[**Capítulo 19. MODELOS DE PROBABILIDAD** 713](#_Toc301776368)

[19.1. Características 713](#_Toc301776369)

[19.2. Modelo Lineal de Probabilidad 714](#_Toc301776370)

[19.3 Modelos de Probabilidad con distribución acumulativa 716](#_Toc301776371)

[19.4 Modelo Logit 719](#_Toc301776372)

[Estimación del modelo 723](#_Toc301776373)

[*Estimación con datos agrupados* 724](#_Toc301776374)

[*Estimación con datos individuales* 727](#_Toc301776375)

[19.5 Modelo Probit 728](#_Toc301776376)

[19.6 Modelo Tobit 732](#_Toc301776377)

[19.7 Evaluación de los modelos de probabilidad 733](#_Toc301776378)

[**CASOS DE ESTUDIO, PREGUNTAS Y PROBLEMAS** 734](#_Toc301776379)

[Caso 19.1: Vulnerabilidad social en los hogares de Río Cuarto 734](#_Toc301776380)

[**BIBLIOGRAFIA** 736](#_Toc301776381)

### **Capítulo 19. MODELOS DE PROBABILIDAD**

#### 19.1. Características

En el modelo mínimo cuadrático



* Las variables **X** pueden cuantitativas o cualitativas
* La variable **Y** sólo puede ser cuantiativa

En la práctica, en muchas ocasiones, se tiene la necesidad de explicar la ocurrencia de un evento. Esto se representa a través de una variable binaria que asume el valor 1, si el evento sucede, y el valor 0, si no sucede; es decir, produce una respuesta de sí o no, de acuerdo a la presencia de la cualidad que se quiera medir. Este tipo de situaciones se estudian a partir de los Modelos de Probabilidad.

La variable dependiente binaria también se denomina variable dependiente limitada o variable indicadora. Estas variables tienen restringido su rango de variabilidad y muchas conducen a soluciones de esquina, esto significa que resulta óptimo elegir el 0 para un número importante de situaciones.

**Ejemplo**. Supongamos que se desea estudiar la participación de la fuerza laboral de hombres adultos en función de la tasa de desempleo, de la tasa de salarios promedio, del ingreso familiar, de la educación, etc. Una persona o bien está en la fuerza laboral o no está. Por tanto, la variable dependiente que es la participación en la fuerza laboral, solamente puede adquirir dos valores: 1 si la persona está en la fuerza laboral y 0 si no lo está.

Se considerarán los cuatro enfoques de mayor difusión

* Modelo lineal de probabilidad (MPL)
* Modelo Logit
* Modelo Probit
* Modelo Tobit

#### 19.2. Modelo Lineal de Probabilidad

En la década del ´60 y principios de la siguiente, el modelo de probabilidad lineal (MPL) se utilizó ampliamente y la estimación se realizaba a partir del mínimo cuadrado ordinario. En estos modelos, la variable **Y** es indicadora de ocurrencia o no ocurrencia de un evento.

Dado



El modelo se especifica

 (1)

donde,

 es una variable cualitativa

 son variables explicativas de carácter cuantitativo o cualitativo

Si , entonces la esperanza condicional

 (2)

Debe interpretarse como la probabilidad de que ocurra el evento, dado que ocurren las ; es decir,



Modelos de este tipo, que expresan la variable binaria como una función lineal de la o las variables independientes, se denominan Modelo de Probabilidad Lineal puesto que la , puede ser interpretada como la *probabilidad condicional* de que el evento suceda dado .

**Ejemplo**. En un estudio en los hogares, si Y mide la posesión de una casa -*Y*=1 cuando la familia posee casa y *Y=0 cuando* no posee- y X el nivel de ingresos. La da la probabilidad de que una familia posea una casa dado que tiene un ingreso de cierta cantidad 

Haciendo,  -que el evento ocurra- y  - que el evento no ocurra-, la distribución de probabilidad de la variable *Y* (que sigue la distribución de probabilidad Bernoulli[[1]](#footnote-1)) viene dada por:

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Probabilidad* |  |  |
| 0 |  |  |  |
| 1 |  |  | (3) |
| Total | 1 |  |  |

Por consiguiente, por definición de esperanza matemática se obtiene



comparando con , se puede igualar



es decir, la esperanza condicional del modelo puede ser interpretada, de hecho, como la probabilidad condicional de 

Puesto que la probabilidad  debe encontrarse entre 0 y 1, se tiene la restricción

 (4)

es decir, la esperanza condicional o probabilidad condicional debe encontrarse entre 0 y 1.

En aplicaciones prácticas el MPL, especificado en (1), presenta problemas tales como,

* No normalidad de los . Al igual que la variable Y, los errores siguen una distribución Bernoulli:



|  |  |  |
| --- | --- | --- |
| Y | e | Probabilidad |
| 1 |  |  |
| 0 |  |  |

Este problema puede obviarse si se trabaja con un alto número de observaciones, porque se sabe que si n tiende a infinito, los estimadores mínimoS cuadráticos tienden a estar normalmente distribuidos.

