ECONOMETRÍA
APLICADA UTILIZANDO R.

PAPIME PE302513 Libro electrónico y complementos didácticos en medios computacionales, para el fortalecimiento de la enseñanza de la econometría

Capítulo 13.

Modelos Logit y Probit

Luis Quintana Romero y Miguel Ángel Mendoza González

La importancia de las variables categóricas

- En el análisis económico se utilizan variables categóricas, las cuales son indicadoras de la presencia o ausencia de algún atributo. Sobre todo, en la información proveniente de micro datos de individuos, de empresas o de familias es común encontrar este tipo de variables.

- En encuestas demográficas es común encontrar variables como el género de las personas que habitan una vivienda, en las encuestas industriales se reporta si la empresa tuvo acceso o no al crédito del sistema financiero y la variable respectiva es simplemente si o no; en todos estos casos las variables se registran en forma binaria, utilizando el número 1 para indicar la presencia del atributo respectivo y el 0 para su ausencia.

- En inglés es usual denominar a esas variables binarias como dummys, término que en castellano se ha naturalizado y muchos hacen referencia a esas variables, de forma indistinta, como binarias o dummys.

- Para ilustrar este tipo de variables, en el cuadro siguiente se presentan los datos sobre el género de los trabajadores mexicanos obtenidos de la Encuesta Nacional de Ocupación y Empleo (ENOE) para el segundo trimestre de 2015. De esa información se deriva que de un total de 49.6 millones de trabajadores, el 62% son hombres y el 38% son mujeres.

- Cuadro 1

Género de la fuerza de trabajo en México, 2015

Género	Frecuencia	Porcentaje
0	18,738,988	37.8
1	30,832,408	62.2
Total	49,571,396	100

Hombre=1

Mujer=0

Fuente: Con base en datos del INEGI, ENOE, segundo trimestre de 2015

- Un rápido vistazo a los datos de la ENOE permite observar que la variable género simplemente se va reportando con ceros y unos, tal y como se observa en el cuadro siguiente; por ejemplo, en el hogar 1 se entrevistó a una mujer (género=0), con estado civil soltera (casado=0), mientras que en el hogar 2 es un hombre y está casado.

Hogar	Casado	Genero
1	0	0
2	1	1
3	1	0
4	0	1
5	0	0
6	0	0
7	1	1
8	1	0
9	0	1
10	0	0

- En los modelos econométricos este tipo de variables binarias suelen incorporarse, sobre todo cuando la información proviene de micro datos. Por ejemplo, en los estudios sobre las remuneraciones de los trabajadores suelen estimarse ecuaciones conocidas como mincerianas y que fueron propuestas por Jacob Mincer en su conocido libro publicado en 1974.

Fuente: Con base en datos del INEGI, ENOE, segundo trimestre de 2015

- En ese texto Mincer establece la existencia de una relación positiva entre el salario y la escolaridad de los individuos. Con el fin de analizar algún tipo de discriminación salarial para las mujeres, en las ecuaciones mincerianas se incorpora una variable binaria de género como la que ya se ha descrito antes. El modelo se especifica del siguiente modo:

