUDIO2. En Baleares cambian ESTUDIO2 y MURB, el coeficiente de SEXO cambia en Cataluña y Castilla-La Mancha/Castilla-León y en Asturias/Cantabria cambia MURB.

A pesar de cambiar la significatividad de muchos de estos coeficientes, al observar el valor del estadístico de la razón de verosimilitudes se concluye que se debe aceptar el modelo propuesto como correcto para todas las Comunidades.

Sin embargo, al considerar los resultados de la estimación del modelo con las muestras de inquilinos de la vivienda principal, hay muchos cambios con respecto a los resultados de la tabla 27b que se obtuvieron con la muestra completa.

En este caso nos encontramos con grandes cambios tanto en el signo de los coeficientes como en la magnitud de los mismos, y para varias Comunidades se ha encontrado que ninguna variable es significativa en este modelo.

Al realizar la estimación se ha comprobado que el número de hogares inquilinos de la vivienda principal que disponen de vivienda secundaria es muy pequeño comparado con los correspondientes hogares que son propietarios de la vivienda principal. Tal vez estas muestras insuficientes sea la razón por la que se han obtenido resultados tan diferentes y poco significativos.

De hecho al analizar la bondad del ajuste, hay que rechazar el modelo propuesto para las muestras de inquilinos de las Comunidades de Murcia, Asturias/Cantabria y Navarra/País Vasco/La Rioja.

No obstante, el modelo conjunto es aceptado como válido en todas las Comunidades analizadas.

5.3. Análisis del número de viviendas secundarias que posee el hogar

En los apartados anteriores se han analizado las características que determinan si un hogar dispone o no de vivienda secundaria. Tras la decisión de disponer de segunda vivienda un hogar podría plantearse la posibilidad de disponer de más de una vivienda secundaria.

Con la muestra de los 2.130 hogares que sí disponen de vivienda secundaria se van a analizar qué características de los mismos son las que influyen en el hecho de disfrutar de una única vivienda secundaria o de más de una, independientemente del régimen de tenencia de las mismas.

Se plantea un modelo *logit binomial* para modelizar la variable dicotómica VASECU que indicará si el hogar dispone de una vivienda secundaria con el valor 0 y tomará el valor 1 para aquellos hogares que disponen de más de una vivienda secundaria.

Ya se ha comentado en el epígrafe 4.2 dedicado a las muestras de los análisis, la dificultad encontrada para analizar el número de viviendas secundarias al disponer de muy pocos hogares con más de una.

De nuevo se utilizan las mismas variables independientes y los resultados de la estimación del modelo están en la tabla 32.

El primer resultado destacable es la no significatividad de las variables, ya que el único coeficiente estimado significativamente diferente de cero es el asociado a la variable LRENTA.

Se puede concluir que el nivel de renta del hogar es el factor determinante en la elección entre disponer de una vivienda secundaria o de más de una. Con esta variable se consigue explicar la decisión anterior, y no hay ninguna otra característica del sustentador principal ni del hogar o su entorno que influya en esta decisión.

La variable LRENTA posee un coeficiente estimado con signo positivo que permite concluir que son los hogares con mayor nivel de renta los que han optado por disfrutar de más de una vivienda secundaria, por el contrario si disminuye el nivel de renta aumenta la probabilidad de disponer de una única vivienda secundaria.

El signo de este coeficiente era esperado. El nivel de renta está relacionado directamente con el número de viviendas secundarias.

Las viviendas secundarias suponen un gasto adicional para el hogar y mantener más de una vivienda de este tipo necesitará un determinado nivel económico. Así, es razonable que la conclusión del análisis sea que la probabilidad de disponer de más de una vivienda secundaria aumenta con el nivel de renta.

TABLA 32

Análisis sobre la disponibilidad de una vivienda secundaria o más de una con la muestra global

Variable dependiente: VASECU

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-30,66200	4,1060	-7,468	0,00000
Sexo	0,35553	0,4134	0,860	0,38984
Estudio1	0,31461	0,2857	1,101	0,27076
Estudio2	0,04645	0,3903	0,119	0,90527
Edad	0,08390	0,0848	0,989	0,32255
Edad2	-0,00053	0,0008	-0,699	0,48433
Lrenta	1,58590	0,2323	6,826	0,00000
Murb	-0,05617	0,2443	-0,230	0,81818
Miemhog	0,07807	0,0698	1,119	0,26328
N° observaciones	2.130			
Log-verosimilitud	-351,2189			
Log-veros. restrin.	-385,2294			
Chi-cuadrado (8)	68,02091			
Nivel signific.	0,1E-06			
% predic. correc.	95,63%			

Observando el valor del estadístico de la razón de verosimilitudes que proporciona el programa LIMDEP se llega a la conclusión que hay que aceptar el modelo estimado frente a un modelo constante para cualquier nivel de significación.

5.4. Análisis del régimen de tenencia de la vivienda secundaria

Una vez analizado el modelo que discrimina a los hogares según si disponen o no de segunda vivienda, el siguiente paso es encontrar qué características presentan aquellos hogares que, disponiendo de vivienda secundaria, disfrutan de ella en régimen de propiedad o en régimen de alquiler.

Para este análisis, del total de hogares que disponen de vivienda secundaria se han eliminado aquellos que disfrutan de la misma en régimen de cesión gratuita o semigratuita, así como aquellos que han heredado la vivienda secundaria.

De nuevo se plantea un modelo *logit binomial* para analizar el régimen de tenencia de la vivienda secundaria y se modeliza la probabilidad de ser propietario de la misma, ya que esta categoría lleva asociado el valor 1 de la variable dependiente. Entre las variables explicativas se incluye para este modelo la variable ficticia que refleja el régimen de tenencia de la vivienda principal del hogar (TENENCIA). Se ha considerado que el régimen de tenencia de la vivienda principal es una variable interesante cuando se pretende analizar el régimen de tenencia de la vivienda secundaria.

Las otras variables explicativas incluidas en el modelo que recogen las características del sustentador principal, las características económicas y las características del hogar, coinciden con las variables de los modelos comentados anteriormente.

Nótese que el hecho de utilizar la variable TENENCIA, que recoge el régimen de tenencia de la vivienda principal, como variable explicativa en el modelo planteado, restringe la muestra de hogares que disponen de vivienda secundaria de 1.598 a 1.355, de los cuales 45 son inquilinos de la vivienda secundaria y 1.310 propietarios de la misma, ya que se han eliminado los hogares que disfrutan de la vivienda principal gratuita o semigratuitamente.

Los resultados de la estimación (coeficientes y niveles críticos) se encuentran en la tabla 33. Observando esa tabla puede verse el escaso número de variables significativas, únicamente las variables MURB, LRENTA y TENENCIA tienen un nivel crítico inferior al 10%.

TABLA 33

Análisis del régimen de tenencia de la vivienda secundaria con la muestra global

Variable dependiente: TENENVS

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-7,49600	5,8150	-1,289	0,19734
Sexo	0,40410	0,4534	0,891	0,37283
Estudio1	0,62843	0,4247	1,480	0,13892
Estudio2	-0,03436	0,4754	-0,072	0,94238
Edad	-0,00349	0,0983	-0,036	0,97161
Edad2	0,00002	0,0009	0,021	0,98299
Lrenta	0,71403	0,3740	1,909	0,05625
Murb	-0,74498	0,4088	-1,822	0,06841
Miemhog	-0,12096	0,1225	-0,987	0,32360
Tenencia	0,78460	0,3723	2,113	0,03463
N° observaciones	1.355			
Log-verosimilitud	-189,6829			
Log-veros. restrin.	-197,4646			
Chi-cuadrado (9)	15,563447			
Nivel signific.	0,07665764			
% predic. correc.	96,68%			

A la vista de los resultados se ha realizado la estimación de nuevos modelos eliminando aquellas variables que no son significativas y los signos correspondientes a los coeficientes estimados no han variado. Puesto que las conclusiones para las variables significativas no varían, mantenemos el primer modelo que permite hacer comparaciones entre los diversos análisis realizados.

Observando el signo de los coeficientes estimados se puede concluir que los hogares cuya residencia habitual está ubicada en un ámbito urbano la tendencia que presentan es disfrutar de la vivienda secundaria en régimen de alquiler, ya que el coeficiente estimado para la variable MURB es negativo.

El coeficiente asociado a la variable LRENTA es positivo. Esto indicará que los hogares de mayor renta son los que poseen la vivienda secundaria en propiedad, y los de menor renta disfrutan de ella en régimen de alquiler.

Este resultado era previsible: comprar una segunda vivienda implica, en general, realizar una inversión monetaria mucho más elevada que ser inquilino de la misma. En consecuencia los hogares que deciden comprar una vivienda secundaria deberán tener más posibilidades económicas que aquellos que eligen la alternativa de alquilar dicha vivienda.

Para la variable TENENCIA (régimen de propiedad o de alquiler de la vivienda principal) el signo del coeficiente estimado es positivo, lo que indicaría que los hogares que disfrutan de una vivienda principal en propiedad tienen mayor probabilidad de disponer de una vivienda secundaria en propiedad que aquellos hogares que son inquilinos de la vivienda principal.

Este resultado parece bastante lógico, los hogares que poseen la vivienda principal de su propiedad es más frecuente que se planteen adquirir una segunda vivienda. Cuando no se dispone de ninguna vivienda en propiedad resulta más coherente plantearse adquirir antes la vivienda principal que la secundaria.

El valor del estadístico de la razón de verosimilitudes lleva a aceptar el modelo propuesto como correcto para niveles de significación mayores que 0,0766.

Se ha realizado el análisis del modelo con la muestra completa de los 1598 hogares sin incluir la variable TENENCIA y no se modifica ningún resultado, salvo el que hace referencia a la bondad de ajuste del modelo, puesto que en este caso ha resultado no ser aceptable.

El modelo del análisis del régimen de tenencia de la vivienda secundaria sólo se ha realizado para la muestra global de todos los hogares que disfrutan de una vivienda secundaria. Para las muestras desagregadas según el ámbito donde se ubica la residencia del hogar (rural y urbano) y para la muestra de la Comunidad Valenciana, el número de hogares cuya vivienda secundaria está en régimen de alquiler es muy pequeño (10, 39 y 3 hogares, respectivamente) y no se considera suficientemente representativo para realizar los pertinentes análisis. Por la misma razón tampoco se ha realizado este análisis desagregando por Comunidades Autónomas.

Con fines comparativos se ha estimado la elección del régimen de tenencia de la vivienda secundaria considerando ahora tres modalidades de tenencia para la misma: herencia, propiedad y alquiler. El objetivo es encontrar características diferenciadoras entre los hogares que son propietarios de la vivienda secundaria porque han decidido comprarla en algún momento de su vida y aquellos hogares que también son propietarios de una vivienda secundaria pero por haberla heredado. No se tiene en cuenta el régimen de tenencia de la vivienda principal.

En este caso se estima un modelo *logit multinomial* con tres posibles valores para la variable respuesta: el valor 0 para inquilinos, el valor 1 para propietarios y el valor 2 para los que la han heredado.

De nuevo el programa LIMDEP considera el valor 0 a todos los coeficientes asociados a la categoría de inquilinos que corresponde al valor 0 de la variable dependiente. Así, los coeficientes estimados deben ser interpretados por comparación con estos.

Los coeficientes de la categoría de propietarios indicarán la relación entre comprar y alquilar y los coeficientes asociados a la categoría de herencia se referirán a la comparación entre alquilar y heredar.

La muestra consta de los 2.096 hogares anteriormente comentados y en la estimación del modelo correspondiente se han encontrado los mismos resultados en los coeficientes que discriminan entre inquilinos y propietarios que en el modelo anterior (tanto en el signo como en la magnitud y significatividad) y para la categoría de los hogares que han heredado la vivienda se han encontrado coeficientes no significativos para todas las variables.

Ante la no significatividad de las características del hogar para la alternativa que representa ser propietario de una segunda vivienda por haberla heredado, el modelo estimado ahora no se ha considerado en el trabajo, ya que es comparable al analizado en primer lugar y que únicamente consideraba las alternativas de comprar y alquilar.

5.5. Análisis del tipo de vivienda secundaria: unifamiliar/no unifamiliar

5.5.1. Análisis de la muestra global

En el análisis del mercado de las viviendas secundarias un aspecto que puede resultar interesante es determinar las características de los hogares que disponen de una segunda vivienda unifamiliar, frente a los que disfrutan de una vivienda secundaria no unifamiliar.

Para determinar qué características diferenciadoras presentan los hogares de ambos grupos, se ha estimado un modelo *logit binomial*, en el que la variable dependiente, UNIFAM, toma dos valores, unifamiliar (valor 1) o no unifamiliar (valor 0) y como variables independientes se han utilizado las mismas características sociales y demográficas y factores económicos del hogar de los modelos de los epígrafes anteriores (Murb, Sexo, Estudio1, Estudio2, Miemhog, Lrenta, Edad y Edad2).

Análogamente a los análisis anteriores, el programa LIMDEP realiza la estimación para los coeficientes asociados al valor 1 de la variable dependiente, en este caso al tipo unifamiliar.

Los resultados de la estimación se encuentran en la tabla 34. Como puede observarse, casi todas las variables son altamente significativas. Únicamente las correspondientes al número de miembros del hogar (MIEMHOG) y a la EDAD del sustentador principal son no significativas.

El coeficiente de la variable SEXO es positivo, lo cual lleva a afirmar que cuando el sustentador principal es hombre el hogar tiene mayor probabilidad de disfrutar de una vivienda secundaria unifamiliar que si es mujer.

Los coeficientes estimados para las variables que recogen el nivel de estudios del sustentador principal (ESTUDIO1 y ESTUDIO2) son positivos y decrecientes con el nivel de estudios.

El signo y la magnitud de estos coeficientes permite decir que los hogares cuyo sustentador principal tiene como mucho estudios primarios son los que más valoran las viviendas secundarias de tipo unifamiliar. Por el contrario, los hogares

cuyo sustentador principal tiene estudios universitarios son los que menos probabilidad tienen de disponer de este tipo de viviendas secundarias.

TABLA 34

Análisis del tipo de vivienda secundaria con la muestra global

Variable dependiente: UNIFAM

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	6,23960	1,6770	3,721	0,00020
Sexo	0,32959	0,1459	2,259	0,02388
Estudio1	0,36583	0,1318	2,776	0,00551
Estudio2	0,29665	0,1692	1,753	0,07960
Edad	0,01375	0,0297	0,464	0,64300
Edad2	-0,00003	0,0003	-0,122	0,90277
Lrenta	-0,49409	0,1064	-4,642	0,00000
Murb	0,43571	0,0998	4,365	0,00001
Miemhog	0,03145	0,0359	0,875	0,38140
Nº observaciones	2.096			
Log-verosimilitud	-1.352,441			
Log-veros. restrin.	-1.388,664			
Chi-cuadrado (8)	72,44573			
Nivel signific.	0,1E-06			
% predic. correc.	63,03%			

Para la variable que recoge la renta disponible por el hogar (LRENTA) se ha obtenido sorprendentemente un coeficiente estimado negativo. Esto indica que los hogares que disponen de una vivienda secundaria de tipo unifamiliar son los de menor nivel de renta. A medida que aumenta el nivel de renta disminuye la probabilidad hacia las viviendas secundarias unifamiliares.

Una posible explicación de este resultado es que las viviendas secundarias unifamiliares, que están ocupadas por los hogares con menor nivel de renta, serán seguramente viviendas antiguas o de menor calidad, quizás sean viviendas que el hogar ha heredado y por tanto no ha necesitado realizar ninguna inversión en ellas. Tal vez si estos hogares no hubieran heredado esta vivienda no tendrían a su disposición ninguna vivienda secundaria por no disponer de una renta lo suficientemente elevada.

Otra posible explicación de este resultado es un fenómeno muy usual en las viviendas. La situación de las mismas puede encarecerlas. Por ejemplo, un

apartamento en primera línea de playa puede tener un precio mucho más elevado que un chalé con las mismas o mejores características que esté más alejado. Serán hogares con niveles de renta elevados los que accedan al apartamento de primera línea.

El signo positivo del coeficiente estimado de la variable MURB, permite concluir que los hogares que disponen de segunda vivienda, pero cuya residencia principal está ubicada en un ámbito urbano presentan mayor tendencia a disfrutar de una vivienda secundaria de tipo unifamiliar que los hogares que disponen de vivienda secundaria pero viven en zonas no urbanas.

Con este resultado se observa la tendencia que presentan las familias que viven habitualmente en ciudades a buscar viviendas unifamiliares, y por tanto más independientes de otras familias, para su ocio. Esta actitud podría verse como el reflejo de intentar satisfacer la necesidad de aislamiento y evitar la aglomeración que puede representar el vivir en una ciudad.