* Heterocedasticidad de . En la distribución Bernoulli, la media es p y la varianza es p(1-p); lo cual revela que la varianza es función de la media, por lo que la varianza del error es heterocedástica.

En el MLP, si  y la varianza es P(1-P), quiere decir que la varianza de  es heterocedástica porque depende de los valores de X.

Una forma de resolver el modelo es transformar (1) dividiendo ambos miembros por ; es decir:



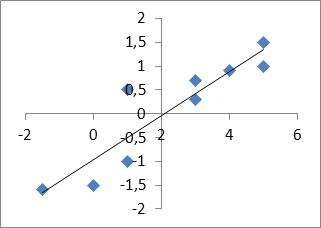
El término de error es homocedástico; la estimación se realiza mediante mínimos cuadrados ponderados, lo cual significa aplicar mínimos cuadrados ordinarios a un modelo ponderado por W.

* La posibilidad de que  se encuentre fuera del rango 0 – 1. Aquí radica el problema real en la estimación mínimo cuadrática del MPL. En la práctica se resuelve de la siguiente manera

Si  se asume que 

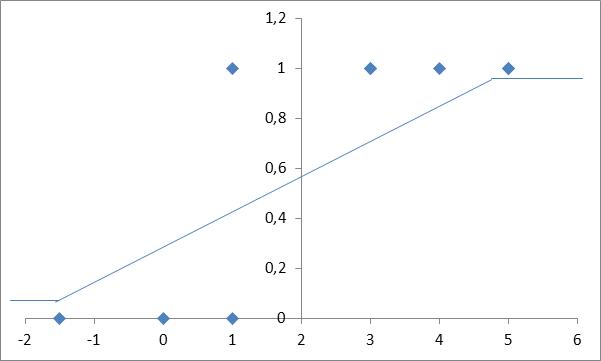
Si  se asume que 

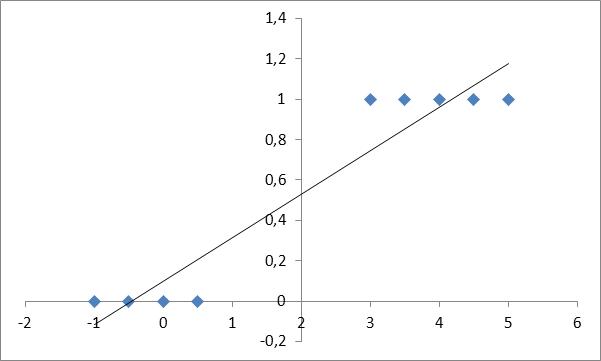
* Valores generalmente bajos de . El  no es considerado como medida de bondad de ajuste porque los valores observados de Y serán 0 o 1; mientras que, los valores estimados abarcan todo el recorrido de los números reales. Esto significa que:





En la práctica, se realiza una estimación restringida o truncada:





Si  se asume que 

Si  se asume que 

El valor de  no puede indicar la bondad de la estimación porque habrá un grupo de observaciones que asumirán un valor estimado que no es el correcto.

Cuando la dispersión de las observaciones está muy concentrado en dos puntos, puede esperarse un  cercano al 0,8. En general, se aconseja no tener en cuenta el  como medida de bondad de ajuste

Aldrich y Nelson sostienen que *el uso del coeficiente de determinación como estadístico resumen debe evitarse en modelos con variable dependiente cualitativa.*

Aún salvando estos problemas el MPL no es un modelo muy atractivo porque supone que aumenta linealmente con *X*, es decir el efecto marginal o incremental de *X* permanece constante todo el tiempo. En realidad se esperaría que  estuviera relacionado en forma no lineal con 

**Ejemplo**. Si se aplica el modelo de propiedad de la vivienda y se encuentra que  significaría que a medida que aumenta una unidad (supongamos en miles de pesos), la probabilidad de ser propietario de una vivienda aumenta en la misma cantidad constante de 0.10. Esto es así para niveles de ingreso de $ 8.000; $15.000 o $50.000. Esto no parece ser realista.

Para ingresos muy bajos una familia no poseerá una casa, pero a un nivel de ingresos suficientemente alto, por ejemplo $a es muy probable que ésta sí posea una casa. Cualquier aumento en el ingreso más allá de $a tendrá un efecto pequeño sobre la probabilidad de poseer una casa. Así a ambos extremos de la distribución de ingresos, la probabilidad de poseer una casa no se verá afectada, virtualmente, por un pequeño incremento en .