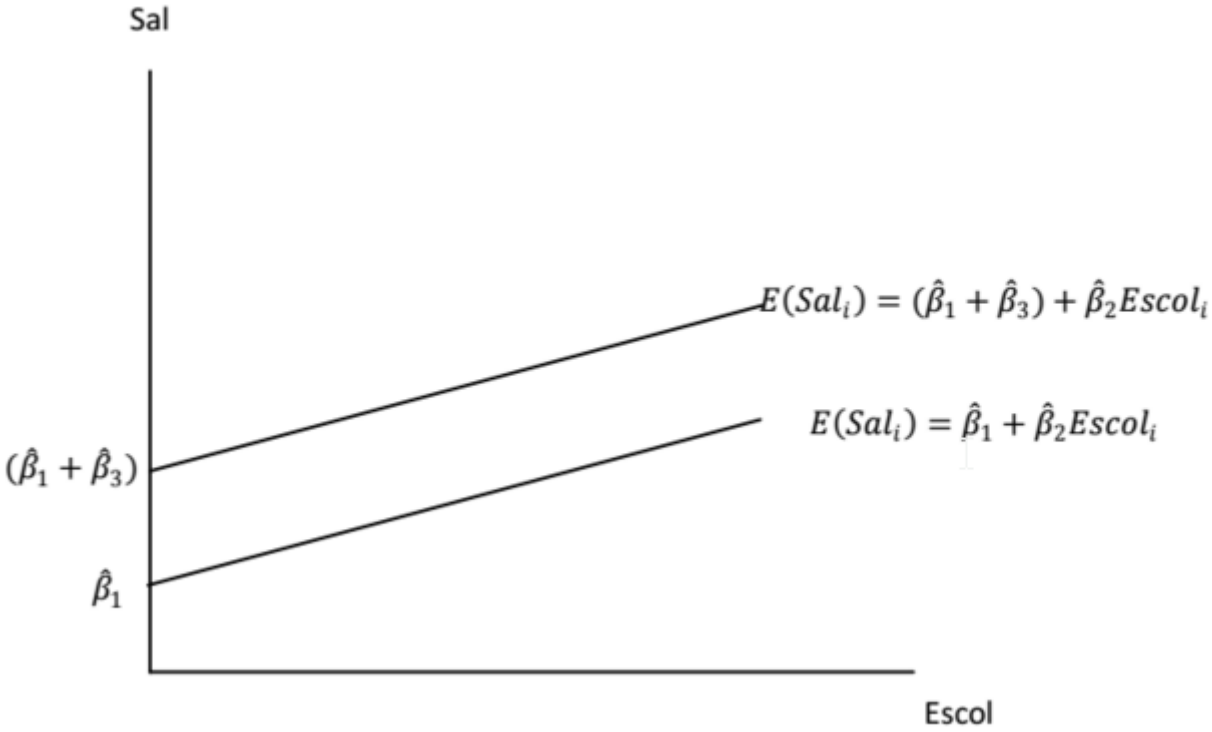
$$Sal_i = \beta 1 + \beta 2 Escol_i + \beta 3 Gen_i + ui \quad (1)$$

donde:

Sal es el salario en pesos, Escol son los años de educación, Gen es una variable binaria con 1 cuando es hombre y 0 cuando es mujer y u es un término de perturbación aleatoria.

- Al estimar un modelo como el de la ecuación (1) la recta de regresión tendría un intercepto diferente para hombres y para mujeres (si los valores estimados para los coeficientes β1 y β2 del modelo son positivos y significativos), pero se mantendría la misma pendiente ya que la variable escolaridad no está diferenciada por género.

- Gráfica 1 Regresión con variable binaria explicativa



- En estos modelos las variables binarias operan como uno más de los regresores de la ecuación. Sin embargo, si estas variables las utilizamos como variables dependientes es necesario considerar otro tipo de modelos, los cuales se revisaran en este capítulo en las secciones siguientes.

Modelos Logit y Probit

- Cuando la variable binaria es la variable dependiente a explicar, el modelo de regresión se interpreta como probabilidades. Retomando el ejemplo de la ecuación minceriana, el modelo se podría reformular considerando una variable salarial binaria; igual a la unidad para salarios por encima de la media y cero para salarios por debajo de la media. La especificación del modelo se muestra a continuación.

$$Sal\ i = \beta 1 + \beta 2 Escol\ i + \beta 3 Gen\ i + u\ i \quad (2)$$

En la ecuación previa la variable sal *i* es igual a 1 cuando el salario del individuo i está por encima de la media y 0 en otro caso.