Por el contrario, los hogares que residen en zonas no urbanas no presentan la misma necesidad de aislamiento, ya que habitualmente no tendrán estos inconvenientes que presenta la vida urbana.

El modelo propuesto en este análisis resulta más adecuado que un modelo constante. El valor del estadístico de la razón de verosimilitudes indica que hay que aceptar el modelo propuesto como correcto para cualquier nivel de significación, y además todas las variables consideradas como explicativas son muy determinantes en la elección del tipo de vivienda secundaria que disfruta el hogar.

5.5.2. Análisis de las muestras desagregadas según el ámbito rural y urbano

Se ha considerado interesante, como en los apartados anteriores, realizar el análisis sobre el tipo de vivienda que prefiere el hogar, desagregando la muestra global en dos muestras correspondientes al ámbito rural y al ámbito urbano. A priori, se espera que exista alguna diferencia en el comportamiento de los hogares en ambas situaciones ya que la variable MURB en el análisis anterior era altamente significativa.

En la tabla 35, se encuentran los resultados de los coeficientes estimados y los niveles críticos obtenidos con ambas muestras. Los resultados se comentan

comparando esta tabla con la tabla 34, donde se presentaban los resultados para la muestra global.

Para la muestra del ámbito urbano, las variables independientes que son significativas coinciden con las de la muestra global. Mientras que, para la muestra del ámbito rural hay algunos cambios, ya que las variables SEXO y ESTUDIOS del sustentador principal no son significativas y sí que lo es MIEMHOG.

TABLA 35

Análisis del tipo de vivienda secundaria con las muestras del ámbito rural y del ámbito urbano

Variable dependiente: UNIFAM

	RURAL			URBANO		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	9,21990	3,272	0,00107	5,00250	2,359	0,01831
Sexo	0,33605	1,286	0,19841	0,36769	2,074	0,03811
Estudio1	0,39220	1,453	0,14621	0,36767	2,396	0,01658
Estudio2	0,22022	0,598	0,54978	0,33275	1,730	0,08370
Edad	-0,00082	-0,017	0,98621	0,02354	0,617	0,53744
Edad2	0,00014	0,310	0,75643	-0,00015	-0,422	0,67315
Lrenta	-0,70130	-3,895	0,00010	-0,38384	-2,881	0,00397
Miemhog	0,11566	1,958	0,05020	-0,02060	-0,448	0,65394
N° observaciones	721			1.375		
Log-verosimilitud	-472,2481			-877,5646		
Log-veros. restrin.	-489,5577			-895,5337		
Chi-cuadrado (7)	34,61909			35,93809		
Nivel signific.	0,0000132			0,0000074		
% predic. correc.	60,33%			64,51%		

En cuanto al signo de los coeficientes estimados con ambas muestras se observan muy pocos cambios con respecto a los obtenidos con la muestra global (tabla 34). De hecho estos cambios están en variables no significativas.

Observando los resultados de la estimación se podría concluir que sí que hay variaciones en el comportamiento de los hogares que habitualmente residen en un ámbito rural y el de los que residen en un ámbito urbano cuando se plantean la elección entre una vivienda unifamiliar o no unifamiliar como vivienda secundaria.

La principal diferencia se encuentra en las variables que son determinantes de la elección en cada uno de los ámbitos comentados. Mientras el ámbito urbano se comporta igual que la muestra global no ocurre así en el ámbito rural, en el que únicamente son significativas el número de miembros del hogar y la renta. Las características que determinan la elección entre vivienda secundaria unifamiliar o no unifamiliar no son las mismas en un ámbito rural que en un ámbito urbano, y además, algunas influyen de forma contraria.

Para la bondad del ajuste, en los dos modelos el test de la razón de verosimilitudes para contrastar la hipótesis nula H_0: *modelo constante* frente a la hipótesis alternativa H_1: *modelo propuesto*, lleva a rechazar la hipótesis nula; es decir, se debe aceptar el modelo propuesto en el análisis como adecuado para cualquier nivel de significación.

5.5.3. Análisis de la muestra global de la Comunidad Valenciana

Considerando únicamente el ámbito geográfico de la Comunidad Valenciana se ha estimado también un modelo que permita determinar las características que tienen mayor influencia para discriminar entre los hogares cuya vivienda secundaria es de tipo unifamiliar y los hogares cuya vivienda secundaria es de tipo colectivo.

Los resultados de la estimación de este modelo se encuentran en la tabla 36.

Las variables que no resultan significativas para el modelo son las que recogen las características del sustentador principal: SEXO, ESTUDIOS y EDAD.

Comparando estos resultados con los niveles críticos de la tabla 33, se puede decir que en todo el conjunto nacional las variables que determinan el tipo de vivienda de los hogares no coinciden con las que lo determinan en la Comunidad Valenciana, ya que la significatividad de las variables no es la misma. Sin embargo, si se comparan los signos de los coeficientes estimados con ambas muestras se comprueba que coinciden, aunque presentan diferencias respecto a las magnitudes.

En la Comunidad Valenciana las únicas características del hogar que son determinantes en la elección de vivienda secundaria unifamiliar o no unifamiliar son MURB, MIEMHOG y LRENTA.

TABLA 36

Análisis del tipo de vivienda secundaria con la muestra de la Comunidad Valenciana

Variable dependiente: UNIFAM

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	14,52100	5,3460	2,716	0,00661
Sexo	0,24289	0,4352	0,558	0,57680
Estudio1	0,68135	0,4505	1,513	0,13039
Estudio2	0,35817	0,5672	0,631	0,52777
Edad	0,01017	0,0771	0,132	0,89506
Edad2	-0,00008	0,0007	-0,109	0,91298
Lrenta	-1,11150	0,3502	-3,174	0,00151
Murb	0,67479	0,2887	2,337	0,01942
Miemhog	0,32118	0,1364	2,355	0,01852
N° observaciones	286			
Log-verosimilitud	-167,0572			
Log-veros. restrin.	-179,6446			
Chi-cuadrado (8)	25,1749			
Nivel signific.	0,14519E-02			
% predic. correc.	71,33%			

La interpretación del coeficiente estimado de la variable MURB permite decir que en las zonas urbanas es donde hay una mayor probabilidad por las viviendas secundarias unifamiliares.

El coeficiente estimado para el número de miembros del hogar, MIEMHOG, es positivo, lo que indica que los hogares con un mayor número de miembros tienden a disponer de viviendas secundarias de tipo unifamiliar.

El logaritmo de la RENTA disponible por el hogar tiene el coeficiente estimado con signo negativo, igual a como ocurría con la muestra global. Así los hogares con niveles de renta altos prefieren viviendas secundarias de tipo colectivo y son los hogares con los niveles de renta inferiores los que disponen de viviendas secundarias unifamiliares.

El valor del estadístico de la razón de verosimilitudes para este modelo lleva a que hay que aceptar como correcto el modelo propuesto en el análisis frente a un modelo constante para niveles de significación superiores al valor 0,001452.

5.5.4. Análisis de las muestras desagregadas según el ámbito rural y urbano en la Comunidad Valenciana

Los resultados obtenidos con la estimación del modelo de elección del tipo de vivienda secundaria (unifamiliar/no unifamiliar) para las submuestras del ámbito rural y del ámbito urbano de la Comunidad Valenciana se encuentran en la tabla 37.

Vamos a comparar estos resultados con la muestra global de la Comunidad Valenciana y con las muestras del ámbito rural y urbano de toda España.

TABLA 37

Análisis del tipo de vivienda secundaria con la muestra del ámbito rural y urbano de la Comunidad Valenciana

Variable dependiente: UNIFAM

	RURAL			URBANO		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	24,48900	2,493	0,01265	9,85560	1,485	0,13746
Sexo	0,94772	1,187	0,23507	-0,07089	-0,130	0,89658
Estudio1	0,99805	0,773	0,43948	0,67576	1,382	0,16691
Estudio2	0,05502	0,038	0,97008	0,66640	1,010	0,31230
Edad	-0,00986	-0,083	0,93402	0,01988	0,188	0,85073
Edad2	0,00004	0,034	0,97304	-0,00014	-0,133	0,89435
Lrenta	-1,78890	-2,870	0,00410	-0,77525	-1,769	0,07683
Miemhog	0,25661	1,202	0,22946	0,38387	2,113	0,03456
N° observaciones	115			171		
Log-verosimilitud	-66,02927			-98,3645		
Log-veros. restrin.	-74,90992			-104,2024		
Chi-cuadrado (7)	17,7613			11,6755		
Nivel signific.	0,0130946			0,1117446		
% predic. correc.	69,57%			69,01%		

El número de variables significativas obtenidas ahora, en comparación con las otras muestras ya analizadas, ha disminuido. En la muestra del ámbito rural la única variable significativa es la RENTA que posee el hogar, y en la muestra del ámbito urbano las variables significativas son LRENTA y el número de miembros del hogar, MIEMHOG.

Comparando el signo de los coeficientes estimados para el ámbito rural con los obtenidos con la muestra del ámbito rural en toda España, se observa que no hay ningún cambio. Las preferencias de los hogares de zonas rurales por las viviendas unifamiliares en la Comunidad Valenciana y en toda España coinciden. Tampoco hay diferencias significativas con respecto al comportamiento observado con la muestra global de la Comunidad Valenciana.

En el ámbito urbano de la Comunidad Valenciana, comparándolo con el correspondiente ámbito urbano a nivel nacional, se observa un cambio de signo en las variables SEXO y MIEMHOG. En las ciudades de la Comunidad Valenciana, si el sustentador principal es mujer hay mayores preferencias hacia viviendas secundarias de tipo unifamiliar, contrariamente a lo que ocurre en las zonas urbanas de toda España.

El número de miembros del hogar tiene un coeficiente estimado positivo en la muestra del ámbito urbano de la Comunidad Valenciana, lo que indica que si aumenta el número de miembros aumenta la probabilidad de disponer viviendas secundarias unifamiliares. En la muestra del ámbito urbano de toda España la tendencia es inversa.

Si se comparan los resultados obtenidos para la muestra anterior con los de toda la Comunidad Valenciana se puede ver que tampoco ahora hay diferencias significativas.

Con respecto a la bondad del ajuste se tiene que, en el ámbito rural, se acepta el modelo propuesto en el análisis frente a un modelo constante para niveles de significación mayores del 1,3% y en el ámbito urbano sólo puede aceptarse el modelo para niveles de significación mayores del 11%.

5.5.5. Análisis de las muestras desagregadas según la Comunidad Autónoma

En este apartado se pretende analizar el tipo de vivienda secundaria que elige el hogar según la Comunidad Autónoma en la que reside.

Al igual que en el análisis realizado en el apartado 5.1.5 también aquí se han agregado las muestras para algunas Comunidades con el fin de evitar posibles muestras de tamaño insuficiente.

De nuevo se ha estimado el modelo *logit binomial* para modelizar la elección entre vivienda unifamiliar o no unifamiliar. En esta ocasión los resultados

obtenidos muestran que prácticamente ninguna de las variables introducidas en el modelo como variables explicativas, es significativa. Por este motivo no se presentan aquí los resultados numéricos y únicamente se indican algunos comentarios.

Comparando con los resultados obtenidos con la muestra completa (tabla 34) se pueden apreciar diferencias para todas las Comunidades y todas las variables.

Se podría concluir que el comportamiento de los hogares frente a la elección entre una vivienda secundaria unifamiliar o de tipo colectivo, varía de una Comunidad Autónoma a otra. No hay ningún comportamiento generalizado con respecto a las características del hogar y su entorno, aunque las características geográficas y económicas de cada Comunidad pueden ser factores muy determinantes en los diferentes comportamientos.

En lo que respecta a la bondad del ajuste se han encontrado resultados diferentes según la Comunidad Autónoma analizada.

El valor del estadístico de la razón de verosimilitudes que proporciona el programa LIMDEP lleva a aceptar el modelo como correcto frente a un modelo nulo con las muestras de las siguientes Comunidades y agrupaciones de Comunidades: Andalucia/Extremadura, Aragón, Asturias/Cantabria, Castilla-La Mancha/Castilla-León y Murcia.

En el resto, Baleares, Cataluña, Galicia, Madrid y Navarra/País Vasco/La Rioja, hay que rechazar el modelo propuesto para niveles de significación razonables.

5.6. Análisis del tamaño de la vivienda secundaria

Un factor importante en un estudio de la vivienda es el tamaño de la misma. En el mercado de las viviendas secundarias puede resultar bastante interesante analizar que tipo de hogares son los que demandan viviendas secundarias grandes o pequeñas, de esta forma se podrá encuadrar a cada hogar en la categoría de vivienda pequeña, media o grande atendiendo a sus características.

El primer planteamiento que puede realizarse en este estudio es considerar como variable que mide el tamaño de la vivienda la variable VSM2UT que recoge los metros cuadrados útiles de la misma. Esta variable es de carácter continuo y por ello es posible utilizar las técnicas habituales de regresión lineal para encontrar la influencia que las características sociodemográficas y económicas del hogar y su entorno, tienen sobre el tamaño de la vivienda secundaria que el hogar tiene a su disposición.

Este modelo lineal estimado permitirá establecer una primera interpretación acerca del objetivo perseguido. Un segundo planteamiento es realizar comparaciones entre las viviendas clasificadas en pequeñas, medias y grandes. Para este fin es adecuada la metodología de los modelos de respuesta cualitativa, cualquier modelo de elección discreta con una variable dependiente que tome tres valores, uno para cada categoría de tamaño puede utilizarse para el análisis.

Bajo el punto de vista de los modelos de elección discreta hay dos opciones: plantear un modelo multinomial considerando que la variable dependiente, TAMAÑO, toma tres valores correspondientes a tres categorías de vivienda diferentes o utilizar un modelo que tenga en cuenta el hecho de que la variable VSM2UT que define los tamaños establece un orden entre las tres posibles categorías. En este segundo caso el modelo de elección discreta adecuado sería un modelo para alternativas ordenadas.

Al igual que en los apartados anteriores, el análisis del tamaño de la vivienda secundaria se realizará utilizando la muestra de los hogares que disponen de vivienda secundaria en propiedad, incluyendo los que la han heredado, o en alquiler. También se realizará este análisis para la muestra desagregada según el ámbito rural o urbano en el que reside habitualmente el hogar.

5.6.1. Análisis de regresión lineal

En este apartado se realiza el análisis del tamaño de la vivienda secundaria considerando un modelo de *regresión lineal*, utilizando como variable dependiente los metros cuadrados útiles de la vivienda, VSM2UT, y como variables independientes las usuales características del hogar y del entorno en el que reside.

Muestra global

En la tabla 38a se encuentran las estimaciones de los coeficientes del modelo de *regresión lineal* obtenidos con la muestra completa de los 2.096 hogares que disponen de vivienda secundaria.

TABLA 38a

Análisis del tamaño de la vivienda secundaria mediante un modelo de regresión lineal con la muestra global

Variable dependiente: VSM2UT (estimación mínimos cuadrados ordinarios)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-125,76000	45,7300	-2,750	0,00596
Sexo	0,04251	4,0720	0,010	0,99167
Estudio1	-17,23100	3,7230	-4,628	0,00000
Estudio2	-4,94390	4,7640	-1,038	0,29938
Edad	-2,19240	0,8141	-2,693	0,00708
Edad2	0,02241	0,0075	2,986	0,00283
Lrenta	18,64700	2,8920	6,448	0,00000
Murb	3,09560	2,7530	1,124	0,26088
Miemhog	0,91683	0,9953	0,921	0,35695
Nº observaciones	2.096			
Log-verosimilitud	-11.474,22			
Log-veros. restrin.	-11.543,56			
R^2 Ajustado	0,06043465			
Grd. libertad	(8 , 2.087)			
F(n , m)	17,8443			

Observando la tabla se puede concluir que el tamaño de la vivienda secundaria está determinado fundamentalmente por la renta disponible por el hogar y por la edad del sustentador principal, ya que los coeficientes asociados a las variables LRENTA, EDAD y EDAD2 son significativamente diferentes de cero.

También es sigificativa la variable ESTUDIO1, lo que permite decir que el tamaño de la vivienda secundaria que demandan los hogares cuyo sustentador principal posee estudios primarios es diferente del tamaño que demandan los universitarios, mientras que no hay diferencia de estos últimos con los de la categoría de estudios secundarios.

En cuanto a la forma en que estas variables influyen en el tamaño de la vivienda secundaria se puede ver que no hay ningún resultado sorprendente.

Los estudios del sustentador principal tienen asociados coeficientes negativos y crecientes con el nivel de estudios. Este resultado indica que los hogares cuyo sustentador principal tiene estudios primarios son los que ocuparán las viviendas secundarias de menor tamaño, y a medida que aumenta el nivel de estudios también aumenta el tamaño de la vivienda secundaria.