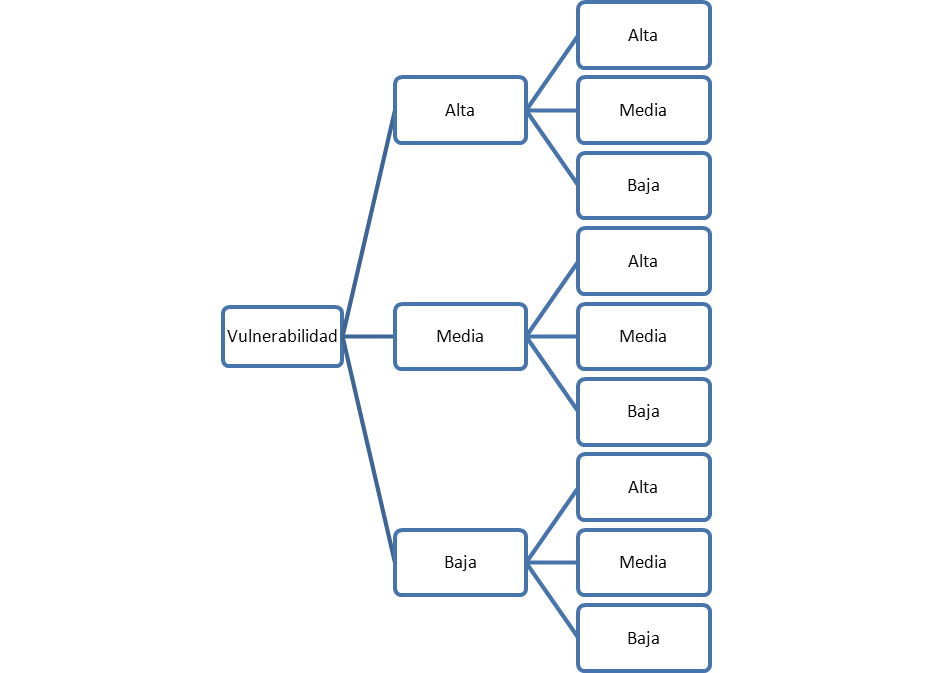
En síntesis, al modelo especificado en (1) se lo denomina lineal de probabilidad porque la probabilidad de respuesta es lineal en los parámetros β. A la expresión se la denomina probabilidad de respuesta e indica que las probabilidades (tanto de ocurrencia como de no ocurrencia) dependen de los valores de X.

Para solucionar estos problemas se presentan a continuación los modelos Logit y Probit (Normit).

**Ejemplo:** Estudio sobre la vulnerabilidad en la ciudad de Río Cuarto

Se cuenta con información para 1549 hogares de la ciudad de Río Cuarto con hijos nacidos durante 2005.

A partir de un estudio exploratorio (AFCM) se identificaron nueve grupos sociales que se clasificaron en



Esta partición permitió la construcción de la variable Alta Vulnerabilidad (ALTAVUL) de la siguiente manera:

El objetivo es identificar los fenómenos de mayor impacto en la determinación de la alta vulnerabilidad en la ciudad de Río Cuarto.

El modelo se especifica

Donde

MAD indica la situación de la madre según su edad

NBI indica las necesidades básicas insatisfechas que tiene el hogar observado

EDAD edad actual de la madre

AI número de personas que aportan ingresos en el hogar

IM ingreso mensual del hogar en pesos de 2005

IPC ingreso per cápita diario del hogar en pesos de 2005

La estimación del modelo se realiza por mínimo cuadrados ordinarios, luego se le dan valores de referencia a las variables explicativas y el valor indicará la probabilidad de pertenecer a un grupo social de alta vulnerabilidad.

La estimación del ejemplo no está exenta de lo que se mencionó anteriormente: no normalidad, heterocedasticidad y poca relevancia de . Frente a esto, lo que se necesita es un modelo probabilístico que tenga:

1. Probabilidad acotada al intervalo 0-1
2. Probabilidad creciente conforme aumentan los valores de X
3. Relación no lineal entre la probabilidad y X, de modo que al acercarse a 1 o 0 lo haga a tasa decreciente.

Las funciones que tienen estas características son las distribuciones acumulativas en forma de S. Aunque hay casos que cumplen con estos requisitos, las más usadas son la logística y la normal, que dan lugar a los modelos de probabilidad Logit y Probit o Normit.

#### 19.3 Modelos de Probabilidad con distribución acumulativa

Estos modelos se especifican de la siguiente manera

Donde es una variable latente que se define como la proyección, capacidad o posibilidad de que el evento bajo estudio ocurra.

La variable latente no se puede observar, por lo que debe aproximarse a través de una variable relacionada que sí se pueda observar. Esta variable es la que indica la presencia o ausencia del evento en una población.

**Ejemplo:** Estudio sobre la vulnerabilidad en la ciudad de Río Cuarto

La variable observable es el individuo que presenta situación de vulnerabilidad, la variable no observable es la posibilidad de que el individuo esté en una situación de alta vulnerabilidad.

El método de estimación de estos modelos es el de máxima verosimilitud (MV). La verosimilitud de la iésima observación es la probabilidad de que Y tome el valor que se observa que toma, de modo que la verosimilitud de cada observación depende del valor de los parámetros β y de si el evento se presenta o no.

Si se tienen n observaciones idénticas e independientemente distribuidas que siguen el modelo

La función de verosimilitud es la probabilidad de que tome el valor que se observa para cada elemento de la muestra condicionado a los valores de X.

La función de verosimilitud puede escribirse como

Lo cual implica que la probabilidad de que un grupo de observaciones tome efectivamente los valores observados de Y es igual al producto de las probabilidades de cada observación. De modo que:

Con

El estimador de máxima verosimilitud de β es el que maximiza el logaritmo de la función de verosimilitud

Que será un estimador consistente, asintóticamente normal y asintóticamente eficiente.