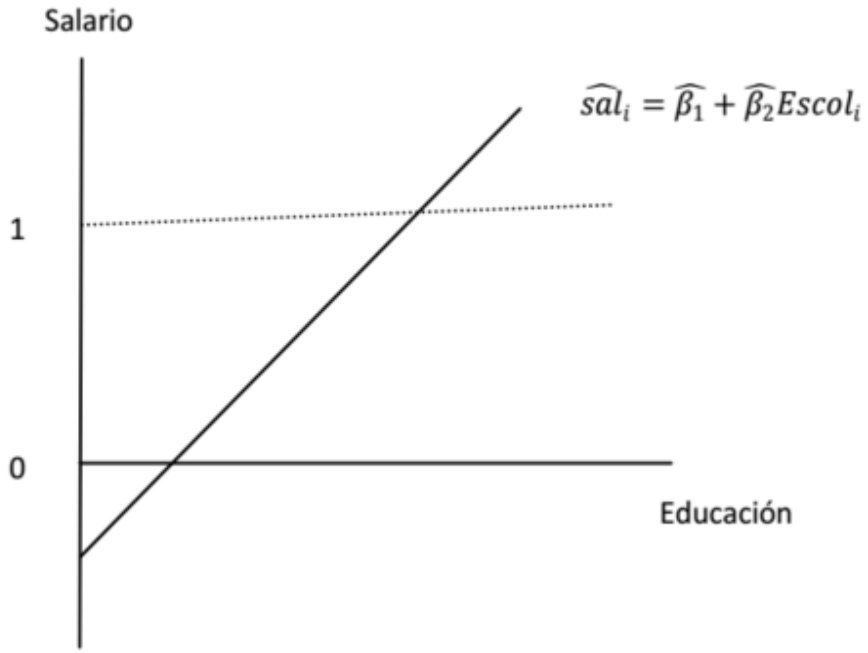
- Las preguntas que podríamos responder con esta ecuación son, por ejemplo, ¿cuál es la probabilidad de que el salario se encuentre por encima de la media cuando el individuo tiene cierto número de años de educación? o bien ¿cuál es la probabilidad de que el salario este por encima de la media cuando el individuo es mujer?

- Para simplificar la explicación suponga que se estima una versión más restringida de la ecuación previa:

$$Sal_i = \beta 1 + \beta 2 Escol_i + u_i \quad (3)$$

- El resultado de la estimación por mínimos cuadrados ordinarios podría graficarse de la siguiente manera:

Gráfica 2: Regresión con variable dependiente binaria



Si tomamos las probabilidades tendríamos:

P(sal=1|Educ) = F β1 + β2Escol *i*

Objetivo

Hacer una introducción a los modelos panel, así como conocer los diferentes tipos del mismo.

- Tal como se observa en la gráfica 2, la probabilidad de que el salario sea mayor a la media, dados los años de educación, es una función lineal de la educación. En teoría la función de probabilidad debe tener estrictamente valores entre cero y uno, sin embargo, en la gráfica nada garantiza eso y las probabilidades no tendrían sentido cuando obtenemos valores negativos o mayores a la unidad.

- Este tipo de modelos se conocen como Modelo de Probabilidad Lineal (MPL), pero su utilidad es limitada dado el resultado mencionado antes.

- Para lograr asegurar que las probabilidades estén restringidas a valores entre cero y uno se han sugerido dos modelos fundamentales; el logístico o logit y el probabilístico o probit.