La EDAD del sustentador principal actúa como una parábola negativa: los más jovenes buscan las viviendas de menor tamaño, al aumentar la edad se prefieren viviendas con tamaños cada vez mayores y las personas de edad avanzada buscan de nuevo viviendas más pequeñas.

El logaritmo de la RENTA, como era de esperar, tiene un coeficiente estimado de signo positivo. Los hogares con niveles de renta elevados son los que prefieren las viviendas de mayor tamaño.

Muestras del ámbito rural y del ámbito urbano

Estimando el modelo de *regresión lineal* con la muestra del ámbito rural y la muestra del ámbito urbano separadamente se han obtenido los resultados que muestra la tabla 38b

Comparando estos resultados con los de la tabla 38a se puede observar que con la muestra del ámbito rural se ha obtenido que la variable ESTUDIO2 pasa a ser significativa, pero por el contrario la EDAD del sustentador principal deja de ser un factor determinante del tamaño de la vivienda secundaria.

Con la muestra del ámbito urbano únicamente hay un cambio: el número de miembros del hogar es un factor determinante en este análisis.

Se podría decir que en las ciudades un aumento del número de miembros del hogar provoca un aumento del tamaño de la vivienda secundaria, ya que el coeficiente estimado para la variable MIEMHOG es positivo.

TABLA 38b

Análisis del tamaño de la vivienda secundaria mediante un modelo de regresión lineal con las muestras del ámbito rural y del ámbito urbano

Variable dependiente: VSM2UT (estimación mínimos cuadrados ordinarios)

	RURAL			URBANO		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-92,15800	-1,606	0,10818	-149,45000	-2,341	0,01924
Sexo	0,87867	0,161	0,87192	-0,83449	-0,153	0,87870
Estudio1	-11,83900	-2,071	0,03839	-17,71200	-3,706	0,00021
Estudio2	-14,17100	-1,815	0,06950	-1,64320	-0,276	0,78271
Edad	-0,97627	-0,998	0,31831	-3,08850	-2,675	0,00748
Edad2	0,00986	1,093	0,27449	0,03116	2,929	0,00340
Lrenta	14,81800	4,076	0,00005	21,56100	5,380	0,00000
Miemhog	-1,47520	-1,218	0,22318	2,39650	1,706	0,08798
N° observaciones	721			1375		
Log-verosimilitud	-3.736,231			-7.665,341		
Log-veros. restrin.	-3.751,454			-7.714,783		
R^2 Ajustado	0,03193556			0,06462564		
Grd. libertad	(7 , 713)			(7 , 1.367)		
F(n , m)	4,393162			14,56151		

En términos generales se podría decir que no hay diferencias claras en el tamaño de la vivienda secundaria según si el hogar reside habitualmente en un ámbito rural o urbano. Únicamente destaca el hecho de que los factores determinantes del tamaño en un ámbito rural son el nivel de estudios del sustentador principal y la renta disponible por el hogar. Por el contrario en el ámbito urbano además de los anteriores también están el número de miembros del hogar y la edad del sustentador principal.

En cuanto a la forma en la que se comporta cada variable significativa no se aprecia ninguna diferencia entre las muestras utilizadas. A nivel nacional, a nivel

rural y a nivel urbano se tiene el mismo comportamiento de los hogares frente al tamaño de la vivienda secundaria.

En las tablas anteriores aparece el valor del estadístico F y sus grados de libertad. Desde estos valores se concluye que hay que aceptar como válido el modelo propuesto frente a un modelo constante para cualquier nivel de significación, en los tres casos.

5.6.2. Análisis del modelo Logit Multinomial

A continuación se presentan los resultados de la estimación de un modelo *logit multinomial* que considera las tres categorías de la variable que representa el tamaño como tres categorías independientes.

El modelo *logit multinomial* considera el valor 0 de la variable dependiente a la categoría de tamaño medio, estando el valor 1 y el 2 asociados a las categorías de tamaños pequeños y grande respectivamente. En consecuencia, el programa informático LIMDEP asignará el valor 0 a todos los coeficientes correspondientes a la categoría de tamaño medio, y los coeficientes estimados para las restantes categorías de la variable dependiente (pequeño y grande) deberán interpretarse comparándolos con la categoría de tamaño medio.

Muestra global

En la tabla 39a se encuentran las estimaciones que proporciona el programa LIMDEP para el modelo *logit multinomial.* Las dos primeras columnas son los coeficientes estimados y el nivel crítico asociado a los mismos para la categoría de viviendas pequeñas. Las dos últimas columnas corresponden a la categoría de viviendas grandes.

Respecto a la significatividad de los coeficientes estimados se puede observar en la tabla correspondiente que en la comparación entre el tamaño pequeño y el tamaño medio para las viviendas secundarias, sólo son significativos el nivel de estudios primarios frente a los universitarios (ESTUDIO), la renta disponible por el hogar (LRENTA), el ámbito de residencia habitual del hogar (MURB), y el número de miembros del hogar (MIEMHOG).

Cuando se comparan los tamaños grande e intermedio las variables significativas cambian, ya que en este caso MIEMHOG tampoco es significativa, pero sí que es un factor determinante el término cuadrático de la EDAD.

TABLA 39a

Análisis del tamaño de la vivienda secundaria mediante un modelo logit multinomial con la muestra global

Variable dependiente: TAMAÑO

	PEQUEÑA			GRANDE		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	5,76210	3,139	0,00170	-4,07540	-1,916	0,05537
Sexo	-0,04809	-0,299	0,76529	-0,04579	-0,239	0,81132
Estudio1	0,27133	1,749	0,08037	-0,41415	-2,588	0,00966
Estudio2	0,14301	0,716	0,47376	-0,05998	-0,297	0,76611
Edad	-0,01399	-0,429	0,66758	-0,05665	-1,513	0,13039
Edad2	0,00009	0,305	0,76006	0,00057	1,670	0,09488
Lrenta	-0,36239	-3,113	0,00185	0,33496	2,491	0,01273
Murb	0,21952	2,041	0,04125	0,24297	1,859	0,06305
Miemhog	-0,06648	-1,655	0,09791	-0,01379	-0,308	0,75811
N° observaciones	2.096					
Log-verosimilitud	-2.187,718					
Log-veros. restrin.	-2.239,760					
Chi-cuadrado (16)	104,0851					
Nivel signific.	0,1E-06					
% predic. correc.	43,42%					

Los coeficientes estimados para las variables que recogen los ESTUDIOS del sustentador principal permiten decir que el tamaño de la vivienda secundaria crece con el nivel de estudios. Los hogares cuyo sustentador principal tiene estudios primarios prefieren siempre las viviendas de menor tamaño, en cualquiera de las dos comparaciones. Los hogares clasificados en la categoría de estudios universitarios son los que mayor probabilidad tienen asignada a los tamaños superiores.

Con respecto a la EDAD del sustentador principal, la forma cuadrática estimada permite decir que, en la comparación entre viviendas secundarias grandes y viviendas secundarias de tamaño medio, los hogares cuyo sustentador principal es joven tienen una probabilidad mayor de ocupar viviendas secundarias de tamaño grande. Al aumentar la edad aumenta la probabilidad de los tamaños

medios y los hogares cuyo sustentador principal es de mayor edad vuelven a preferir los tamaños grandes.

Los coeficientes de la variable LRENTA que se observan en la tabla favorecen el tamaño de la vivienda secundaria. Como era previsible, un aumento del nivel de renta del hogar lleva a un aumento en la probabilidad de elegir viviendas secundarias grandes.

Observando las estimaciones obtenidas se podría concluir que en las zonas urbanas la mayor tendencia es hacia los tamaños extremos (pequeño y grande). Por el contrario, en zonas no urbanas el tamaño preferido es el medio en cualquier caso.

El número de miembros del hogar lleva asociado un coeficiente estimado negativo en la categoría de tamaño pequeño, que indicará que un aumento en la variable MIEMHOG implica un aumento en la probabilidad de elegir una vivienda secundaria de tamaño medio frente a la probabilidad de elegir un tamaño pequeño.

Como se puede apreciar, se mantienen las conclusiones obtenidas con el modelo de *regresión lineal* del apartado anterior.

Muestras del ámbito rural y del ámbito urbano

Realizando la estimación de los modelos *logit multinomial* para las muestras desagregadas según el ámbito de residencia del hogar se han encontrado los resultados de las tablas $39b_1$ y $39b_2$.

Observando las tablas correspondientes se puede ver que hay muy pocas variables significativas en las dos muestras. Tal vez el número de observaciones obtenidas en cada categoría de tamaño al realizar la desagregación según el ámbito de residencia habitual del hogar sea insuficiente.

Comparando estos resultados con los obtenidos en la estimación del mismo modelo con la muestra completa se aprecian algunos cambios.

TABLA $39b_1$

Análisis del tamaño de la vivienda secundaria mediante un modelo logit multinomial con la muestra del ámbito rural

Variable dependiente: TAMAÑO

	PEQUEÑA			GRANDE		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	6,04420	2,021	0,04332	-1,50740	-0,403	0,68683
Sexo	0,13732	0,482	0,62981	-0,07897	-0,229	0,81892
Estudio1	0,23435	0,728	0,46654	-0,57160	-1,763	0,07796
Estudio2	-0,22725	-0,529	0,59695	-0,90949	-1,926	0,05408
Edad	-0,02364	-0,460	0,64577	-0,07312	-1,193	0,23273
Edad2	0,00018	0,388	0,69773	0,00065	1,156	0,24761
Lrenta	-0,37688	-1,990	0,04658	0,24295	1,016	0,30973
Miemhog	-0,04861	-0,770	0,44122	-0,10474	-1,307	0,19128
N° observaciones	721					
Log-verosimilitud	-737,9113					
Log-veros. restrin.	-751,7895					
Chi-cuadrado (14)	27,75642					
Nivel signific.	0,1532E-01					
% predic. correc.	45,63%					

TABLA 39b

Análisis del tamaño de la vivienda secundaria mediante un modelo logit multinomial con la muestra del ámbito urbano

Variable dependiente: TAMAÑO

	PEQUEÑA			GRANDE		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	5,85410	2,484	0,01298	-5,50670	-2,083	0,03727
Sexo	-0,12493	-0,634	0,52612	-0,04332	-0,186	0,85225
Estudio1	0,26865	1,495	0,13493	-0,35197	-1,896	0,05793
Estudio2	0,25996	1,141	0,25402	0,18852	0,826	0,40858
Edad	-0,00726	-0,172	0,86356	-0,04608	-0,969	0,33264
Edad2	0,00003	0,075	0,94013	0,00051	1,171	0,24145
Lrenta	-0,35960	-2,416	0,01568	0,40500	2,460	0,01391
Miemhog	-0,07642	-1,466	0,14263	0,03014	0,539	0,58977
N° observaciones	1.375					
Log-verosimilitud	-1.444,210					
Log-veros. restrin.	-1.482,927					
Chi-cuadrado (14)	77,43281					
Nivel signific.	0,1E-06					
% predic. correc.	41,89%					

Para la muestra del ámbito rural se puede decir que la variable LRENTA es la única característica determinante de la elección entre viviendas secundarias de tamaño pequeño y de tamaño medio, y su influencia está enfocada a potenciar el tamaño superior. En la comparación entre los tamaños grande y medio el único factor determinante de la elección es el nivel de estudios del sustentador principal: los hogares con mayor probabilidad de elegir una vivienda secundaria de tamaño grande son los clasificados en la categoría de universitarios, mientras que los de estudios primarios son los que más preferencia presentan por los tamaños intermedios.

Con la muestra del ámbito urbano destaca que la variable LRENTA también es significativa en la elección entre viviendas secundarias de tamaño grande y viviendas secundarias de tamaño pequeño. De nuevo la influencia de esta variable indica un aumento de la probabilidad de los tamaños grandes. El resto de variables funcionan exáctamente igual que con la muestra del ámbito rural.

Un análisis sobre bondad del ajuste del modelo *logit multinomial* planteado lleva a la aceptación de este modelo como correcto frente a un modelo constante para cualquier nivel de significación, en la muestra completa y en la muestra del ámbito urbano.

Con la muestra del ámbito rural se acepta para niveles de significación no inferiores a 0,0015.

5.6.3. Análisis del modelo Probit Ordenado

Antes de proceder a la presentación de los resultados nótese que para la estimación del modelo *logit multinomial* y del *probit ordenado* la variable dependiente (TAMAÑO) se ha definido de forma diferente, la interpretación de la misma debe realizarse atendiendo a estas diferencias.

Para el modelo *probit ordenado* la definición de la variable TAMAÑO está realizada consecuentemente con la magnitud de la variable VSM2UT, ya que el valor 0 está asignado a las viviendas pequeñas, el valor 1 a las de tamaño medio y el valor 2 a las de tamaño grande. En este caso, sólo hay un coeficiente estimado por variable (no uno por categoría como en el *logit multinomial*) y la interpretación del mismo es más sencilla.

Además en el modelo *probit ordenado* se deben estimar tantos coeficientes como categorías menos una se disponen en el modelo. Aquí se tendrían que estimar dos coeficientes, α_1, α_2 . El programa LIMDEP realiza una normalización y estima únicamente el coeficiente $\mu_1 = \alpha_2 - \alpha_1$, cuyo valor indica el cambio de una categoría a otra de la variable respuesta.

Muestra global

En la tabla 40a se encuentran las estimaciones correspondientes al modelo *probit ordenado.*

TABLA 40a

Análisis del tamaño de la vivienda secundaria mediante un modelo probit ordenado con la muestra global

Variable dependiente: TAMAÑO

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-3,75570	0,8857	-4,240	0,00002
Sexo	0,01199	0,0788	0,152	0,87900
Estudio1	-0,29346	0,0721	-4,072	0,00005
Estudio2	-0,09427	0,0901	-1,047	0,29527
Edad	-0,01451	0,0161	-0,903	0,36677
Edad2	0,00017	0,0001	1,159	0,24632
Lrenta	0,29886	0,0561	5,322	0,00000
Murb	-0,01969	0,0544	-0,362	0,71708
Miemhog	0,02426	0,0199	1,219	0,22291
μ_1	1,08560	0,0323	33,611	0,00000
N° observaciones	2.096			
Log-verosimilitud	-2.193,648			
Log-veros. restrin.	-2.239,760			
Chi-cuadrado (8)	92,22402			
Nivel signific.	0,1E-06			
% predic. correc.	42,75%			

El primer resultado notable es la no significatividad de la variable MURB, que sí lo era en los dos modelos estimados en los apartados anteriores.

En este caso las únicas variables significativas son ESTUDIO1 y LRENTA.

El comportamiento de estas variables es el mismo que en los modelos anteriores: los ESTUDIOS influyen en forma negativa, pero creciente, en el tamaño de la vivienda secundaria, y la RENTA tiene una influencia positiva en el tamaño.

Muestras del ámbito rural y del ámbito urbano

Las estimaciones del modelo *probit ordenado* con las muestras del ámbito rural y del ámbito urbano se encuentran en la tabla 40b.

TABLA 40b

Análisis del tamaño de la vivienda secundaria mediante un modelo probit ordenado con las muestras del ámbito rural y del ámbito urbano

Variable dependiente: TAMAÑO

	RURAL			URBANO		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-2,86890	-1,994	0,04620	4,42080	-3,863	0,00011
Sexo	-0,09135	-0,649	0,51655	0,05191	0,532	0,59490
Estudio1	-0,34330	-2,347	0,01891	-0,26048	-3,085	0,00203
Estudio2	-0,25131	-1,215	0,22432	-0,03672	-0,360	0,71883
Edad	-0,01553	-0,603	0,54667	-0,01473	-0,718	0,47295
Edad2	0,00015	0,639	0,52282	0,00019	0,994	0,32037
Lrenta	0,26861	2,951	0,00316	0,32835	4,543	0,00001
Miemhog	-0,01228	-0,371	0,71072	0,04592	1,780	0,07511
	1,19640	20,703	0,00000	1,03270	26,473	0,00000
N° observaciones	721			1.375		
Log-verosimilitud	-741,1699			-1.446,622		
Log-veros. restrin.	-751,7895			-1.482,927		
Chi-cuadrado (7)	21,23934			72,60986		
Nivel signific.	0,0034316			0,0000001		
% predic. correc.	45,08%			42,33%		

Comparando estos resultados con los de la muestra global se puede apreciar que no hay prácticamente ningún cambio. Sólamente hay que matizar que en la muestra del ámbito urbano la variable MIEMHOG es significativa y contribuye de forma positiva en el tamaño de la vivienda secundaria.

Los valores del estadístico de la razón de verosimilitudes calculados para cada muestra analizada llevan a la conclusión que hay que aceptar el modelo propuesto como correcto frente a un modelo constante para cualquier nivel de significación en el caso de la muestra completa y la muestra del ámbito urbano. Para la muestra del ámbito rural se aceptará para niveles de significación superiores a 0,0034.