Las condiciones de primer orden serán

Al ser un problema no lineal, su resolución se alcanza al aplicar métodos iterativos y el estimador se obtiene por métodos numéricos iterativos.

En definitiva, el principio de máxima verosimilitud consiste en escoger valores estimados de los parámetros β que maximicen la función de verosimilitud.

Los estimadores de máxima verosimilitud son los valores por los cuales la probabilidad de “observar la muestra que se observa” es tan grande como sea posible.

Los modelos logit y probit difieren en la especificación del término de error

Donde F es la función de distribución acumulada de

Si la distribución de es simétrica, haciendo

Es posible escribir

La forma funcional de F dependerá de la suposición en torno al término ε; si la distribución que acumula ε es logística, se tiene el logit, mientras que, si es normal se tiene el probit (o normit).

Los modelos logit y probit son modelos de respuesta binaria

La toma valores entre 0 y 1.

Si es logística

Si es logística

Da lugar al modelo logit cuya expresión será

Da lugar al modelo logit cuya expresión será

Si es normal

Donde

Se tiene el modelo probit, cuya expresión será

#### 19.4 Modelo Logit

El modelo logit supone que una variable dependiente ,

Donde se tiene una variable latente , y se distribuye como una logística.

Denominando

Donde X son las variables explicativas e Y las variables dicótomas; si Y=1 significa que el evento ocurre, si Y=0 el evento no ocurre.

es la probabilidad de que el evento se presente se lo conoce como función de distribución logística acumulativa

Si , la

Cuando , y cuando ,

no se relaciona linealmente con X por lo que se satisfacen los requerimientos expuestos anteriormente.

Esta función se comporta diferente al modelo lineal de probabilidad

Si el evento se presenta

Si no se presenta

La razón de probabilidades a favor del evento bajo estudio es

Tomando logaritmo de la razón de probabilidades

A L se lo denomina logit; si es positivo la probabilidad tiende a 1, si es negativo tiende a 0. L se relaciona linealmente con las variables X y los coeficientes β; mientras que, la probabilidad guarda una relación no lineal con las variables X y los coeficientes β.

Para estimar un modelo logit, se debe especificar el modelo

Donde el evento se presenta,

el evento no se presenta,

logística

La estimación da por resultado

indica el cambio marginal que un cambio en X provoca en L, pero no indica en cuanto variará la probabilidad. Esta situación se deriva de las relaciones no lineales planteadas anteriormente y da lugar a que se deba definir una situación de referencia (o situación promedio) para encontrar el valor de Z.

Esto se logra haciendo que:

1. Las variables X cuantitativas adopten el valor promedio
2. Las variables X cualitativas adopten alternativamente el valor 0 y el 1

Con el valor de Z se calcula la probabilidad de la siguiente manera:

aplicando antilogaritmo

La incógnita es la probabilidad, que se resuelve como

Los cambios en X impactan en L en la cuantía β, como se dijo anteriormente, y en la probabilidad de acuerdo a

**Ejemplo:** Propiedad de la vivienda (ejemplo desarrollado en Gujarati (2004))

Ahora se considerará la siguiente especificación





donde  es el ingreso

 significa que la familia es propietaria de una casa

 representa la probabilidad de que el evento se presente.

Por simplicidad, se escribe la ecuación de la siguiente manera:

; donde 

Esta última ecuación es la función de distribución logística (acumulativa).

Es fácil verificar que mientras  se encuentra dentro de un rango de  a ,  se encuentra dentro de un rango de 0 a 1 y que  no está linealmente relacionada con  (es decir, con ), satisfaciendo así los dos requerimientos considerados anteriormente. Pero se crea un problema de estimación porque  no es solamente no lineal en las  sino también en los , como puede verse a partir de la ecuación del modelo.

Esto significaría que no se puede utilizar MCO para estimar los parámetros. Pero, este problema es más aparente que real ya que el modelo es intrínsecamente lineal, lo cual puede verse de la siguiente manera.

Si el evento se presenta 

Si el evento no se presenta la probabilidad es 

La razón de probabilidad a favor del evento bajo estudio se expresa como



La razón de probabilidades a favor de poseer una casa se interpreta como la probabilidad de que una familia posea una casa a la probabilidad de que no la posea. Así si  significa que las probabilidades son 4 a 1 a favor de la familia que posee una casa.

