**Trabajo Práctico N° 2:**

**Extensiones de Modelos Logit y Probit.**

**Ejercicio 1.**

*Considerar la siguiente armación: “La estimación de un modelo de probabilidad lineal es más robusta que Probit o Logit porque el modelo de probabilidad lineal no asume homocedasticidad ni tiene supuestos acerca de la distribución de los errores.”*

En esta afirmación, se propone una comparación que no es adecuada.

**Ejercicio 2: Probit con una Variable no Observable.**

*Considerar el modelo Probit:*

*P (y= 1 | z, q)= ( + q),*

*donde q es independiente de z y distribuido normal (0, 1); el vector z es observado, pero el escalar q no lo es.*

**(a)** *Encontrar el efecto parcial de sobre la probabilidad de respuesta, a saber, .*

= ( + *q*) q.

**(b)** *Mostrar que P (y= 1 | z)= ().*

Se escribe:

= + r,

con r= q + e, donde e (0, 1) y es independiente de (z, q).

Como se asume que q es independiente de z, se tiene:

E (r | z)= E (q + e | z)

E (r | z)= E (q | z) + E (e | z)

E (r | z)= E (q | z) + E (e)

E (r | z)= E (q) + 0

E (r | z)= \* 0 + 0

E (r | z)= 0 + 0

E (r | z)= 0.

Var (r | z)= Var (q + e | z)

Var (r | z)= Var (q | z) + Var (e | z) + 2 Cov (q, e | z)

Var (r | z)= Var (q | z) + Var (e) + 2 \* 0

Var (r | z)= Var (q) + 1 + 0

Var (r | z)= \* 1 + 1 + 0

Var (r | z)= 1 + .

Entonces, se puede armar la distribución de y ver que:

P (y= 1 | z)= ().

**(c)** *Definir . ¿Cómo se testearía la hipótesis : = 0?*

Definiendo , la hipótesis : = 0 se podría testear usando un Score Test o un LM Test.

**(d)** *Si se tuvieran motivos para creer que 0, ¿cómo se estimaría junto con ?*

Si se tuvieran motivos para creer que 0, se estimaría junto con mediante el método de máxima verosimilitud.

**Ejercicio 3.**

*Considerar una gran muestra aleatoria de trabajadores en un momento dado. Sea una variable que vale 1 si la persona i se reportó enferma durante los últimos 90 días y vale 0 en caso contrario. Sea un vector de características del individuo y del empleador. Sea el número de cigarrillos que fuma el individuo i por día (en promedio).*

**(a)** *Explicar el experimento subyacente de interés cuando se quieren examinar los efectos del tabaquismo en los días de trabajo perdidos.*

El experimento subyacente de interés cuando se quieren examinar los efectos del tabaquismo en los días de trabajo perdidos es qué analizar qué efecto tendrá sobre la probabilidad de que una persona se reporte enferma durante los últimos 90 días el cambio exógeno del número de cigarrillos que fuma por día esa persona. En otras palabras, se quiere inferir causalidad, no sólo encontrar una correlación entre el ausentismo en el trabajo y el tabaquismo.

**(b)** *¿Por qué podría estar correlacionada con variables no observables que afectan a ?*

Dado que las personas eligen si fumar y cuánto, ciertamente, no se puede tratar a los datos como si provinieran del experimento que se tiene en mente en el inciso (a). Es decir, no se puede asignar a las personas, aleatoriamente, un consumo de cigarrillos diario.

El consumo de cigarrilos diario puede estar correlacionado con variables no observables que afectan la falta en el trabajo. Por ejemplo, los fumadores pueden ser menos saludables o tener otros atributos que les hagan faltar al trabajo con más frecuencia; o, por el contrario, el consumo de cigarrillos puede estar relacionado con rasgos de la personalidad que hacen que las personas trabajen más. En cualquier caso, el consumo de cigarrillos diarios podría estar correlacionado con elementos no observables de la ecuación.