Comparación entre el modelo Logit Multinomial y Probit Ordenado

En el análisis del tamaño de la vivenda secundaria desde los modelos de elección discreta se han utilizado dos modelos diferentes, el *logit multinomial* y el *probit ordenado*, y los dos son aceptados como correctos frente a un modelo nulo.

En la teoría de los modelos de elección discreta existen medidas de bondad de ajuste que permiten realizar la comparación entre dos modelos cualesquiera. Una posibilidad es elegir aquel modelo que sea más verosímil, es decir, aquél cuyo valor del logaritmo de la función de verosimilitud sea mayor.

Con este razonamiento se podría decir que en el análisis del tamaño de la vivienda secundaria realizado en este trabajo se deberá elegir el modelo *logit multinomial*, ya que el modelo *probit ordenado* presenta un menor valor de la función de verosimilitud en los tres casos analizados.

5.6.4. Análisis de regresión lineal en la Comunidad Valenciana

Siguiendo el esquema del trabajo se ha procedido a analizar el tamaño de las viviendas secundarias de los hogares que residen en la Comunidad Valenciana.

Se presentan en primer lugar los resultados de la estimación del modelo de regresión lineal para la muestra total de los 286 hogares que disfrutan de vivienda secundaria en esta Comunidad y seguidamente se realiza el análisis para los 115 de ellos que residen en un ámbito rural y los 171 de un ámbito urbano.

Muestra global de la Comunidad Valenciana

En la tabla 41a se presentan los resultados de la *regresión lineal* del tamaño de la vivienda secundaria para la muestra de hogares de la Comunidad Valenciana.

TABLA 41a

Análisis del tamaño de la vivienda secundaria mediante un modelo de regresión lineal con la muestra de la Comunidad Valenciana

Variable dependiente: VSM2UT (estimación mínimos cuadrados ordinarios)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-114,02000	94,4800	-1,207	0,22851
Sexo	7,14830	8,2360	0,868	0,38616
Estudio1	-2,68060	8,6410	-0,310	0,75664
Estudio2	-0,40527	10,9800	-0,037	0,97058
Edad	-1,66060	1,4030	-1,183	0,23770
Edad2	0,01428	0,0135	1,059	0,29034
Lrenta	17,23500	6,1270	2,813	0,00526
Murb	5,83470	5,1200	1,140	0,25543
Miemhog	-3,35830	2,3380	-1,436	0,15201
N° observaciones	286			
Log-verosimilitud	-1.459,030			
Log-veros. restrin.	-1.467,526			
R^2 Ajustado	0,03046258			
Grd. libertad	(8 , 277)			
F(n , m)	2,119327			

El número de variables significativas ha disminuido con respecto al análisis realizado a nivel nacional.

Salvo el hecho de que en la Comunidad Valenciana la única variable determinante en el modelo es LRENTA, no se aprecian diferencias significativas entre los resultados obtenidos con esta muestra de hogares y los resultados obtenidos a nivel nacional (tabla 38a). La renta del hogar tiene un coeficiente estimado de signo positivo que indica que un aumento de la renta del hogar conduce a un aumento del tamaño de la vivienda secundaria.

Muestras del ámbito rural y del ámbito urbano de la Comunidad Valenciana

Los resultados de la estimación del modelo de *regresión lineal* con las muestras del ámbito rural y del ámbito urbano se encuentran recogidos en la tabla 41b:

TABLA 41b

Análisis del tamaño de la vivienda secundaria mediante un modelo de regresión lineal con las muestras del ámbito rural y urbano de la Comunidad Valenciana

Variable dependient: VSM2UT (estimación mínimos cuadrados ordinarios)

	RURAL			URBANO		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	221,87000	1,557	0,12244	-259,25000	-2,111	0,03633
Sexo	-0,27542	-0,022	0,98236	10,48700	0,991	0,32377
Estudio1	-76,98400	-3,645	0,00041	11,57400	1,184	0,23822
Estudio2	-51,37500	-2,180	0,03144	1,18920	0,091	0,92776
Edad	-3,13140	-1,603	0,11186	-0,87567	-0,448	0,65443
Edad2	0,03131	1,717	0,08896	0,00446	0,233	0,81606
Lrenta	-0,51654	-0,058	0,95349	26,79100	3,315	0,00113
Miemhog	4,21810	1,308	0,19378	-8,10740	-2,571	0,01103
N° observaciones	115			171		
Log-verosimilitud	-568,6243			-878,0001		
Log-veros. restrin.	-579,1363			-884,9756		
R^2 Ajustado	0,1125903			0,03876601		
Grd. libertad	(7 , 107)			(7 , 163)		
F(n , m)	3,066253			1,979429		

Lo más destacable es que en el ámbito rural de la Comunidad Valenciana la variable LRENTA resulta ser no significativa, contrariamente a lo que ocurre en todos los análisis que sobre el tamaño de la vivienda secundaria han sido realizados hasta el momento. Así, en las zonas rurales de la Comunidad Valenciana la renta disponible por el hogar no es determinante en la elección del tamaño de la vivienda secundaria. El hecho de que la familia posea o no un elevado nivel de renta no es un factor influyente en el tamaño de la vivienda secundaria.

En esta muestra los factores que sí han resultado determinantes del tamaño de la vivienda secundaria son el nivel de estudios (ESTUDIO1 y ESTUDIO2 son significativas) y la edad del sustentador principal (EDAD2 es significativa).

En el ámbito rural si que hay varios cambios en los coeficientes estimados con respecto a la muestra completa de la Comunidad Valenciana, y respecto a la muestra del ámbito rural a nivel nacional. El primero, ya comentado, es la no significatividad de la variable LRENTA.

Observando los coeficientes estimados para el nivel de ESTUDIOS del sustentador principal, se puede establecer el siguiente orden entre los hogares atendiendo a sus preferencias hacia las viviendas grandes: los hogares cuyo sustentador principal tiene estudios universitarios son los que más preferencia muestran por las viviendas grandes, en segundo lugar están los hogares sustentados por un individuo con estudios secundarios y en último lugar están los hogares cuyo sustentador principal tiene estudios primarios.

Para la EDAD del sustentador principal se ha obtenido un comportamiento en forma parábolica positiva que coincide con los resultados globales de la Comunidad y con los del ámbito rural a nivel nacional.

Al observar la estimación correspondiente al ámbito urbano se puede apreciar que, a diferencia de lo que ocurría con la muestra de toda la Comunidad Valenciana, la variable MIEMHOG sí que es significativa.

En este caso las únicas variables que sí que están determinando el tamaño de la vivienda secundaria son el número de miembros del hogar, MIEMHOG, y la renta disponible por el hogar, LRENTA.

Además, la influencia de la variable MIEMHOG es en sentido contrario al esperado, ya que el coeficiente asociado es negativo, indicando con ello que al aumentar el número de miembros del hogar la tendencia es disminuir el tamaño de la vivienda secundaria.

Con respecto a la bondad de ajuste se puede concluir que los valores del estadístico de la razón de verosimilitudes llevan a aceptar como válido el modelo propuesto en los tres casos, para niveles de significación razonables: para la muestra completa de Valencia se acepta para valores de mayores que el 3%, en la muestra del ámbito rural se acepta para cualquier nivel de significación y en la muestra del ámbito urbano el nivel crítico es el más alto situándose alrededor del 5,5%.

5.6.5. Análisis del modelo Probit Ordenado en la Comunidad Valenciana

No se analiza el modelo *logit multinomial* porque al realizar la estimación del mismo con la muestra de hogares que residen en la Comunidad Valenciana y las correspondientes muestras del ámbito rural y urbano, se ha obtenido que ninguna de las variables explicativas es significativa. Esto lleva a pensar la posibilidad de estar trabajando con información muestral insuficiente para este tipo de análisis. De hecho el test de la razón de verosimilitudes lleva a rechazar el modelo *logit multinomial* estimado en favor de un modelo constante. Si que se ha procedido a la estimación del modelo *probit ordenado*, cuyos resultados se comentan a continuación.

Muestra global de la Comunidad Valenciana

En la tabla 42a se encuentran los coeficientes estimados para el modelo *probit ordenado* con la muestra completa de la Comunidad Valenciana.

TABLA 42a

Análisis del tamaño de la vivienda secundaria mediante un modelo probit ordenado con la muestra de la Comunidad Valenciana

Variable dependiente: TAMAÑO

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-5,36700	2,7820	-1,929	0,05373
Sexo	0,01329	0,2389	0,056	0,95560
Estudio1	0,01504	0,2435	0,062	0,95073
Estudio2	0,05102	0,3105	0,164	0,86947
Edad	-0,06789	0,0377	-0,180	0,85727
Edad2	0,00007	0,0004	0,190	0,84921
Lrenta	0,39304	0,1850	2,124	0,03365
Murb	0,14997	0,1483	1,011	0,31193
Miemhog	-0,04694	0,0662	-0,709	0,47831
μ_1	0,97989	0,0839	11,682	0,00000
N° observaciones	286			
Log-verosimilitud	-301,1687			
Log-veros. restrin.	-305,8156			
Chi-cuadrado (8)	9,293694			
Nivel signific.	0,3181293			
% predic. correc.	43,36%			

Al comparar estos resultados con los obtenidos mediante el modelo de *regresión lineal* se observa la equivalencia entre ambos. La única característica determinante de la elección del tamaño de la vivienda secundaria es el nivel de renta disponible por el hogar. Esta variable contribuye de forma positiva en el tamaño.

Muestras del ámbito rural y del ámbito urbano de la Comunidad Valenciana

Los resultados de la estimación correspondientes a estas muestras se recogen en la tabla 42b.

De la tabla de resultados siguiente se puede concluir que hay diferencias importantes entre los hogares de una zona rural y los de una zona urbana de la Comunidad Valenciana cuando se plantean la elección del tamaño de la vivienda secundaria.

TABLA 42b

Análisis del tamaño de la vivienda secundaria mediante un modelo probit ordenado con las muestras del ámbito rural y urbano de la Comunidad Valenciana

Variable dependiente: TAMAÑO

	RURAL			URBANO		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	4,97640	0,916	0,35977	-2,95060	-0,844	0,39878
Sexo	-0,52962	-1,277	0,20155	0,06021	0,202	0,83965
Estudio1	-1,73940	-2,182	0,02913	0,07315	0,270	0,78751
Estudio2	-0,75049	-0,718	0,47269	-0,14085	-0,401	0,68832
Edad	-0,11502	-1,723	0,08485	0,02786	0,490	0,62402
Edad2	0,00121	1,907	0,05657	-0,00039	-0,717	0,47345
Lrenta	-0,08077	-0,231	0,81743	0,21386	0,937	0,34857
Miemhog	0,24872	2,113	0,03463	-0,08264	-0,934	0,35023
	1,15320	7,253	0,00000	0,92273	8,794	0,00000
N° observaciones	115			171		
Log-verosimilitud	-109,1340			-182,6599		
Log-veros. restrin.	-118,3450			-185,7468		
Chi-cuadrado (7)	18,42192			6,173740		
Nivel signific.	0,01020469			0,5196156		
% predic. correc.	55,65%			37,43%		

La primera diferencia está en el hecho de que mientras en el ámbito urbano no se ha encontrado ninguna característica determinante de dicha elección, en el ámbito rural se tiene que el nivel de estudios y la edad del sustentador principal y el número de miembros del hogar son factores que influyen en esta decisión.

En cuanto a los ESTUDIOS se puede decir que influye de forma negativa y creciente en el tamaño de la vivienda secundaria. Por el contrario el número de miembros del hogar contribuye de forma positiva y la edad del sustentador principal también.

En cuanto a la bondad de ajuste se tiene que el modelo *probit ordenado* únicamente es aceptable para la muestra del ámbito rural. Hay que rechazar este modelo frente a un modelo nulo en el caso del ámbito urbano y con la muestra completa de la Comunidad Valenciana.

5.6.6. Análisis del tamaño de la vivienda secundaria para las diferentes Comunidades Autónomas

Al igual que en los análisis anteriores también ahora se ha realizado la estimación de los modelos correspondientes para modelizar la elección del tamaño de la vivienda secundaria según la Comunidad Autónoma de residencia del hogar.

Debido a la semejanza entre los resultados obtenidos con los diversos modelos alternativos que se han utilizado para analizar el tamaño de la vivienda secundaria en los apartados anteriores, para este análisis únicamente se plantea la estimación del modelo *probit ordenado*.

No se presentan aquí las tablas de resultados puesto que en prácticamente ninguna Comunidad Autónoma considerada se han encontrado variables significativas. Las características del hogar y su entorno no son determinantes del tamaño de la vivienda secundaria.

Observando el valor del estadístico de la razón de verosimilitudes que proporciona el programa LIMDEP, se concluye que hay que rechazar el modelo propuesto en favor de un modelo constante en casi todas las muestras analizadas (sólo se acepta para Andalucia/Extremadura, Castilla-La Mancha/Castilla-León y Galicia).

APÉNDICE B: Tablas de resultados para las diferentes Comunidades Autónomas

B1. Análisis sobre la disponibilidad de vivienda secundaria

TABLA 26a

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Andalucia/Extremadura

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-31,84600	2,0850	-15,272	0,00000
Sexo	0,03286	0,1840	0,179	0,85827
Estudio1	-0,37965	0,1812	-2,096	0,03611
Estudio2	0,00622	0,2316	0,027	0,97856
Edad	0,21002	0,0392	5,365	0,00000
Edad2	-0,00183	0,0004	-4,931	0,00000
Lrenta	1,66520	0,1333	12,491	0,00000
Murb	0,03677	0,1289	0,285	0,77550
Miemhog	-0,06965	0,0424	-1,645	0,10006
N° observaciones	4.504			
Log-verosimilitud	-1.001,275			
Log-veros restrin.	-1.187,700			
Chi-cuadrado (8)	372,8485			
Nivel signific.	0,000000			
% predic. correc.	92,58%			

TABLA 26b

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Aragón

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-27,83900	3,6470	-7,634	0,00000
Sexo	0,35535	0,3289	1,080	0,27999
Estudio1	0,12393	0,3101	0,400	0,68946
Estudio2	-0,22505	0,4711	-0,478	0,63283
Edad	0,29298	0,0667	4,397	0,00001
Edad2	-0,00258	0,0006	-4,156	0,00003
Lrenta	1,20070	0,2289	5,244	0,00000
Murb	1,15990	0,2354	4,928	0,00000
Miemhog	-0,18994	0,0852	-2,230	0,02575
N° observaciones	1.105			
Log-verosimilitud	-345,7722			
Log-veros. restrin.	-400,2427			
Chi-cuadrado (8)	108,9410			
Nivel signific.	0,000000			
% predic. correc.	88,33%			

TABLA 26c

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Asturias/Cantabria

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-37,47400	5,5310	-6,775	0,00000
Sexo	0,10228	0,4182	0,245	0,80679
Estudio1	-0,46414	0,4184	-1,109	0,26724
Estudio2	-0,85643	0,5601	-1,529	0,12625
Edad	0,41588	0,1077	3,861	0,00011
Edad2	-0,00358	0,0010	-3,720	0,00020
Lrenta	1,66000	0,3340	4,969	0,00000
Murb	-0,01534	0,2849	-0,054	0,95707
Miemhog	-0,08879	0,0994	-0,893	0,37192
N° observaciones	804			
Log-verosimilitud	-206,4794			
Log-veros. restrin.	-242,4065			
Chi-cuadrado (8)	71,85422			
Nivel signific.	0,1E-06			
% predic. correc.	90,80%			

TABLA 26d

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Baleares

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-36,34000	5,5330	-6,567	0,00000
Sexo	0,98762	0,4358	2,266	0,02343
Estudio1	-0,07613	0,5563	-0,137	0,89115
Estudio2	0,72065	0,5889	1,224	0,22103
Edad	0,15722	0,0689	2,280	0,02262
Edad2	-0,00109	0,0006	-1,721	0,08518
Lrenta	2,02680	0,3627	5,587	0,00000
Murb	-0,22525	0,3018	-0,746	0,45543
Miemhog	-0,21696	0,1418	-1,530	0,12610
N° observaciones	429			
Log-verosimilitud	-158,2927			
Log-veros. restrin.	-197,2624			
Chi-cuadrado (8)	77,93936			
Nivel signific.	0,1E-06			
% predic. correc.	84,85%			

TABLA 26e

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Castilla-La Mancha/Castilla-León

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-29,05800	1,8310	-15,873	0,00000
Sexo	-0,03909	0,1578	-0,248	0,80434
Estudio1	-0,20830	0,1637	-1,272	0,20329
Estudio2	-0,03139	0,2074	-0,151	0,87974
Edad	0,18825	0,0304	6,202	0,00000
Edad2	-0,00153	0,0003	-5,567	0,00000
Lrenta	1,50250	0,1174	12,798	0,00000
Murb	0,29235	0,1150	2,542	0,01102
Miemhog	-0,14583	0,0427	-3,416	0,00063
N° observaciones	4.856			
Log-verosimilitud	-1.286,368			
Log-veros. restrin.	-1.466,716			
Chi-cuadrado (8)	360,6971			
Nivel signific.	0,00000			
% predic. correc.	90,90%			