Tomando logaritmo natural de la razón de probabilidades



L es denominado logit, es lineal en  y lineal en los parámetros:

* Si  a valores crecientes de , se incrementa la probabilidad de ocurrencia del evento
* Si  a valores crecientes de , disminuye la probabilidad de ocurrencia del evento

A continuación se enuncian los aspectos que caracterizan al modelo Logit:

1. A medida que  va de 0 a 1 (cuando *Z* varía de  a ) el Logit *L* va de  a . Es decir, aunque las probabilidades (por necesidad) se encuentran entre 0 y 1, los Logit no están limitados en esa forma.
2. Aunque *L* es lineal en *X*, las probabilidades en sí mismas no lo son. Esta propiedad hace contraste con el MLP en donde las probabilidades aumentan linealmente con *X*.
3. Utilizando el cálculo, puede demostrarse que , lo cual muestra que la tasa de cambio en la probabilidad con respecto a *X* contiene no solamente a  sino también al nivel de probabilidad a partir del cual se mide el cambio. A propósito, obsérvese que un cambio unitario en  sobre  es máximo cuando  y mínimo cuando  está cercano a 0 o a 1.
4. La interpretación del modelo LOGIT es la siguiente: , la pendiente, mide el cambio en *L* ocasionado por un cambio unitario en . En el ejemplo, como el logaritmo de las probabilidades a favor de poseer una casa cambia a medida que el ingreso cambia en una unidad (supongamos, $1.000). El intercepto  es el valor del logaritmo de las probabilidades a favor de poseer una casa si el ingreso es cero.
5. Dado un nivel de ingreso determinado, por ejemplo $a, si realmente se desea estimar la probabilidad misma de poseer una casa, *y no las probabilidades a favor de poseer una casa*, esto puede hacerse directamente a partir de la primera ecuación una vez de que se disponga de las estimaciones de  y .
6. Mientras que el MLP supone que  está relacionado linealmente con , el modelo LOGIT supone que el logaritmo de la razón de probabilidades está relacionado linealmente con .

##### Estimación del modelo

A fines de la estimación el modelo se especifica



Para estimar el modelo, además de los valores de , se necesitan los valores del logit  pero se incurre en algunas dificultades. En el caso del ejemplo (y en otros similares) si existe información disponible sobre familias individuales, entonces  si una familia posee una casa y  si una familia no la posee. Pero si se colocan estos valores directamente en el logit  se obtiene

 si una familia posee una casa

 si una familia no posee una casa.

Ambas expresiones carecen de sentido.

Por consiguiente, si la información disponible está a nivel micro o individual, no se puede estimar el modelo mediante la rutina del método de mínimos cuadrados ordinarios. En esta situación se recurre a máxima verosimilitud. Pero es posible estimar por MCO si se tiene datos agrupados

###### *Estimación con datos agrupados*

La información se agrupa siguiendo algún criterio y se estima por mínimos cuadrados ordinarios. En el ejemplo de familias propietarias de viviendas, la información puede agruparse según el nivel de ingresos.

|  |  |  |  |
| --- | --- | --- | --- |
| Obs | ingreso (miles de $) | número de familias con ingreso | número de familias que poseen casa |
| 1 | 6 | 40 | 8 |
| 2 | 8 | 50 | 12 |
| 3 | 10 | 60 | 18 |
| 4 | 13 | 80 | 28 |
| 5 | 15 | 100 | 45 |
| 6 | 20 | 70 | 36 |
| 7 | 25 | 65 | 39 |
| 8 | 30 | 50 | 33 |
| 9 | 35 | 40 | 30 |
| 10 | 40 | 25 | 20 |

Luego se deben seguir los siguientes pasos:

1. Para cada nivel de ingreso  calcular la probabilidad estimada de poseer una casa como ; es decir *la frecuencia relativa*. Se puede utilizar ésta como una estimación del verdadero  correspondiente a cada . Si  es relativamente grande,  será una estimación razonablemente buena de  (de la estadística elemental recuerde que la probabilidad de un evento es el límite de la frecuencia relativa a medida que el tamaño de la muestra se hace infinitamente grande).
2. Utilizando  estimado, se puede obtener el Logit estimado como



1. Por lo tanto, dada la información *agrupada o replicada* (observaciones repetidas), se puede obtener información sobre la variable dependiente, los Logit.
2. La interpretación se realiza de la siguiente manera:



al tomar antilogaritmo a esta expresión se obtiene la razón de probabilidades 

Pero  esto significa que 

Al resultado de evaluar  se le resta 1 y se lo multiplica por 100, este resultado es el cambio porcentual a favor de la ocurrencia del evento ante el cambio en algún regresor.

1. Si se quiere calcular la probabilidad  se debe hacer



1. Puede demostrarse que si  es relativamente grande y cada observación en una clase de ingreso dado  está distribuida en forma independiente como una variable binomial, entonces



por consiguiente, como en el caso del MLP, el término de perturbación es heterocedástico y habrá que utilizar MCP. En esta situación se usará la siguiente transformación del modelo



que se escribe como: 

donde  es el coeficiente que pondera

 es igual a  ponderada

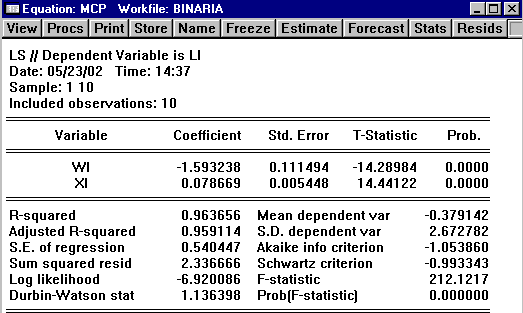
 es igual a  ponderada

 es el término de error ponderado homocedástico

1. Estímese la ecuación transformada mediante MCO. Tenga en cuenta que en esta ecuación no hay término de intercepto introducido explícitamente, por lo que se tendrá que usar el procedimiento de regresión a través del origen.
2. Establézcanse intervalos de confianza y/o pruebas de hipótesis dentro de lo usual para MCO, cuyos resultados serán válidos sólo para muestras relativamente grandes.