- El logístico se especifica a través de una función logística de la siguiente manera:

Función logística: $F(z) = \frac{Exp\ (z)}{[1+Exp\ (z)]} = Pi$ (4)

donde $Zi = \beta 1 + \beta 2X1,i + \cdots + \beta kXk,i$

- La expresión previa es simplemente una función de distribución acumulada para una variable aleatoria logística Z. Por lo cual el modelo de regresión Logit quedaría especificado de la siguiente manera:

$$Y\ i = \frac{Exp\ (z)}{[1+Exp\ (z)]} + \varepsilon i$$

- Con base en la probabilidad es posible construir la razón de probabilidades (Gujarati, 2014):

$$\frac{\mathrm{P}i}{1-\mathrm{P}i} = \frac{1+ Exp\ (z)}{[1+Exp\ (-z)]} = \mathrm{Exp}\ (\mathrm{z}) \qquad (5)$$

- Por consiguiente al tomar logaritmos en (5) se obtiene el logit:

$$\mathrm{L}i = ln\ [\frac{Pi}{1-Pi}]$$

- En el ejemplo de los salarios de la ecuación (4) la razón de probabilidades indicaría la razón de la probabilidad de tener un salario por arriba de la media en relación a tenerlo por debajo de la media, dado el nivel educativo.

- La regresión logística no supone linealidad como en los modelos de regresión clásica, tampoco requiere del supuesto de normalidad ni del de homocedasticidad (Garson, 2014). Sin embargo, si requiere que las observaciones sean independientes y que las variables explicatorias estén relacionadas linealmente al logito de la variable dependiente, tal y como se expresa esa relación en la ecuación (4).

- El probit, por otro lado, se especifica a través de la siguiente función de distribución acumulada normal:

$$\text{Modelo Probit: } F(z) = \Psi z = \int_{\infty}^{1} \Psi\ (v)dv \quad (5)$$

En donde Ψ(z) es la distribución normal estándar: $\varphi v = (2\pi)^{-1/2} \exp\left(-\frac{Z2}{2}\right)$

Por lo cual el modelo de regresión Probit quedaría especificado de la siguiente manera:

$$Yi = \Psi(z) + \varepsilon i = \int_{\infty}^{1} \Psi\ (v)dv + \varepsilon i = \int_{\infty}^{1} (2\pi)^{-1/2} \exp\left(-\frac{Z2}{2}\right)\ dv + \varepsilon i$$

.

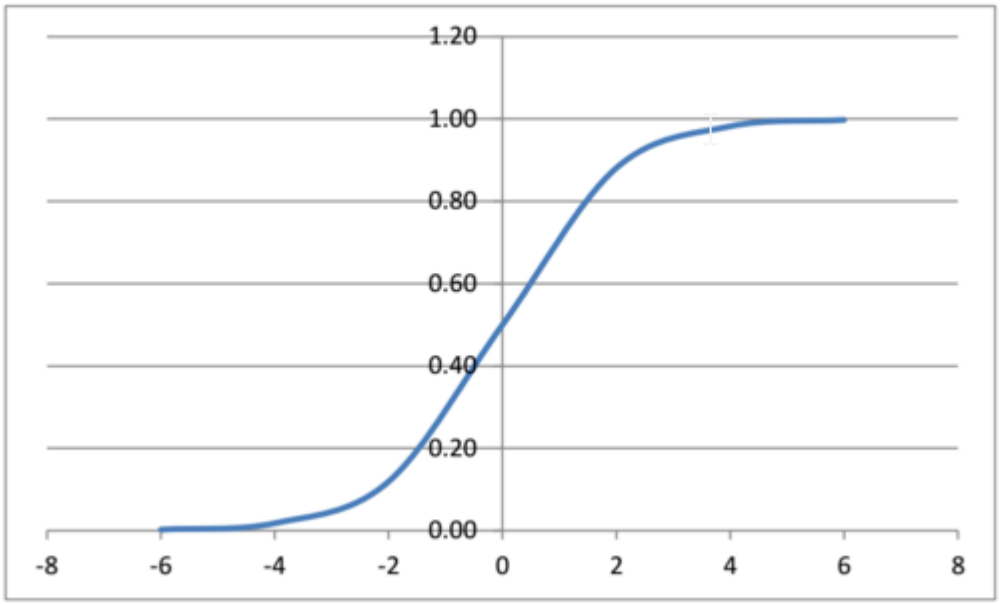
- En general, los resultados de los modelos logit y probit permiten llegar a las mismas conclusiones ya que sus coeficientes sólo difieren en escala; los coeficientes logit son aproximadamente 1.8 veces los que se obtienen en el probit. Tal vez la desventaja más visible de los probit es que sus coeficientes son más difíciles de interpretar y además, debido al supuesto de normalidad, no se recomienda su uso cuando las observaciones se concentran mucho en alguna de las colas de la distribución (Garson, 2012).