TABLA 26f

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Cataluña

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-31,21000	2,9870	-10,448	0,00000
Sexo	0,13719	0,2680	0,512	0,60878
Estudio1	-0,56300	0,2299	-2,448	0,01435
Estudio2	-0,35852	0,3004	-1,193	0,23272
Edad	0,23139	0,0515	4,493	0,00001
Edad2	-0,00187	0,0005	-3,875	0,00011
Lrenta	1,55860	0,1936	8,051	0,00000
Murb	0,71443	0,1679	4,256	0,00002
Miemhog	-0,11555	0,0730	-1,584	0,11326
N° observaciones	1.644			
Log-verosimilitud	-517,3496			
Log-veros. restrin.	-620,3824			
Chi-cuadrado (8)	206,0657			
Nivel signific.	0,00000			
% predic. correc.	87,41%			

TABLA 26g

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Galicia

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-26,96100	3,4500	-7,816	0,00000
Sexo	0,19024	0,2785	0,683	0,49455
Estudio1	-0,73568	0,2881	-2,553	0,01068
Estudio2	0,10093	0,3528	0,286	0,77478
Edad	0,22792	0,0603	3,781	0,00016
Edad2	-0,00191	0,0006	-3,472	0,00052
Lrenta	1,28450	0,2152	5,969	0,00000
Murb	0,16499	0,2131	0,774	0,43884
Miemhog	-0,13803	0,0765	-1,805	0,07101
N° observaciones	1.739			
Log-verosimilitud	-378,5867			
Log-veros. restrin.	-428,7431			
Chi-cuadrado (8)	100,3127			
Nivel signific.	0,00000			
% predic. correc.	93,16%			

TABLA 26h

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Madrid

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-36,75000	4,7020	-7,816	0,00000
Sexo	0,38380	0,3864	0,993	0,32053
Estudio1	0,22482	0,3468	0,648	0,51682
Estudio2	0,03335	0,3985	0,084	0,93330
Edad	0,22077	0,0757	2,915	0,00355
Edad2	-0,00179	0,0007	-2,538	0,01114
Lrenta	1,85200	0,2981	6,213	0,00000
Murb	0,75035	0,5615	1,336	0,18143
Miemhog	-0,03976	0,0987	-0,403	0,68707
N° observaciones	764			
Log-verosimilitud	-248,6391			
Log-veros. restrin.	-298,3768			
Chi-cuadrado (8)	99,47539			
Nivel signific.	0,00000			
% predic. correc.	87,83%			

TABLA 26i

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Murcia

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-38,54100	5,1040	-7,551	0,00000
Sexo	0,49448	0,4462	1,108	0,26774
Estudio1	-0,26434	0,4508	-0,586	0,55765
Estudio2	-0,11126	0,5526	-0,201	0,84042
Edad	0,12499	0,0789	1,583	0,11339
Edad2	-0,00096	0,0007	-1,286	0,19844
Lrenta	2,23060	0,3328	6,702	0,00000
Murb	1,12810	0,3451	3,269	0,00108
Miemhog	-0,07560	0,1028	-0,736	0,46187
N° observaciones	526			
Log-verosimilitud	-170,5133			
Log-veros. restrin.	-237,5309			
Chi-cuadrado (8)	134,0352			
Nivel signific.	0,00000			
% predic. correc.	86,12%			

TABLA 26j

Análisis sobre la disponibilidad de vivienda secundaria con la muestra de Navarra/País Vasco/La Rioja

Variable dependiente: SECUND

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-29,7490	2,7610	-10,775	0,00000
Sexo	0,38962	0,2392	1,629	0,10331
Estudio1	0,13989	0,2277	0,614	0,53898
Estudio2	0,12399	0,2833	0,438	0,66159
Edad	0,30056	0,0556	5,494	0,00000
Edad2	-0,00268	0,0005	-5,109	0,00000
Lrenta	1,29010	0,1689	7,639	0,00000
Murb	0,33703	0,1596	2,112	0,03466
Miemhog	-0,07659	0,0564	-1,358	0,17435
N° observaciones	2.084			
Log-verosimilitud	-643,9199			
Log-veros. restrin.	-725,8101			
Chi-cuadrado (8)	163,7804			
Nivel signific.	0,00000			
% predic. correc.	88,82%			

B2. Análisis conjunto de la elección del régimen de tenencia de la vivienda principal y la elección entre disponer o no de una vivienda secundaria

TABLA $31a_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Andalucía / Extremadura

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-25,29300	1,6080	-15,729	0,00000
Sexo	0,35862	0,1343	2,670	0,00759
Estudio1	1,68820	0,1835	9,201	0,00000
Estudio2	0,70478	0,2148	3,281	0,00104
Edad	0,15106	0,0197	7,675	0,00000
Edad2	-0,00117	0,0002	-6,227	0,00000
Lrenta	1,56640	0,1079	14,523	0,00000
Murb	-1,07360	0,1098	-9,778	0,00000
Miemhog	-0,12563	0,0372	-3,373	0,00074
Nº observaciones	4.107			
Log-verosimilitud	-1.310,176			
Log-veros. restrin.	-1.569,867			
Chi-cuadrado (8)	519,3816			
Nivel signific.	0,000000			
% predic. correc.	88,19%			

TABLA $31a_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Andalucía/Extremadura

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-33,51400	-13,855	0,00000	-28,83900	-4,451	0,00001
Sexo	0,01488	0,073	0,94146	0,18398	0,285	0,77565
Estudio1	-0,42780	-2,131	0,03308	-0,58316	-0,975	0,32974
Estudio2	0,02409	0,095	0,92416	-0,20089	-0,267	0,78960
Edad	0,21396	4,772	0,00000	0,26383	2,144	0,03206
Edad2	-0,00184	-4,376	0,00001	-0,00226	-1,898	0,05762
Lrenta	1,76740	11,666	0,00000	1,42430	3,354	0,00080
Murb	0,15651	1,108	0,26775	-1,13820	-2,361	0,01824
Miemhog	-0,08467	-1,799	0,07202	-0,15737	-0,968	0,33312
Nº observaciones	3.582			525		
Log-verosimilitud	-827,9789			-81,97566		
Log-veros. restrin.	-992,2870			-100,5081		
Chi-cuadrado (8)	328,6163			37,06497		
Nivel signific.	0,000000			0,1119E-04		
% predic. correc.	92,07%			95,43%		

TABLA $31b_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Aragón

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-20,92900	3,3380	-6,270	0,00000
Sexo	0,53068	0,2684	1,977	0,04801
Estudio1	0,68493	0,3362	2,037	0,04162
Estudio2	0,08577	0,4026	0,213	0,83129
Edad	0,15043	0,0413	3,642	0,00027
Edad2	-0,00110	0,0004	-2,832	0,00462
Lrenta	1,28570	0,2301	5,588	0,00000
Murb	-0,49293	0,2116	-2,330	0,01982
Miemhog	-0,25250	0,0831	-3,040	0,00236
N° observaciones	1.038			
Log-verosimilitud	-353,3261			
Log-veros. restrin.	-397,3654			
Chi-cuadrado (8)	88,07872			
Nivel signific.	0,1E-06			
% predic. correc.	87,77%			

TABLA $31b_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Aragón

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-34,68800	-7,561	0,00000	-16,91300	-1,567	0,11703
Sexo	0,59152	1,500	0,13359	0,39645	0,464	0,64298
Estudio1	0,41093	1,127	0,25983	-1,52020	-1,654	0,09809
Estudio2	-0,18723	-0,320	0,74934	0,16946	0,174	0,86201
Edad	0,36519	4,242	0,00002	0,23851	1,625	0,10424
Edad2	-0,00317	-4,021	0,00006	-0,00178	-1,327	0,18444
Lrenta	1,47710	5,437	0,00000	0,52137	0,725	0,46846
Murb	1,62960	5,464	0,00000	0,35954	0,547	0,58425
Miemhog	-0,27700	-2,729	0,00635	0,24605	1,031	0,30255
N° observaciones	905			133		
Log-verosimilitud	-258,3786			-41,04744		
Log-veros. restrin.	-316,6161			-48,88066		
Chi-cuadrado (8)	116,4751			15,66644		
Nivel signific.	0,000000			0,4741E-01		
% predic. correc.	88,62%			88,72%		

TABLA $31c_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Asturias / Cantabria

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-25,76600	3,6350	-7,089	0,00000
Sexo	-0,34026	0,3145	-1,082	0,27926
Estudio1	0,76193	0,4151	1,836	0,06641
Estudio2	0,23631	0,4617	0,512	0,60881
Edad	0,08309	0,0467	1,779	0,07524
Edad2	-0,00055	0,0004	-1,276	0,20213
Lrenta	1,73830	0,2593	6,703	0,00000
Murb	-0,49113	0,2302	-2,134	0,03286
Miemhog	-0,10595	0,0834	-1,270	0,20417
N° observaciones	749			
Log-verosimilitud	-289,6578			
Log-veros. restrin.	-331,2296			
Chi-cuadrado (8)	83,1465			
Nivel signific.	0,1E-06			
% predic. correc.	85,18%			

TABLA $31c_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Asturias/Cantabria

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-37,09200	-6,037	0,00000	-77,25900	-0,109	0,91327
Sexo	0,03435	0,081	0,93564	10,09500	0,028	0,97752
Estudio1	-0,78241	-1,763	0,07783	13,19800	0,025	0,98039
Estudio2	-1,46770	-2,274	0,02297	12,82100	0,024	0,98095
Edad	0,47853	3,937	0,00008	0,36355	0,669	0,50324
Edad2	-0,00409	-3,822	0,00013	-0,00373	-0,683	0,49485
Lrenta	1,53360	4,263	0,00002	2,21760	1,241	0,21449
Murb	-0,06701	-0,226	0,82153	12,01600	0,041	0,96727
Miemhog	-0,05363	-0,510	0,61035	-0,36668	-0,532	0,59447
N° observaciones	628			121		
Log-verosimilitud	-181,7983			-9,69961		
Log-veros. restrin.	-213,2274			-14,05403		
Chi-cuadrado (8)	62,85818			8,708854		
Nivel signific.	0,1E-06			0,3674488		
% predic. correc.	89,65%			97,52%		

TABLA $31d_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Baleares

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-24,50500	4,1810	-5,861	0,00000
Sexo	0,30441	0,3276	0,929	0,35280
Estudio1	0,28621	0,5024	0,570	0,56892
Estudio2	1,06040	0,6187	1,714	0,08653
Edad	0,08979	0,0504	1,782	0,07475
Edad2	-0,00044	0,0005	-0,899	0,36885
Lrenta	1,53500	0,2807	5,468	0,00000
Murb	-0,14024	0,2540	-0,552	0,58094
Miemhog	-0,18441	0,1109	-1,663	0,09634
N° observaciones	394			
Log-verosimilitud	-202,4568			
Log-veros. restrin.	-238,7871			
Chi-cuadrado (8)	72,66073			
Nivel signific.	0,1E-06			
% predic. correc.	75,63%			

TABLA $31d_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Baleares

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-40,26000	-5,792	0,00000	-45,21900	-2,976	0,00292
Sexo	1,02490	1,989	0,04667	1,21260	1,234	0,21738
Estudio1	-0,03353	-0,048	0,96170	-0,03979	-0,037	0,97023
Estudio2	0,82830	1,125	0,26072	0,87336	0,650	0,51549
Edad	0,23467	2,577	0,00995	0,06979	0,443	0,65799
Edad2	-0,00179	-2,175	0,02962	-0,00036	-0,225	0,82175
Lrenta	2,21050	4,881	0,00000	2,72970	2,788	0,00531
Murb	-0,64572	-1,751	0,07990	1,11460	1,462	0,14361
Miemhog	-0,45669	-2,365	0,01805	-0,15482	-0,626	0,53157
N° observaciones	278			116		
Log-verosimilitud	-111,0797			-31,68325		
Log-veros. restrin.	-139,6646			-42,72243		
Chi-cuadrado (8)	57,16984			22,07836		
Nivel signific.	0,1028E-08			0,004772		
% predic. correc.	84,89%			87,93%		

TABLA $31e_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Castilla-La Mancha / Castilla-León

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-19,31200	1,5480	-12,476	0,00000
Sexo	0,20058	0,1328	1,511	0,13090
Estudio1	1,08050	0,1577	6,849	0,00000
Estudio2	0,95586	0,1996	4,788	0,00000
Edad	0,16060	0,0192	8,369	0,00000
Edad2	-0,00117	0,0002	-6,442	0,00000
Lrenta	1,12900	0,1053	10,718	0,00000
Murb	-1,07370	0,1060	-10,134	0,00000
Miemhog	-0,11000	0,0384	-2,867	0,00414
N° observaciones	4.478			
Log-verosimilitud	-1.508,192			
Log-veros. restrin.	-1.733,772			
Chi-cuadrado (8)	451,1596			
Nivel signific.	0,00000			
% predic. correc.	87,47%			

TABLA $31e_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Castilla-La Mancha / Castilla-León

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-28,63900	-13,506	0,00000	-27,57000	-5,920	0,00000
Sexo	-0,07197	-0,407	0,68422	-0,23056	-0,568	0,57017
Estudio1	-0,33203	-1,809	0,07042	0,17350	0,382	0,70251
Estudio2	-0,14123	-0,602	0,54723	0,20860	0,358	0,72033
Edad	0,16982	4,831	0,00000	0,20479	2,458	0,01395
Edad2	-0,00135	-4,320	0,00002	-0,00177	-2,191	0,02846
Lrenta	1,50080	11,196	0,00000	1,41060	4,722	0,00000
Murb	0,38609	2,976	0,00292	0,11710	0,305	0,76062
Miemhog	-0,13157	-2,704	0,00684	-0,20220	-1,722	0,08500
N° observaciones	3.894			584		
Log-verosimilitud	-1.036,983			-144,6786		
Log-veros. restrin.	-1.181,642			-165,9086		
Chi-cuadrado (8)	289,3181			42,45998		
Nivel signific.	0,0000000			0,1109E-05		
% predic. correc.	90,78%			91,95%		

TABLA $31f_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Cataluña

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-20,49700	2,2060	-9,290	0,00000
Sexo	-0,06473	0,1869	-0,346	0,72902
Estudio1	0,79430	0,2148	3,697	0,00022
Estudio2	0,64782	0,2588	2,503	0,01232
Edad	0,08083	0,0263	3,068	0,00215
Edad2	-0,00052	0,0003	-2,044	0,04092
Lrenta	1,30350	0,1511	8,628	0,00000
Murb	-0,70267	0,1297	-5,418	0,00000
Miemhog	-0,02959	0,0595	-0,497	0,61911
N° observaciones	1.551			
Log-verosimilitud	-769,9034			
Log-veros. restrin.	-850,9717			
Chi-cuadrado (8)	162,1367			
Nivel signific.	0,00000			
% predic. correc.	78,21%			

TABLA $31f_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Cataluña

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-30,64000	-8,608	0,00000	-37,28500	-5080	0,00000
Sexo	-0,03455	-0,115	0,90883	0,21621	0,322	0,74762
Estudio1	-0,48549	-1,804	0,07122	-0,49183	-0899	0,36873
Estudio2	-0,37517	-1,064	0,28735	0,02165	0,033	0,97403
Edad	0,18666	3,190	0,00142	0,50471	3,419	0,00063
Edad2	-0,00143	-2,651	0,00802	-0,00469	-3,220	0,00128
Lrenta	1,57900	6,860	0,00000	1,66890	3,543	0,00040
Murb	1,02980	5,251	0,00000	-0,60171	-1,493	0,13535
Miemhog	-0,08869	-1,052	0,29261	-0,41788	-2,359	0,01831
N° observaciones	1.182			369		
Log-verosimilitud	-388,9871			-94,5338		
Log-veros. restrin.	-466,7546			-120,1752		
Chi-cuadrado (8)	155,5350			51,28281		
Nivel signific.	0,000000			0,2251E-07		
% predic. correc.	86,80%			90,51%		

TABLA $31g_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Galicia

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-18,40700	2,4990	-7,517	0,00000
Sexo	0,15674	0,1941	0,807	0,41939
Estudio1	1,04180	0,2666	3,908	0,00009
Estudio2	0,14176	0,3034	0,467	0,64031
Edad	0,10769	0,0321	3,355	0,00079
Edad2	-0,00066	0,0003	-1,968	0,04909
Lrenta	1,11050	0,1629	6,818	0,00000
Murb	-1,25440	0,1685	-7,442	0,00000
Miemhog	0,05215	0,0628	0,830	0,40638
N° observaciones	1.619			
Log-verosimilitud	-566,2928			
Log-veros. restrin.	-686,3576			
Chi-cuadrado (8)	240,1296			
Nivel signific.	0,000000			
% predic. correc.	85,98%			