Aunque paquetes tales como el EVIEWS estiman directamente estos modelos, apliquemos el razonamiento anterior para comprender algunos resultados.

La estimación por MCP sin ordenada al origen da:



Esto es,



El intercepto estimado es 

Como muestra esta regresión, el coeficiente de pendiente estimado sugiere que para un incremento unitario ($1.000) en el ingreso ponderado, el logaritmo ponderado de las probabilidades a favor de poseer una casa aumenta en alrededor de 0.08.

Tomando antilogaritmo de 0.0787, se obtiene aproximadamente 1.0818, lo cual significa que para un incremento unitario en los ingresos ponderados, las probabilidades ponderadas a favor de poseer una casa aumentan en 1.0818 o alrededor de 8.18%.

¿Se puede calcular la probabilidad de poseer una casa, dado el ingreso, a partir de la razón de probabilidades?

Este cálculo puede hacerse fácilmente. Supóngase que se desea estimar la probabilidad de poseer una casa para el nivel de ingreso de $20.000. Se tiene el dato observado en la Tabla 16.1 para la observación 6, donde ,  y , y la estimación que surge de la Tabla 16.1.

 (1)

Se debe tener en cuenta que

 (2)



Si  entonces  (3)

Reemplazando (2) y (3) en (1) 

Pero , por lo que .

Ahora bien, 

Tomando el antilogaritmo de 

 (4)

Es decir,  (5).

Igualando (4) y (5) se obtiene



Es decir, la probabilidad de que una familia con un ingreso de $20.000 posea una casa es de alrededor de 0,50.

###### *Estimación con datos individuales*

Dado el modelo



El modelo estimado con datos individuales por Logit será



Este resultado se interpreta de la siguiente manera. En primer lugar debe tenerse en cuenta que



Al tomar antilogaritmo en la expresión anterior, tendremos la razón de probabilidades:



Para encontrar la probabilidad de ocurrencia del evento dado que las variables explicativas se comportan de una manera determinada, se procede de la siguiente manera:



Habitualmente, para encontrar el valor de  se le asigna a las variables cuantitativas el valor medio; con las cualitativas se trabaja asignando el valor 1 o el valor 0, de acuerdo a que se quiera encontrar la probabilidad del evento cuando la cualitativa está presente o ausente.

La tasa marginal de cambio en la probabilidad de ocurrencia del evento ante cambios en las variables explicativas cuantitativas viene dado por:



#### 19.5 Modelo Probit

Como se ha mencionado, para explicar el comportamiento de una variable dependiente binaria es preciso usar una función de distribución acumulada seleccionada apropiadamente.

Para el caso del modelo Logit se usó la función logística acumulativa. La función de distribución acumulada normal también brinda utilidad a estos efectos dando lugar a lo que se conoce como *modelo Probit o Normit*.

El modelo probit supone que una variable dependiente Y=1 si una variable latente o indicadora viene determinada por

Con que se distribuye normal de media 0 y desvío 1.

Cuanto más alto sea el valor de , mayor será la probabilidad de que ocurra el evento

Es decir, la función de distribución acumulativa que permite explicar el comportamiento de una variable dicótoma es la normal. El valor de Z se obtiene de igual manera que en el logit, se consideran los valores promedios de las variables cuantitativas y, alternativamente, los valores 1 o 0 en las variables cualitativas lo que permite definir una situación de referencia.

**Ejemplo:** Propiedad de la vivienda (ejemplo desarrollado en Gujarati (2004))

Supóngase que la decisión de la *i–ésima* familia de poseer una casa o de no poseerla depende de un *índice de conveniencia no observable* , que está determinado por una o varias variables explicativas, por ejemplo, el ingreso , de tal manera que cuando mayor sea el valor del índice, mayor será la probabilidad de que la familia posea vivienda.

De esta manera:



¿Cómo se relaciona el  no observable con la decisión de poseer una casa?

Igual que antes, sea  si la familia posee una casa e  si no la posee.

Ahora bien, es razonable suponer que para cada familia hay un nivel crítico o umbral del índice, que se puede denominar , tal que si



la familia poseerá una casa, de lo contrario no lo hará.

Si se supone que el índice y el umbral se distribuyen normales con igual media y varianza, será posible estimar los parámetros del modelo y obtener alguna información adicional.

Dado el supuesto de normalidad, la probabilidad de que  sea menor o igual que  puede ser calculada a partir de la FDA normal estandarizada como



donde *t* es una variable normal estandarizada, es decir *t* ~ *N(0,1)*.

Ahora, para estimar el modelo, deberá tenerse en cuenta que:



Al igual que el logaritmo de la razón de probabilidades (en el modelo Logit), la inversa de la FDA normal sirve para hacer lineal al modelo Probit.