- Utilizando los valores para Z que vienen en la siguiente tabla y auxiliándose con una simple calculadora se podría aplicar la fórmula (4) del modelo logístico y construir una función logística.

Cuadro 1: Función logística

Z	Logit
6	1.00
4	0.98
2	0.88
0	0.50
-2	0.12
-4	0.02
-6	0.00

Gráfica 4: Función logística



En el caso del modelo probit se pueden sustituir las medias de las variables explicatorias en la ecuación (5) para obtener las estimaciones de los valores Z y luego simplemente buscar en la tabla de la normal los niveles de probabilidad que les corresponden.

Estimación por MV

- Son modelos estimados por MV debido a su no linealidad. Este método tiene ventajas estadísticas en virtud de que sus estimaciones son consistentes, eficientes y para muestras grandes son insesgadas y su distribución se aproxima a una normal (Garson, 2014).

- Para estimarlos es necesario tener la densidad de y dada x, la cual es una función binaria de éxito y fracaso:

$$f\,(y|xi,\,\beta) = [F\,(xi\beta)]_{y} \quad [1 - F\,(\,xi\beta)\,]^{\,1-y}$$

Al tomar logaritmos tenemos la logMV:

$$l\,(\beta) = y[F(\,xi\beta)\,] + (1 - y)\,[1 - F\,(xi\beta)\,]$$

- La ecuación se maximiza de manera usual tomando las condiciones de primero y segundo orden, se igualan a cero y se resuelve el sistema de ecuaciones resultante. Sin embargo, es un sistema de ecuaciones no lineales, por lo cual se debe utilizar algún algoritmo de optimización que permita a los estimadores la convergencia.

Pruebas de hipótesis

- Se pueden aplicar pruebas de restricciones tipo Wald. Una prueba usual en este sentido consiste en comparar la razón de verosimilitud (LR) del modelo que se está estimando en relación al modelo nulo, en el cual los coeficientes de las variables explicativas están restringidos a ser nulos. Si el LR es significativamente diferente de cero tendremos evidencia de que el modelo que se está estimando es diferente al nulo.

- La bondad de ajuste se obtiene con base en el porcentaje correctamente predicho por el modelo: se define un valor predicho de uno si la probabilidad predicha es de menos 0.5 y de cero en caso contrario. El porcentaje predicho correctamente es el número de veces en que el valor estimado es igual al real.

- En ese sentido las R cuadradas son en realidad seudo R cuadradas. Las más usuales son las siguientes.

$$\text{McFadden (1974)} = \frac{\log MV}{\log MV\ (0)}$$

- Es decir, toma las funciones log verosimilitud no restringida (logMV) y la restringida logMV(0) (con sólo la pendiente). Si las variables no explican nada logMV=logMV(0) y por ende la seudo Rcuadrada será cero.

- Otras alternativas toman correlaciones entre las variables estimadas y las reales, lo cual es más cercano al espíritu de la R cuadrada en modelos de MCO.

El curso corresponde al proyecto PAPIME: Libro electrónico y complementos didácticos en medios computacionales, para el fortalecimiento de la enseñanza en la econometría.

FES Acatlán UNAM

Coordinación general:

Dr. Luis Quintana Romero.

Dr. Miguel Ángel Mendoza González.

Voz en off: Nancy Nayely Morales Parada

Coordinación de edición:

Mtro. José Antonio Huitrón Mendoza.

Edición:

Claudia Viridiana Torres Trejo.