TABLA $31g_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Galicia

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-27,94000	-6,747	0,00000	-39,58900	-3,927	0,00009
Sexo	0,07910	0,250	0,80270	1,22810	1,418	0,15629
Estudio1	-1,09150	-3,395	0,00069	0,81331	0,951	0,34165
Estudio2	-0,30058	-0,711	0,47734	2,02960	2,245	0,02476
Edad	0,28416	3,552	0,00038	0,08561	0,618	0,53675
Edad2	-0,00246	-3,409	0,00065	-0,00027	-0,200	0,84175
Lrenta	1,28040	5,171	0,00000	2,17020	3,523	0,00043
Murb	0,09561	0,393	0,69459	0,14082	0,229	0,81868
Miemhog	-0,12869	-1,493	0,13549	-0,01492	-0,071	0,94379
N° observaciones	1.375			244		
Log-verosimilitud	-290,3459			-51,58101		
Log-veros. restrin.	-335,0168			-66,74212		
Chi-cuadrado (8)	89,34182			30,32221		
Nivel signific.	0,1E-06			0,1854E-03		
% predic. correc.	93,45%			93,44%		

TABLA $31h_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Madrid

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-28,45100	3,6000	-7,903	0,00000
Sexo	0,68729	0,2788	2,465	0,01371
Estudio1	1,67300	0,3347	4,998	0,00000
Estudio2	1,02640	0,3699	2,775	0,00552
Edad	0,09001	0,0480	1,875	0,06084
Edad2	-0,00069	0,0004	-1,495	0,13483
Lrenta	1,80440	0,2442	7,390	0,00000
Murb	-0,21389	0,3828	-0,559	0,57630
Miemhog	-0,16143	0,0965	-1,672	0,09450
N° observaciones	715			
Log-verosimilitud	-288,8993			
Log-veros. restrin.	-337,5008			
Chi-cuadrado (8)	97,20314			
Nivel signific.	0,00000			
% predic. correc.	84,34%			

TABLA $31h_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Madrid

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-33,16900	-6,395	0,00000	-100,70000	-3,387	0,00071
Sexo	0,60188	1,310	0,19019	-0,65633	-0,653	0,51362
Estudio1	-0,08716	-0,227	0,82075	3,69580	2,073	0,03822
Estudio2	-0,35181	-0,803	0,42216	4,31980	2,354	0,01859
Edad	0,26890	2,858	0,00427	0,41304	1,617	0,10585
Edad2	-0,00221	-2,549	0,01081	-0,00349	-1,420	0,15570
Lrenta	1,51850	4,627	0,00000	5,96600	3,319	0,00090
Murb	0,86800	1,357	0,17491	-1,97330	-1,277	0,20153
Miemhog	-0,02679	-0,246	0,80539	-0,60721	-1,356	0,17520
N° observaciones	586			129		
Log-verosimilitud	-207,2126			-21,23828		
Log-veros. restrin.	-242,6204			-39,92257		
Chi-cuadrado (8)	70,81552			37,36857		
Nivel signific.	0,1E-06			0,984E-05		
% predic. correc.	86,18%			96,12%		

TABLA $31i_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Murcia

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-17,88100	3,9770	-4,496	0,00001
Sexo	0,06206	0,4033	0,154	0,87772
Estudio1	1,42240	0,5290	2,689	0,00717
Estudio2	0,06483	0,5683	0,114	0,90917
Edad	0,13139	0,0524	2,509	0,01212
Edad2	-0,00111	0,0005	-2,185	0,02886
Lrenta	1,06800	0,2734	3,907	0,00009
Murb	-0,33876	0,2910	-1,164	0,24436
Miemhog	-0,03869	0,0996	-0,388	0,69767
N° observaciones	472			
Log-verosimilitud	-184,9272			
Log-veros. restrin.	-206,6598			
Chi-cuadrado (8)	43,46513			
Nivel signific.	0,7172E-06			
% predic. correc.	84,32%			

TABLA $31i_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Murcia

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-40,44600	-6,580	0,00000	-30,64600	-1,535	0,12475
Sexo	0,41419	0,835	0,40361	-0,85914	-0,595	0,55159
Estudio1	-0,53253	-1,009	0,31288	0,03266	0,016	0,98690
Estudio2	-0,22612	-0,336	0,73680	0,48669	0,292	0,77013
Edad	0,12749	1,328	0,18423	0,36445	1,362	0,17306
Edad2	-0,00101	-1,118	0,26363	-0,00263	-1,155	0,24811
Lrenta	2,35760	5,949	0,00000	1,20560	0,961	0,33640
Murb	1,53300	3,594	0,00033	0,88335	0,673	0,50075
Miemhog	-0,07107	-0,559	0,57638	-0,17733	-0,392	0,69473
N° observaciones	397			75		
Log-verosimilitud	-127,1841			-16,23734		
Log-veros. restrin.	-187,9596			-20,90770		
Chi-cuadrado (8)	121,5510			9,340721		
Nivel signific.	0,00000			0,314373		
% predic. correc.	86,40%			93,33%		

TABLA $31j_1$

Análisis del régimen de tenencia de la vivienda principal con la muestra de Navarra / País Vasco / La Rioja

Variable dependiente: TENENCIA (primera etapa del proceso de eliminación)

Variables	Coeficientes	Error Std	Estad.t	Nivel
Constante	-29,34400	2,5570	-11,477	0,00000
Sexo	0,31336	0,2066	1,517	0,12933
Estudio1	0,94624	0,2476	3,822	0,00013
Estudio2	0,53665	0,2916	1,840	0,06575
Edad	0,12518	0,0325	3,847	0,00012
Edad2	-0,00095	0,0003	-3,085	0,00204
Lrenta	1,91730	0,1724	11,124	0,00000
Murb	-0,40188	0,1600	-2,511	0,01204
Miemhog	-0,22517	0,0625	-3,604	0,00031
N° observaciones	1.994			
Log-verosimilitud	-602,7665			
Log-veros. restrin.	-702,7037			
Chi-cuadrado (8)	199,8744			
Nivel signific.	0,00000			
% predic. correc.	88,97%			

TABLA $31j_2$

Análisis sobre la disponibilidad de vivienda secundaria con las muestras de propietarios e inquilinos de la vivienda principal de Navarra / País Vasco / La Rioja

Variable dependiente: SECUND (segunda etapa del proceso de eliminación)

	PROPIETARIOS			INQUILINOS		
Variable	Coef	Est. t	Nivel	Coef	Est. t	Nivel
Constante	-31,79000	-10,143	0,00000	-11,53900	-1,234	0,21728
Sexo	0,35920	1,431	0,15254	0,35993	0,305	0,76058
Estudio1	0,11923	0,483	0,62877	0,77581	0,664	0,50699
Estudio2	0,14573	0,479	0,63217	-0,03389	-0,023	0,98154
Edad	0,35770	5,578	0,00000	0,12893	0,620	0,53538
Edad2	-0,00312	-5,203	0,00000	-0,00169	-0,758	0,44841
Lrenta	1,32880	7,119	0,00000	0,37199	0,644	0,51942
Murb	0,39014	2,295	0,02174	-0,00848	-0,012	0,99068
Miemhog	-0,07704	-1,274	0,20275	0,04275	0,186	0,85249
N° observaciones	1.769			225		
Log-verosimilitud	-571,4194			-35,25711		
Log-veros. restrin.	-646,5356			-37,78743		
Chi-cuadrado (8)	150,2325			5,060649		
Nivel signific.	0,000000			0,751075		
% predic. correc.	88,07%			96%		

6. CONCLUSIONES

En este trabajo en primer lugar, se presenta una revisión metodológica de los modelos de elección discreta y de los modelos de variable dependiente limitada, junto con técnicas de estimación y contraste de hipótesis que completan el estudio de los mismos.

En segundo lugar se realiza un estudio sobre la demanda de vivienda secundaria en España, analizando con datos reales los aspectos más relevantes de la vivienda secundaria. Este estudio empírico se ha realizado utilizando modelos de respuesta cualitativa.

La revisión metodológica que se incluye en esta tesis presenta un proceso original de clasificación de los modelos de respuesta cualitativa y de los modelos de variable dependiente limitada. Para la presentación de los modelos se ha seguido el criterio de considerarlos agrupados por familias, según la distribución de probabilidad que se ha asignado al correspondiente modelo subyacente. El planteamiento general es el de considerar los modelos de elección discreta desde el punto de vista de la maximización de la utilidad que los define asumiendo que el individuo elegirá como respuesta aquella alternativa que le produzca el mayor beneficio.

Atendiendo a las características del problema que se desea analizar se pueden considerar diferentes criterios de clasificación. Una primera posibilidad es intentar reflejar únicamente la estructura interna del conjunto de las alternativas de elección. Según el tipo de relación existente entre las alternativas disponibles será más adecuado un modelo u otro, ya que cada uno tiene unas características que le permiten reflejar más fielmente una situación determinada.

La segunda alternativa para la elección y clasificación de los modelos se apoya en el proceso de decisión que sigue el individuo. Así se tendrán modelos que, además de considerar la estructura interna de las alternativas, reflejan los pasos seguidos por el decisor hasta llegar a la respuesta final.

Los modelos de variable dependiente limitada permiten analizar cualquiera de las situaciones que involucran decisiones continuas y decisiones discretas al mismo tiempo, estando asociados con las situaciones reales en las que es necesario analizar variables continuas truncadas o censuradas.

El proceso inferencial ligado a los modelos propuestos ha sido desarrollado analizando todas las posibilidades que pueden aparecer en una situación práctica, según el tipo de muestreo utilizado y según la información inicial que acerca de la población será conocida. Se han planteado tres tipos de muestreo (aleatorio simple, estratificado exógenamente y basado en la elección) y se presenta la función de verosimilitud asociada a cada uno de ellos. El metodo de estimación que se ha desarrollado en esta tesis para los modelos de respuesta cualitativa es el de maximizar la función de verosimilitud.

Aunque existen estudios realizados por diversos autores como Manski o Cosslett, que plantean la estimación en este tipo de modelos dentro de las técnicas no paramétricas, en este trabajo se ha optado por técnicas paramétricas. Las investigaciones en este campo están aún abiertas y presentan una gran variedad de posibilidades.

Estudios recientes concluyen que algunos de los problemas computacionales detectados pueden ser adecuadamente resueltos adoptando un enfoque bayesiano. Ésta podría ser una buena solución a los problemas de búsqueda del máximo de la función de verosimilitud, no obstante su aplicación necesita un conocimiento de las distribuciones de probabilidad "prior" y "posterior", que resta interés a este planteamiento. Sin embargo éste es un campo de trabajo no muy desarrollado, en el cual futuras investigaciones pueden llevar a resultados muy interesantes en los modelos de respuesta cualitativa y de variable dependiente limitada.

Los datos utilizados en este estudio han sido obtenidos de la Encuesta de Presupuestos Familiares (EPF) 1990/91, que proporciona información correspondiente a 21.155 familias repartidas por toda la geografía española.

La muestra de hogares que proporciona la encuesta ha sido desagregada en varias direcciones, considerando el ámbito de residencia habitual del hogar y considerando la situación geográfica (Comunidad Autónoma) donde habita la familia, dedicando un especial interés a la Comunidad Valenciana.

A pesar de que el número de hogares entrevistados es elevado, el número de familias con una vivienda secundaria no lo es. Únicamente 2.130 hogares de los recogidos en la encuesta disfrutan de una vivienda secundaria. Esto ha hecho que algunos de los análisis propuestos no hayan podido realizarse para todos los niveles de desagregación deseados, al obtener muestras de pequeño tamaño.

El análisis de la demanda de vivienda secundaria se ha planteado desde diversos aspectos. En primer lugar se han analizado las características familiares

El análisis de la demanda de vivienda secundaria se ha planteado desde diversos aspectos. En primer lugar se han analizado las características familiares que determinan si una familia disfruta o no de vivienda secundaria, sin distinguir el tipo de régimen de tenencia de la misma, y en caso de disponer de vivienda secundaria, en qué número.

En segundo lugar, se ha centrado el análisis en el régimen de tenencia en el que las familias disfrutan de su vivienda secundaria.

El tercer objetivo planteado en este análisis es determinar qué tipo de viviendas secundarias demandan los hogares españoles, tanto en lo referente a vivienda unifamiliar o colectiva, como en el tamaño de las mismas (pequeña, media, grande).

El carácter cualitativo de las variables dependientes que se desean analizar lleva a la utilización de modelos de elección discreta en todo el análisis.

Las decisiones binarias que se han planteado han sido modelizadas con el modelo *logit binomial*. Aquí se incluye la decisión de disponer o no de una vivienda secundaria, la elección del número de viviendas secundarias, la elección del régimen de tenencia y el tipo de vivienda secundaria (unifamiliar o no).

La decisión de disponer o no de vivienda secundaria se ha analizado también conjuntamente con el régimen de tenencia de la vivienda principal mediante un modelo de *eliminación jerárquica*.

El análisis del número de viviendas secundarias no se ha podido realizar mediante un modelo multinomial, ya que el número de hogares que disfrutan de más de una vivienda secundaria es demasiado pequeño, y se ha analizado la elección entre disponer de una vivienda secundaria y de más de una.

La elección del tamaño de la vivienda secundaria se ha analizado con tres modelos alternativos: *logit multinomial*, *probit ordenado* y *regresión lineal*. Los tres modelos proporcionan resultados equivalentes (epígrafe 5.6.).

Los análisis de las decisiones binarias anteriores se podrían haber realizado con el modelo *probit binomial* y las conclusiones obtenidas no diferirían de las que proporciona el modelo *logit binomial* (ver epígrafe 3.1.).

A continuación se presentan los resultados más destacados que se han obtenido con los análisis realizados sobre la vivienda secundaria en España.

1. Como primera conclusión global se podría decir que en el análisis planteado acerca de la demanda de vivienda secundaria la única característica determinante es el nivel de RENTA del hogar. Esta variable resulta ser significativa en todos los análisis propuestos. De hecho, salvo en el análisis sobre la disponibilidad de vivienda secundaria y su régimen de tenencia, es prácticamente el único factor determinante de las elecciones que sobre la vivienda secundaria realiza un hogar.

   El comportamiento de esta variable ha sido el esperado en todos los casos:

   a) son los hogares con los mayores niveles de renta los que presentan la mayor tendencia a tener a su disposición una vivienda secundaria.

   b) influye directamente en el número de viviendas secundarias que el hogar tiene a su diposición: un aumento en el nivel de renta proporciona más posibilidades de disfrutar de más de una vivienda secundaria.

   c) la tendencia a comprar la vivienda secundaria aumenta con la renta disponible por el hogar.

   d) en la elección entre vivienda secundaria unifamiliar o no unifamiliar la influencia de la variable renta va dirigida a potenciar las viviendas no unifamiliares. Este resultado, aparentemente sorprendente, está justificado convenientemente en el epígrafe 5.5. correspondiente.

   e) los hogares que disfrutan de las viviendas secundarias de mayor tamaño son los que tienen un mayor poder adquisitivo. En este análisis la RENTA es la única característica significativa, pero además sólo lo es en el análisis realizado a nivel nacional. En cualquier desagregación planteada, ya no puede considerarse como un factor tan determinante.

2. La variable EDAD únicamente ha resultado ser significativa en el análisis de la disponibilidad de vivienda secundaria. Su influencia está caracterizada por la forma cuadrática negativa, que indica una evolución en la tendencia hacia la segunda vivienda paralela al ciclo de vida: son los jóvenes y los más ancianos los que menos predisposición presentan a disponer de la vivienda secundaria, mientras que son los hogares cuyo cabeza de familia tiene una edad intermedia los que más posibilidades tienen de disfrutar de una segunda vivienda.

Sin embargo, una vez reducido el estudio a los hogares que sí disponen de vivienda secundaria, la característica EDAD ya no aparece como un factor influyente en prácticamente ningún análisis realizado.

Tal vez esto sea debido a que el conjunto de hogares que están recogidos en la muestra de la EPF y que disponen de vivienda secundaria, posean la característica de que el sustentador principal tenga una edad similar en todos los casos, lo que llevaría a que la variable edad no sea determinante de estas decisiones analizadas.

3. Otro resultado que merece una atención especial es el hecho de que el nivel de ESTUDIOS del sustentador principal no influye en ninguno de los análisis realizados.

4. En cuanto al resto de variables explicativas consideradas en los diferentes modelos, no pueden establecerse comportamientos generalizados respecto a la significatividad de las mismas. Dependiendo de la muestra analizada los resultados son diferentes.