De esta forma el modelo a estimar resulta de conocer las probabilidades, en este caso



Luego, aplicar MCP.

Notas:

* La perturbación , al igual que antes, es heterocedástica. Se puede demostrar que su varianza está dada por  donde  es la función de densidad normal estándar evaluada en . Por lo tanto, habrá que ponderar el modelo para aplicar mco.
* La variable no observable es conocida como desviación equivalente normal o simplemente normit. Puesto que normit será negativo siempre que , en la práctica se agrega el número 5 al normit y el resultado se denomina probit.
* Los dos métodos presentados son bastante similares, generalmente por conveniencia matemática se prefiere el logit. Pero como lo sugieren algunos autores una estimación logit de un parámetro multiplicada por 0.625 proporciona una aproximación relativamente buena de la estimación probit del mismo parámetro.
* También, se puede demostrar que , excepto para el intercepto que es .
* Todas las aproximaciones anteriores funcionan bien cuando el valor promedio de la probabilidad de que suceda el evento no este lejana de 0.5.
* Se debe tener cuidado al interpretar el coeficiente de pendiente. En el mpl el coeficiente de pendiente mide directamente el cambio en la probabilidad de que ocurra un evento como resultado de un cambio unitario en el valor del regresor. En el logit la tasa de cambio en la probabilidad está dada por  donde  es el coeficiente del *j–ésimo* regresor. En el probit, la tasa de cambio en la probabilidad es algo complicada y está dada por , donde  es la función de densidad de la variable normal estándar y donde  o sea, el modelo de regresión utilizado en el análisis.

En síntesis, dado el modelo

Donde si el evento se presenta, si el evento no se presenta y .

La estimación da por resultado

Calculando

Donde indica valor promedio de la variable si es cuantitativa y presencia o ausencia de la variable en la observación si es cualitativa.

La probabilidad se calcula haciendo

Encontrar equivale a buscar el valor de Z en la tabla de la normal y determinar su probabilidad asociada.

El efecto marginal de cambio en la probabilidad ante cambios en una variable cuantitativa viene dado por

Donde es la función de densidad de la distribución normal de que el evento promedio ocurra y el coeficiente de la variable para la cual se analiza el cambio.

#### 19.6 Modelo Tobit

Es una extensión del modelo Probit, desarrollado por el Nobel J. Tobin.

Continuando con el ejemplo de la vivienda, supóngase ahora que se desea encontrar la cantidad de dinero que el consumidor gasta en comprar una casa en relación con su ingreso (y otras variables económicas).

Ahora se tiene un problema: *si un consumidor no compra una casa, obviamente no se tiene información sobre el gasto en vivienda, se tiene tal información solamente sobre los consumidores que efectivamente compran casa.*

Por lo tanto se tiene dos grupos de consumidores. Unos, digamos  sobre quienes se posee información sobre los regresores y la variable dependiente y otros, supongamos  sobre quienes solamente se tiene información sobre los regresores.

Cuando en una muestra la información para la variable dependiente está disponible solamente para algunas observaciones, ésta se conoce como muestra censurada. Por consiguiente el modelo Tobit también se conoce como modelo de regresión censurada.

En términos matemáticos se puede expresar el modelo Tobit como

; si tiene datos

; en los demás casos

Ante esta situación el modelo sólo se puede estimar por Máxima Verosimilitud, ya que el término de error no cumple con la propiedad de media nula. Esto se debe a que sólo se incluyen en la muestra las observaciones para las cuales , que puede verse si se escribe el modelo en forma de desviaciones.

#### 19.7 Evaluación de los modelos de probabilidad

Independientemente de si se ha especificado y estimado un modelo logit o un modelo probit, para evaluar un modelo de probabilidad se tiene en cuenta:

* Significatividad individual a la prueba Z, estos modelos presuponen la existencia de gran número de datos por lo que la habitual prueba t tiende a la prueba Z
* Significatividad conjunta a través del estadístico de Wald, el que se distribuye con una chi cuadrado con q grados de libertad (siendo q el número de restricciones). La hipótesis nula es
* Significatividad conjunta a través del estadístico de la razón de verosimilitud (RV), el que se distribuye como una chi cuadrada con k-1 grados de libertad (k es el número de parámetros)

Porcentaje de predicciones correctas, es el número de veces en que coincide el valor observado de Y con el valor estimado, respecto del total de observaciones.

* Pseudo (de Mc Fadden), se basa el logaritmo de la función de verosimilitud

Donde es el logaritmo de la función de verosimilitud para el modelo estimado y es el logaritmo de la función de verosimilitud para un modelo estimado solo de término constante. El Pseudo varía entre 0 y 1, pero generalmente no alcanza valores cercanos a 1.

* Criterios de información de Akaike, Schwarz o Hanna Quinn (HQ), estos son indicadores válidos cuando se tiene que elegir entre varios modelos alternativos, siempre debe escogerse aquel que tenga menor valor en estos indicadores.

Gujarati afirma que en un modelo de regresión binaria es importante que los signos de los coeficientes coincidan con lo esperado y que sean significativamente distintos de 0. La bondad de ajuste a través del Pseudo tiene importancia secundaria.

### **CASOS DE ESTUDIO, PREGUNTAS Y PROBLEMAS**

#### Caso 19.1: Vulnerabilidad social en los hogares de Río Cuarto

El objetivo del trabajo es identificar los fenómenos de mayor impacto en la determinación de la vulnerabilidad en la ciudad de Río Cuarto

El diccionario de la Real Academia Española define vulnerable como aquel que puede ser herido o recibir una lesión física o moral

Roberto Pizarro[[2]](#footnote-3), consultor de la División de Estadística y Proyecciones Económicas de CEPAL, define la vulnerabilidad social como un estado de inseguridad e indefensa que experimentan los individuos en sus condiciones de vida, en el manejo de recursos y en las estrategias que utilizan para enfrentar las consecuencias del impacto provocado por algún tipo de evento económico social

La población estudiada consta de 1686 hogares que tienen bebés nacidos en el año 2005. Las características cualitativas medidas en estos hogares reúnen 527 modalidades.

Dado que la vulnerabilidad es un concepto multidimensional, se optó por realizar un análisis factorial de correspondencia múltiple utilizando 90 variables activas con 381 modalidades asociadas y 9 variables ilustrativas con 146 modalidades asociadas. La partición del espacio de observación y la posterior clasificación de los hogares dio lugar a la conformación de 9 grupos, tal como puede observarse en el Gráfico.



Los hogares que forman los grupos de vulnerabilidad alta (segmentados en los subgrupos Alta-Alta, Alta-Media y Alta-Baja) se caracterizan por tener

•Baja edad y bajo nivel educativo de la madre y situaciones de maternidad en cuasi-soledad

•Alto número de hijos, aun en madres muy jóvenes, y comienzo tardío en la atención durante el embarazo

•Hogares numerosos con alto número de menores de 15 años, bajos ingresos con pocos aportantes por hogar, no hay aportes jubilatorios del jefe de hogar y se observan jóvenes entre 15 y 24 años que no estudian ni trabajan.

•Presentan déficit en la infraestructura sanitaria de la vivienda, no cuentan con gas natural, y se proveen de energía eléctrica de manera irregular, no son propietarios, los hogares tienen hacinamiento

•Están expuestos a riesgos por accidentes con electricidad, los bebés carecen de identificación y no cuentan con cobertura de salud

•Los barrios Alberdi, Banda Norte y Santa Teodora es la ubicación geográfica de estos grupos.

Además se observó que existían variables comunes a todos los grupos, independientes del nivel de vulnerabilidad; estas eran

•Edad de la madre

•Edad de la madre al tener el primer hijo

•Existencia de baño en la vivienda

•Provisión de agua potable

•Cantidad de cuartos

•Número de integrantes del hogar

•Ingreso del hogar

•Ingreso per cápita mensual

•Personas que aportan ingresos

Ahora bien, ¿qué nivel de impacto tienen estos factores en los niveles de vulnerabilidad de los hogares?

Para responder a esto se provee de la Tabla 16.1 que contiene:

* Unidades de observación: 1549 hogares con hijos nacidos en 2005
* Características observadas

Altavul: 1 pertenencia a grupo de alta vulnerabilidad

0 no pertenencia a grupo de alta vulnerabilidad

MAD: 1 madre adolescente

0 madre adulta

NBI: 1 hogar con NBI

0 hogar sin NBI

AI: cantidad de personas que aportan ingresos en el hogar *(variable continua)*

IM: ingreso mensual del hogar *(variable continua)*

IPC: ingreso per cápita diario *(variable continua)*

Edad: edad actual de la madre *(variable continua)*

Se solicita

* la estimación a través del Modelo Logit y el Modelo Probit
* realizar comparaciones entre los resultados de ambos modelos

### **BIBLIOGRAFIA**

* Chou, Y.L. (1977) “Análisis Estadístico” Segunda Edición. Interamericana. México
* Gujarati, D. (2004) "Econometría". 4°Edición. Mc.Graw Hill. México.
* Gujarati, D. (2010) "Econometría". 5°Edición. Mc.Graw Hill. México.
* Maddala, G. S. “Introducción a la Econometría” 2º Edición. Prentice Hall. México.
* Perez Lopez, C. (2006). “Problemas Resueltos de Econometría”. Editorial Thomson Paraninfo.
* Schmidt, S. J. (2005) “Econometría”. Mc-Graw Hill Interamericana. México.

1. El modelo de probabilidad Bernoulli se aplica a una variable que puede asumir solo dos valores y donde la suma de las probabilidades de ocurrencia de cada uno de esos valores es igual a 1. El modelo binomial es la suma de n variables Bernoulli.

   |  |  |  |
   | --- | --- | --- |
   |  | Bernoulli | Binomial |
   | Media | p | np |
   | Varianza | p(1-p) | np(1-p) |

   Ya-Lu Chou (1977) pág. 152 [↑](#footnote-ref-1)
2. Pizarro, Roberto (2001). La vulnerabilidad social y sus desafíos. Una mirada desde América Latina. Estudios Estadísticos y Prospectivos. Serie 6. CEPAL [↑](#footnote-ref-3)