   Únicamente hay que destacar que el régimen de tenencia de la vivienda principal (TENENCIA) es un factor muy determinante en la elección del régimen de tenencia de la vivienda secundaria. En este análisis se puede decir que los que tienen más posibilidades de comprar una segunda vivienda son los que ya son propietarios de la vivienda principal. Sin embargo, esta variable TENENCIA no tiene ninguna influencia sobre la decisión entre disponer o no de vivienda secundaria.

5. Los resultados obtenidos para las muestras desagregadas según el ámbito de residencia habitual del hogar, rural o urbano, proporcionan conclusiones similares a las obtenidas con la muestra global. Sin embargo, la desagregación por Comunidades Autónomas sí que lleva a conclusiones diferentes en el comportamiento de los hogares. Aunque para la Comunidad Valenciana no se ha encontrado gran diferencia con respecto al estudio realizado a nivel nacional.

   Este resultado reflejará las diferencias económicas, sociales y culturales que separan a las Comunidades entre sí. En cada Comunidad se lleva una forma de vida diferente que repercute en la actitud del hogar frente a las decisiones analizadas.

REFERENCIAS BIBLIOGRAFICAS

AGRESTI, A. (1984), *Analysis of Ordinal Categorical Variables*, New York: Wiley.

ALDRICH, J.H., NELSON, F.D. (1984), *Linear Probability, Logit and Probit Models*, Sage Publications, Beverly Hills.

AMEMIYA, T. (1973), Regression Analysis when the Dependent Variable is Truncated Normal, *Econometrica*, 41, 997-1016.

AMEMIYA, T. (1978), The Estimation of a Simultaneous Equation Generalized Probit Model, *Econometrica*, 46, 1193-1206.

AMEMIYA, T. (1979), The Estimation of a Simultaneous-Equation Tobit Model, *International Economic Review*, 20, 169-181.

AMEMIYA, T. (1980), Selection of Regressors, *International Economic Review*, 21, 331-354.

AMEMIYA, T. (1981), Qualitative response models: A survey, *Journal of Economic Literature*, 19, 1483-1536.

AMEMIYA, T. (1984), Tobit Models: A Survey, *Journal of Econometrics*, 24, 3-61.

AMEMIYA, T. (1988), Modelos de Respuesta Cualitativa: Un examen, *Cuadernos Económicos de I.C.E.*, 39, 173-246.

AMEMIYA, T., BOSKIN, M. (1974), Regresion Analysis when the Dependent Variable is Truncated Lognormal, with an Application to the Determinants of the Duration of Welfare Dependency, *International Economic Review*, 15, 485-496.

AMEMIYA, T., NOLD, F. (1975), A Modified Logit Model, *The Review of Economics and Statistics*, LVII (notes), 255-257.

AMEMIYA, T., POWELL, J.L. (1983), A Comparison of the Logit Model and Normal Discriminant Analysis when the Independent Variables are Binary, *Studies in Econometrics, Time Series and Multivariate Statistics*, Stanford University.

ANAS, A., MOSES, L.N. (1984), Qualitative Choice and the Blending of Discrete Alternatives, *The Review of Economics and Statistics*, LXVI, 547-555.

ANDERSON, G.J. (1987), Prediction Tests in Limited Dependent Variable Models, *Journal of Econometrics*, 34 Annals1, 253-261.

ASHFORD, J.R., SOWDEN, R.R. (1970), Multivariate Probit Analysis, *Biometrics*, 43, 535-546.

BERKSON, J. (1944), Application of the Logistic function to Bio-Assay, *Journal of the American Statistical Association*, 39, 357-365.

BERKSON, J. (1951), Why I Prefer Logits to Probits, *Biometrics*, 7, 327-339.

BERKSON, J. (1953), A Statistically Precise and Relatively Simple Method of Estimating the Bioassay with Quantal Response, Based on the Logistic Function, *American Statistical Association Journal*, 45, 565-599.

BLACKLEY, P., ONDRICH, J. (1988), A Limiting Joint-Choice Model for Discrete and Continuous Housing Characteristics, *The Review of Economics and Statistics*, LXX, 266-274.

BLUNDELL, R.W., SMITH, R.J. (1989), Estimation in a Class of Simultaneous Equation Limited Dependent Variable Models, *Review of Economic Studies*, 56, 37-57.

BOLDUC, D. (1992), Generalized Autoregressive Errors in the Multinomial Probit Model, *Transportation Research*, 26 B, 155-170.

BÖRSCH-SUPAN, A. (1987), *Econometric Analysis of Discrete Choice. With Applications on the Demand for Housing in the U.S. and West-Germany*, Ed. M. Beckmann, W. Krelle, Springer-Velarg.

BÖRSCH-SUPAN, A. (1990), On the Compatibility of Nested Logit Models with Utility Maximization, *Journal of Econometrics*, 43, 371-388.

BÖRSCH-SUPAN, A., PITKIN, J. (1988), On Discrete Choice Models of Housing Demand, *Journal of Urban Economics*, 24, 153-172.

BÖRSCH-SUPAN, A., POLLAKOWSKI, H.O. (1990), Estimating Housing Consumption Adjustments from Panel Data, *Journal of Urban Economics*, 27, 131-150.

BÖRSCH-SUPAN, A., STAHL, K. (1991), Do Savings Programs Dedicated to Home-Ownership Increase Personal Savings? An Analysis of the West German Bausparkassen System, *Journal of Public Economics*, 44, 265-297.

BOURASSA, S.C. (1995), A Model of Housing Tenure Choice in Australia, *Journal of Urban Economics*, 37, 161-175.

BOWER, O. (1993), Un Modelo Empírico de la Evolución de los Precios de la Vivienda en España (1976-1991), *Investigaciones Económicas*, XVIII(1), 65-83.

BREUSCH, T.S., PAGAN, A.R. (1980), The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics, *Review of Economic Studies*, 42, 239-253.

BROWN, B. (1985), Location and Housing Demand, *Journal of Urban Economics*, 17, 30-41.

BROWNSTONE, D., ENGLUND, P. (1991), The Demand for Housing in Sweden: Equilibrium Choice of Tenure and Type of Dwelling, *Journal of Urban Economics*, 29, 267-281.

BROWNSTONE, D., ENGLUND, P., PERSSON, M. (1988), A Microsimulation Model of Swedish Housing Demand, *Journal of Urban and Economics*, 23, 179-198.

BUNCH, D.S., BATSELL, R.R. (1989), A Monté Carlo Comparison of Estimators for the Multinomial Logit Model, *Journal of Marketing Research*, XXVI, 56-68.

BUSE, A. (1973), Goodness of Fit in Generalized Least Squares Estimation, *American Statistician*, 27, 106-108.

BUTLER, J.S., MOFFITT, R. (1982), Notes and Comments. A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model, *Econometrica*, 50, 761-764.

CALZOLARI, G., FIORENTINI, G. (1993), Alternative Covariance Estimators of the Standard Tobit Model, *Economics Letters*, 42, 5-13.

CARLINER, G. (1973), Income Elasticity of Housing Demand, *The Review of Economics and Statistics*, LV, 528-532.

CARROLL, S., RELLES, D. (1976), A Bayesian Model of Choice Among Higher Education Institutions, *RAND Corporation Report R-2005-NIE/LE*, Santa Monica, California.

CATALANO, P.J., RYAN, L.M. (1992), Bivariante Latent Variable Models for Clustered Discrete and Continuous Outcomes, *Journal of the American Statistical Association*, 87, 651-658.

CHAPMAN, R.G., STAELIN, R. (1982), Exploiting Rank Ordered Choice Set Data within the Stochastic Utility Model, *Journal of Marketing Research*, XIX, 288-301.

CHESHER, A. (1984), Improving the Efficiency of Probit Estimators, *The Review of Economics and Statistics*, LXVI (notes), 523-527.

CHESHER, A., IRISH, M. (1987), Residual Analysis in the Grouped and Censored Normal Lineal Model, *Journal of Economics*, 34, 33-61.

CHIANG, J., LEE, L-F. (1992), Discrete/Continuous Models of Consumer Demand with Binding Nonnegativity Constraints, *Journal of Econometrics*, 54, 79-93.

CHINTAGUNTA, P.K., JAIN, D.C., VILCASSIM, N.J. (1991), Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data, *Journal of Marketing Research*, XXVIII, 417-428.

CLARK, C. (1961), The Greast of a Finite Set of Random Variables, *Operations Research*, 9, 145-162.

CLARK, W.A., DEURLOO, M.C., DIELEMAN, F.M. (1994), Tenure Changes in the Context of Micro-Level Family and Macro-Level Economic Shifts, *Urban Studies*, 31, 137-154.

CONSIDINE, T.J., MOUNT, T.D. (1984), The Use of Linear Logit Models for Dynamic Input Demand Systems, *The Review of Economics and Statistics*, LXVI, 434-443.

COSSLETT, S.R. (1981a), Efficient Estimation of Discrete Choice Models, en *Structural Analysis of Discrete Data with Econometrics Applications*, Ed. C.F. Manski, D. McFadden, London: MIT Press.

COSSLETT, S.R. (1981b), Maximum Likelihood Estimator for Choice-Based Samples, *Econometrica*, 49, 1289-1316.

COSSLETT, S.R. (1983), Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model, *Econometrica*, 51, 765-782.

COSSLETT, S.R. (1987), Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models, *Econometrica*, 55, 559-585.

COSSLETT, S.R., LEE, L-F. (1985), Serial Correlation in Latent Discrete Variable Models, *Journal of Econometrics*, 27, 79-97.

COX, D.R., WERMUTH, N. (1992), Response Models for Mixed Binary and Qualitative Variables, *Biometrika*, 79, 441-461.

DAGANZO, C. (1979), *Multinomial Probit. The Teory and Its Application to Demand Forecasting,* Academic Press Inc. New York.

DAVIDSON, R., MacKINNON, J.G. (1984), Convenient Specification Test for Logit and Probit Models, *Journal of Econometrics*, 25, 241-262.

De LEEUW, F. (1971), The Demand for Housing: A Review of Cross-Sectional Evidence, *The Review of Economics and Statistics*, LIII, 1-10.

DEATON, A., IRISH, M. (1984), Statistical Models for Zero Expenditures in Household Budgets, *Journal of Public Economics*, 23, 59-80.

DHRYMES, J.P. (1984), Limited Dependent Variables, en *Handbook of Econometrics*, Ed. Z. Griliches, M.D. Intriligator, Vol. III.

DUBIN, J.A., McFADDEN, D.L. (1984), An Econometric Analysis of Residential Electric Appliance Holdings and Consumption, *Econometrica*, 52, 345-362.

DURBIN, J. (1970), Testing for serial Correlation in Least Squares Regression When Some of the Regressors are Lagged Dependent Variables, *Econometrica*, 38, 410-421.

DUSANSKY, R., WILSON, P.W. (1993), The Demand for Housing: Theoretical Considerations, *Journal of Economic Theory*, 61, 120-138.

DYNARSKI, M. (1985), Housing Demand and Desequilibrium, *Journal of Urban and Economics*, 17, 42-57.

EBBELER, D.H. (1975), On the Probability of Correct Model Selection Using the Maximun $\overline{R}^2$ Choice Criterion, *International Economic Review*, 16, 516-520.

EDIN, P.A., ENGLUND, P. (1991), Moving Costs and Housing Demand. Are recent movers really in equilibrium, *Journal of Public Economics*, 44, 299-320.

EFRON, B. (1978), Regression and ANOVA with Zero-One Data: Measures of Residual Variation, *Journal of the American Statistical Association*, 73, 113-121.

ENBERG, J., GOTTSCHALK, P., WOLF, D. (1990), A Random-Effects Logit Model of Work-Welfare Transitions, *Journal of Econometrics*, 43, 63-75.

ENGLE, R.F. (1984), Wald, Likelihood Ratio, and Lagrange Multiplier Test in Econometrics, en *Handbook of Econometrics* Ed. Z. Griliches, M.D. Intriligator, Vol. II.

EVEN, W. E. (1988), Testing Exogeneity in a Probit Model, *Economics Letters*, 26, 125-128.

FEINSTEIN, J., McFADDEN, D. (1989), The Dynamics of Housing Demand by the Elderly: Wealth, Cash Flow, and Demographic Effects, en *Economics of Againg* Ed. D.A. Wise. National Bureau of Economic Research of Chicago Press.

FISHBURN, P.C., FALMAGNE, J-C. (1989), Binary Choice Probabilities and Rankings, *Economics Letters*, 31, 113-117.

FITZMAURICE, G.M., LAIRD, N.M. (1993), A Likelihood-Based Method for Analysing Longitudinal Binary Responses, *Biometrika*, 80, 141-151.

GABRIEL, S.A., ROSENTHAL, S.S. (1989), Household Location and Race: Estimates of a Multinomial Logit Model, *The Review of Economics and Statistics*, 71, 140-249.

GOLDFELD, S.M., QUANDT, R.E. (1972), *Nonlinear Methods in Econometrics,* North-Holland, Amsterdam

GOODMAN, A.C. (1988), An Econometric Model of Housing Price, Permanent Income, Tenure Choice, and Housing Demand, *Journal of Urban and Economics*, 23, 327-353.

GOODMAN, A.C., KAWAI, H. (1982), Permanent Income, Hedonic Prices and Demand for Housing: New Evidence, *Journal of Urban and Economics*,12, 214-237.

GOODMAN, A.C., KAWAI, H. (1984), Estimation and Policy Implications of Rental Housing Demand, *Journal of Urban and Economics*,16, 76-90.

GOULIAS, K.G., KITAMURA, R. (1993), Analysis of Binary Choice Frequencies with Limit Cases: Comparison of Alternative Estimation Methods and Application to Weekly Household Mode Choice, *Transportation Research*, Part B, 27B, 65-78.

GREENE, W.H. (1981a), On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model, *Econometrica*, 49, 505-513.

GREENE, W.H. (1981b), Sample Selection Bias as a Specification Error: Comment, *Econometrica*, 49, 795-798.

GREENE, W.H. (1983), Estimation of Limited Dependent Variable Models by Ordinary Least Squares and the Method of Moments, *Journal of Econometrics*, 21, 195-212.

GREENE, W.H. (1991), *LIMDEP. User's Manual and Reference Guide*, (versión 6.0), Econometric Software, Inc. New York.

GRIZZLE, J.E. (1971), Multivariate Logit Analysis, *Biometrics*, 317 (notes), 1.057-1.062.

GROOTAERT, C., DUBOIS, J.L. (1988), Tenure Choice and the Demand for Rental Housing in the Cities of the Ivory Coast, *Journal of Urban and Economics*, 24, 44-63.

GUILKEY, D.K., MURPHY, J.L. (1993), Estimation and Testing in the Random Effects Probit Model, *Journal of Econometrics*, 59,301-317.

GUILKEY, D.K., SCHMIDT, P. (1979), Some Small Sample Properties of Estimators and Test Statistics in the Multivariate Logit Model, *Journal of Econometrics*, 10, 33-42.

GUMBEL, E.J. (1961), Bivariate Logistic Distributious, *Journal of the American Statistical Association*, 56, 335-349.

GYOURKO, J., LINNEMAN, P. (1990), Rent Controls and Rental Housing Quality: A Note on the Effects of New York City's Old Controls, *Journal of Urban and Economics*, 27, 398-409.

HAINES, H.R., GOODMAN, A.C. (1992), Housing Demand in the United States in the Late Nineteenth Century: Evidence from the Commissioner of the Labor Survey, 1889/1890, *Journal of Urban and Economics*, 31, 99-122.

HALL, B.H. (1984), Software for the Computation of Tobit Model Estimates, *Journal of Econometrics*, 24, 215-222.

HAUSER, J.R. (1978), Testing the Accuracy, Usefulnes, and Significance of Probabilistic Choice Models: An Information-Teoretic Approach, *Operations Research*, 26, 406-421.

HAUSER, J.R. (1986), Agendas and Consumer Choice, *Journal of Marketing Research*, XXIII, 199-212.

HAUSMAN, J.A. (1978), Specification Tests in Econometrics, *Econometrica*, 46, 1251-1271.

HAUSMAN, J.A., HALL, B.H., GRILICHES, Z. (1984), Econometric Models Four Count Data with an Application to the Patents-R & D Relationship, *Econometrica*, 52, 909-938.

HAUSMAN, J.A., LEONARD, G.K., McFADDEN, D. (1995), A Utility-Consistent, Combined Discrete Choice and Count Data Model. Assessing Recreational Use Losses Due to Natural Resource Damage, *Journal of Public Economics*, 56, 1-30.

HAUSMAN, J.A., McFADDEN, D. (1984), Specification Tests for the Multinomial Logit Model, *Econometrica*, 52, 1219-1240.

HAUSMAN, J.A., TAYLOR, W. (1981), A Generalized Specification Test, *Economics Letters*, 6, 239-245.

HAUSMAN, J.A., WISE, D.A. (1978), A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences, *Econometrica*, 46, 403-426.

HAUSMAN, J.A., WISE, D.A. (1980), Discontinuous Budget Contraints and Estimation: The Demand for Housing, *Review of Economic Studies*, XLVII,75-96.

HAUSMAN, J.A., WISE, D.A. (1988), Un Modelo Probit Condicional de Respuesta Cualitativa: Decisiones con Número Restringido de Alternativas Teniendo en Cuenta la Interdependencia y Heterogeneidad de las Preferencias, *Cuadernos Económicos de ICE*, 39, 327-352.

HECKMAN, J.J. (1978), Dummy Endogenous Variables in a Simultaneous Equation System, *Econometrica*, 46, 931-959.

HECKMAN, J.J. (1979), Sample Selection Bias a Specification Error, *Econometrica*, 47, 153-161.

HECKMAN, J.J. (1984), The χ^2 Goodness of Fit Statistic for Models with Parameters Estimated from Microdata, *Econometrica*, 52, 1543-1547.

HECKMAN, J.J. (1990), Varieties of Selection Bias, *American Economic Review*, 80, 313-318.

HENDERSON, J.V., IOANNIDES, Y.M. (1983), A Model of Housing Tenure Choice, *American Economic Review*, 73, 98-113.

HENDERSON, J.V., IOANNIDES, Y.M. (1986), Tenure Choice and the Demand for Housing, *Economica*, 53, 231-246.

HENDERSON, J.V., IOANNIDES, Y.M. (1987), Ower Occupancy: Investment vs Consumtion Demand, *Journal of Urban Economics*, 21, 228-241.

HENDERSON, J.V., IOANNIDES, Y.M. (1989), Dynamic Aspects of Consumer Decisions in Housing Markets, *Journal of Urban Economics*, 26, 212-230.

HONORÉ, B.E. (1993), Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables, *Journal of Econometrics*, 59, 35-61.

HORIOKA, C.Y. (1988), Tenure Choice and Housing Demand in Japan, *Journal of Urban Economics*, 24, 289-309.

HOROWITZ, J.L. (1980), The Accuracy of the Multinomial Logit Model as an Approximation to the Multinomial Probit Model of Travel Demand, *Transportation Research*, 14-B, 331-341.

IMBENS, G.W. (1992), An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling, *Econometrica*, 60, 1187-1214.

JAEN, M., MOLINA, A. (1994), Un Análisis Empírico de la Tenencia y Demanda de Vivienda en Andalucía, *Investigaciones Económicas*, XVIII, 143-164.

JAEN, M., MOLINA, A. (1995), *Modelos Econométricos de Tenencia y Demanda de Vivienda,* Ed. Universidad de Almeria.

JENNRICH, R.I. (1969), Asymptotic Properties of Non-Linear Least Squares Estimators, *Annals of Mathematical Statistics*, 40, 633-643.

JOHNSON, N.L., KOTZ, S. (1969), *Discrete Distributions,* Distributions in Statistics, Wiley, New York.

JOHNSON, N.L., KOTZ, S. (1970a), *Continuous Univariate Distributions-1,* Distributions in Statistics, Wiley, New York.

JOHNSON, N.L., KOTZ, S. (1970b), *Continuous Univariate Distributions-2,* Distributions in Statistics, Wiley, New York.

JOHNSON, N.L., KOTZ, S. (1972), *Continuous Multivariate Distributions,* Distributions in Statistics, Wiley, New York.

KAHN, L.M., MORIMUNE, K. (1979), Unions and Employment Stability: A Sequential Logit Approach, *International Economic Review*, 20, 217-235.

KAIN, J., QUIGLEY, J. (1972), Housing Market Discrimination, Homeownership, and Savings Behavior, *American Economic Review*, 62, 263-277.

KING, M.A. (1980), An Econometric Model of Tenure Choice and Demand for Housing as a Joint Decision, *Journal of Public Economics*, 141, 137-159.

KRUMM, R. (1987), Intertemporal Tenure Choice, *Journal of Urban Economics*, 22, 263-275.

LAITILA, T. (1993), A Pseudo-R^2 Measure for Limited and Qualitative Dependent Variable Models, *Journal of Econometrics*, 56, 341-358.

LANCASTER, T. (1984), The Covariance Matrix of the Information Matrix Test, *Econometrica*, 52 (notes), 1051-1053.

LASHERAS, M.A., SALAS, R., PEREZ, E. (1991), Fiscal Incentives to Home Ownership in Spain: The Effect of Extending Tax Benefits to Second House Purchases, *Instituto de Estudios Fiscales* (Proyecto Provisional).

LAVE, C.A. (1970), The Demand for Urban Mass Trasportation, *The Review of Economics and Statistics*, 37, 320-323.

LEE, L-F. (1979), Identification and Estimation in Binary Choice Models with Limited Dependent Variables, *Econometrica*, 47, 977-996.

LEE, L-F. (1981), Simultaneous Equations Models with Discrete and Censored Dependent Variables, en *Structural Analysis of Discrete Data with Econometric Applications*, Ed. C. Manski, D. McFadden, Cambridge: MIT Press.

LEE, L-F. (1982), Specification Error in Multinomial Logit Models, *Journal of Econometrics*, 20, 197-209.

LEE, L-F., MADDALA, G.S. (1985), The Common Structure of Tests for Selectivity Bias, Serial Correlation, Heteroscedasticity and Non-Normality in the Tobit Model, *International Economic Review*, 26, 1-20.

LEE, L-F., MADDALA, G.S., TROST, R.P. (1980), Asymptotic Covariance Matrices of Two-Stage Probit and Two-Stage Tobit Methods for Simultaneous Equations Models with Selectivity, *Econometrica*, 48, 491-503.

LEE, L-F., TROST, R.P. (1978), Estimation of Some Limited Dependent Variable Models with Application to Housing Demand, *Journal of Econometrics*, 8, 357-382.

LEE, T-H. (1968), Housing and Permanent Income: Tests Based on a Three-year Reinterview Survey, *The Review of Economics and Statistics*, 50, 480-490.

LEE, T-H., KONG, C.M. (1977), Elasticities of Housing Demand, *Southern Economic Journal*, 44, 298-305.

LERMAN, S.R., MANSKI, C.F. (1981), On the Use of Simulated Frequencies to Approximate Choice Probabilities, en *Structural Analysis of Discrete Data with Econometric Applications*, Ed. C.F. Manski, D. McFadden, Cambridge: MIT Press.

LI, M.M. (1977), A Logit Model of Homeownership, *Econometrica*, 45, 1081-1097.

LIEN, D., REARDEN, D. (1990), Missing Measurements in Discrete Response Models, *Economics Letters*, 32, 231-235.

LODHI, A., PASHA, H.A. (1991), Housing Demand in Developing Countries: A Case-Study of Karachi in Pakistan, *Urban Studies*, 28, 623-634.

LOPEZ, M.A. (1992), Algunos Aspectos de la Economía y la Política de la Vivienda, *Investigaciones Económicas*, XVI, 3-41.

LUCE, R.D. (1959), *Individual Choice Behavior. A Theoretical Analysis*, New York: Wiley.

LWIN, T., MARTIN, P.J. (1989), Probits of Mixtures, *Biometrics*, 45, 721-732.

MADDALA, G.S. (1977), Self-Selectivity Problems in Econometric Models, en *Applications of Statistics*, Ed. P.R. Krishnaiah, North-Holland Publishing Company

MADDALA, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, New York.

MADDALA, G.S. (1985), A Survey of the Literature on Selectivity Bias as it Pertains to Health Care Markets, *Advances in Health Economics and Health Services Research*, 6, 3-18.

MAISEL, S.J., BURNHAM, J.B., AUSTIN, J.S. (1971), The Demand for Housing: A Comment, *The Review of Economics and Statistics*, 53, 410-413.

MALHOTRA, N.K. (1984), The Use of Linear Logit Models in Marketing Research, *Journal of Marketing Research*, XXI, 20-31.

MANSKI, C.F. (1975), Maximum Score Estimation of the Stochastic Utility Model of Choice, *Journal of Econometrics*, 3, 205-228.

MANSKI, C.F. (1988), Identification of Binary Response Models, *Journal of the American Statistical Association*, 83, 729-738.

MANSKI, C.F., LERMAN, S.R. (1977), The Estimation of Choice Probabilities from Choice Based Samples, *Econometrica*, 45, 1977-1988.

MANSKI, C.F., McFADDEN, D. (1981), Alternative Estimators and Sample Designs for Discrete Choice Analysis, en *Structural Analysis of Discrete Data with Econometric Applications,* Ed. C.F. Manski, D. McFadden, Cambridge: MIT Press.

MANTEL, N., BROWN, C. (1973), A Logistic Reanalysis of Asford and Sowden's Data on Respiratory Symptoms in British Coal Miners, *Biometrics*, 29, 649-665.

MAYO, S.K. (1981), Theory and Estimation in the Economics of Housing Demand, *Journal of Urban Economics*, 10, 95-116.

McFADDEN, D. (1973), Conditional Logit Analysis of Qualitative Choice Behavior, en *Frontiers in Econometrics*, Ed. P. Zarembka, New York: Academic Press.

McFADDEN, D. (1974), The Measurement of Urban Travel Demand, *Journal of Public Economics*, 3, 303-328.

McFADDEN, D. (1976), Quantal Choice Analysis: A Survey, *Annals of Economic and Social Measurement*, 5, 363-390.

McFADDEN, D. (1978), Modelling the Choice of Residential Location, en *Spatial Interaction Theory and Residential Location*, Ed. A. Karlgvist et al. North Holland: Amsterdam.

McFADDEN, D. (1980), Econometric Models for Probabilistic Choice among Products, *Journal of Business*, 53, 513-529.

McFADDEN, D. (1981), Econometric Models of Probabilistic Choice, en *Structural Analysis of Discrete Data with Econometric Applications*, Ed. C.F. Manski, D. McFadden, Cambridge: MIT Press.

McFADDEN, D. (1988), El Análisis Econométrico de los Modelos de Respuesta Cualitativa, *Cuadernos Económicos de I.C.E.*, 39, 247-305.

McFADDEN, D. (1989), A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration, *Econometrica*, 57, 995-1026.

MINISTERIO DE OBRAS PÚBLICAS TRANSPORTE Y MEDIO AMBIENTE (1994), *Precio medio del m^2 de las viviendas: datos obtenidos de las tasaciones hipotecarias 1987-1993.*

MOORE, W.L., LEHMANN, D.R. (1989), A Paired Comparison Nested Logit Model of Individual Preference Structures, *Journal of Marketing Research*, XXVI, 420-428.

MORIIZUMI, Y., TAKAGI, S. (1983), The Income Elasticity of Housing Demand in Japan, *Economic Studies Quarterly*, 34, 70-86.

MORIMUNE, K. (1979), Comparisons of Normal and Logistic Models in the Bivariate Dichotomous Analysis, *Econometrica*, 47, 957-976.

MUTH, R.F. (1960), The Demand for Non-Farm Housing, en *The Demand for Durable Goods*, Ed. A.C. Harberger, Chicago.

MUTH, R.F. (1969), *Cities and Housing*, The University of Chicago Press.

NAKAGAHI, Y., PEREIRA, A.M. (1991), Housing Appreciation, Mortgage Interest Rates, and Homeowner Mobility, *Journal of Urban Economics*, 30, 271-292.

NELSON, F.D., OLSON, L. (1978), Specification and Estimation of Simultaneous Equation with Limited Dependent Variables, *International Economic Review*, 19, 685-710.

NEWEY, W.K. (1987), Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables, *Journal of Econometrics*, 36, 231-250.

PETERS, H., WAKKER, P. (1991), Independence of Irrelevant Alternatives and Revealed Group Preferences, *Econometrica*, 59, 1787-1801.

PRESS, S.J., WILSON, S. (1978), Choosing Between Logistic Regression and Discriminant Analysis, *Journal of the American Statistical Association*, 73, 699-705.

QUANDT, R.E. (1974), A Comparison of Methods for Testing Nonnested Hypotheses, *The Review of Economics and Statistics*, 56, 92-99.

RAO, C.R. (1973), *Linear Statistical Inference and Its Application*, Wiley, New York.

REID, M. (1962), *Housing and Income*, Chicago: Chicago University Press.

ROSEN. H.S. (1979), Housing Decisions and the U. S. Income Tax: An Econometric Analysis, *Journal of Public Economics*, 11, 1-23.

ROSENTHAL, S.S. (1988), A Residence Time Model of Housing Markets, *Journal of Public Economics*, 36, 87-109.

RUUD, P.A. (1983), Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models, *Econometrica*, 51, 225-228.

RUUD, P.A. (1986), Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution, *Journal of Econometrics*, 32, 157-187.

SAWA, T. (1978), Information Criteria for Discriminating among Alternative Regression Models, *Econometrica*, 46, 1273-1291.

SHEFFI, Y. (1979), Estimating Choice Probabilities Among Nested Alternatives, *Transportation Research*, 13B, 189-205.

SMALL, K.A. (1987), A Discrete Choice Model for Ordered Alternatives, *Econometrica*, 55, 409-424.

SMALL, K.A., ROSEN, H.S. (1981), Applied Welfare Economics with Discrete Choice Models, *Econometrica*, 49, 105-130.

SMITH, R.J. (1989), On the Use of Distributional Mis-Specification Checks in Limited Dependent Variable Models, *The Economic Journal*, 99, 178-192.

SOOFI, E.S. (1992), A Generalizable Formulation of Conditional Logit with Diagnostics, *Journal of the American Statistical Association*, 87, 812-816.

SRINIVASAN, V. (1980), Comments on On Conjoint Analysis and Quantal Choice Models, *Journal of Business*, 53, 547-550.

STERN, S. (1989), Rules of Thumb for Comparing Multinomial Logit and Multinomial Probit Coefficients, *Economics Letters*, 31, 235-238.

STOKER, T.M. (1989), Test of Additive Derivative Constraints, *Review of Economic Studies*, 56, 535-552.

TAUCHEN, G. (1985), Diagnostic Testing and Evaluation of Maximum Likelihood Models, *Journal of Econometrics*, 28, 415-443.

THEIL, H. (1969), A Multinomial Extension of the Linear Logit Model, *International Economic Review*, 10, 251-259.

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

TRAIN, K. (1980), A Structured Logit Model of Auto Ownership and Mode Choice, *Review of Economic Studies*, XLVII, 357-370.

TRAIN, K. (1986), *Qualitative Choice Analysis: Theory Econometrics and an Application to Automobile Demand*, MIT Press Cambridge-Massachusetts.

TSE, Y.K. (1988), A Proportional Random Utility Approach to Qualitative Response Models, *Journal of Business & Economic Statistics*, 7, 61-65.

TVERSKY, A. (1972a), Elimination by Aspects: A Theory of Choice, *Psychological Review*, 79, 281-299.

TVERSKY, A. (1972b), Choice by Elimination, *Journal of Mathematical Psychology*, 9, 341-367.

VEALL, M.R., ZIMMERMANN, K.F. (1992a), Pseudo-R^2's in the Ordinal Probit Model, *Journal of Mathematical Sociology*, 16, 333-342.

VEALL, M.R., ZIMMERMANN, K.F. (1992b), Performance Measures from Predicion-Realization Tables, *Economics Letters*, 39, 129-134.

VITON, P.A. (1985), On the Interpretation of Income Variables in Discrete-Choice Models, *Economics Letters*, 17, 203-206.

WALES, T.J., WOODLAND, A.D. (1983), Estimation of Consumer Demand Systems with Binding Non-Negativity Constraints, *Journal of Econometrics*, 21, 263-285.

WEISS, A.A. (1993), Some Aspects of Measurement Error in a Censored Regression Model, *Journal of Econometrics*, 56, 169-188.

WESTIN, R.B. (1974), Predictions from Binary Choice Model, *Journal of Econometrics*, 2, 1-16.

WINDHEIJER, F.A.G. (1995a), Goodness-of-Fit Measures in Binary Choice Models, *Econometric Reviews*, 14, 101-116.

WINDHEIJER, F.A.G. (1995b), Comments on Goodness-of-Fit Measures in Binary Choice Models, *Econometric Reviews*, 14, 117-120.

WINGER, A.R. (1968), Housing and Income, *Western Economic Journal*, 6, 226-232.

WU, D-M. (1965), An Empirical Analysis of Household Durable Goods Expenditure, *Econometrica*, 33, 761-780.

ZORN, P.M. (1988), An Analysis of Household Mobility an Tenure Choice: An Empirical Study of Korea, *Journal of Urban and Economics*, 24, 113-128.

ZORN, P.M. (1993), The Impact of Mortage Qualification Criteria on Households' Housing Decisions: An Empirical Analysis Using Microeconomic Data, *Journal of Housing Economics*, 3, 51-75.