**(c)** *Una forma de escribir el modelo de interés es:*

*P (sick= 1 | z, cigs, )= (+ cigs + ),*

*donde es un subconjunto de z y es una variable no observable que, posiblemente, esté correlacionada con cigs. ¿Qué sucede si se ignora y se estima el Probit de sick sobre y cigs?*

Lo que sucede si se ignora y se estima el Probit de *sick* sobre y *cigs* es que los estimadores serán incosistentes.

**(d)** *¿Puede cigs tener una distribución normal condicional en la población? Explicar.*

Dado que, en la población, hay muchas personas que no fuman, la distribución (condicional o incondicional) de consumo de cigarrillos diarios se “apila” en cero. Además, la variable *cigs* toma valores enteros positivos, por lo que no puede tener una distribución normal condicional en la población.

**(e)** *Explicar cómo probar si cigs es exógeno. ¿Esta prueba se basa en cigs que tienen una distribución normal condicional?*

Para probar si *cigs* es exógeno, se puede utilizar el procedimiento de dos etapas de Rivers y Vuong (1988).

**(f)** *Suponer que algunos de los trabajadores viven en estados que, recientemente, implementaron leyes de no fumar en el lugar de trabajo. ¿La presencia de las nuevas leyes sugiere un buen candidato IV para cigs?*

Suponiendo que las personas no se mudarán, inmediatamente, de su estado de residencia cuando el estado implemente leyes de no de fumar en el lugar de trabajo y que ese estado de residencia es, aproximadamente, independiente de la salud general de la población, un indicador *dummy* que diga si la persona trabaja en un estado con una nueva ley puede funcionar como una variable exógena. Estas situaciones, a menudo, se denominan “experimentos naturales”. Además, es probable que la variable *cigs* esté correlacionada con el indicador de la ley estatal porque las personas no podrán fumar tanto como lo harían de no existir la ley. Por tanto, la presencia de las nuevas leyes sugiere un buen candidato IV para *cigs*.

**Ejercicio 4.**

*Utilizar el conjunto de datos “BWGHT.dta” para este problema.*

**(a)** *Definir una variable binaria, smokes, si la mujer fuma durante el embarazo. Estimar un modelo Probit que relacione smokes con motheduc, white y log(faminc). En white= 0 y faminc evaluado en el promedio de la muestra, ¿cuál es la diferencia estimada en la probabilidad de fumar para una mujer con 16 años de educación y una con 12 años de educación?*

Probit:

Probit regression Number of obs = 1,387

LR chi2(3) = 92.67

Prob > chi2 = 0.0000

Log likelihood = -546.76991 Pseudo R2 = 0.0781

------------------------------------------------------------------------------

smokes | Coefficient Std. err. z P>|z| [95% conf. interval]

-------------+----------------------------------------------------------------

motheduc | -.1450599 .0207899 -6.98 0.000 -.1858074 -.1043124

white | .1896765 .1098805 1.73 0.084 -.0256853 .4050383

lfaminc | -.1669109 .0498894 -3.35 0.001 -.2646923 -.0691296

\_cons | 1.126276 .2504611 4.50 0.000 .6353817 1.617171

------------------------------------------------------------------------------

La diferencia estimada en la probabilidad de fumar para una mujer con 16 años de educación y una con 12 años de educación es -0,086.

**(b)** *¿faminc es exógena en la ecuación de smokes? ¿Qué pasa con motheduc?*

*faminc* puede llegar a ser endógena en la ecuación de *smokes*.

**(c)** *Suponer que motheduc y white son exógenos en el Probit del inciso (a). Suponer, también, que fatheduc es exógeno a esta ecuación. Estimar la forma reducida de log(faminc) para ver si fatheduc está parcialmente correlacionada con log(faminc).*

Probit:

Source | SS df MS Number of obs = 1,191

-------------+---------------------------------- F(3, 1187) = 119.23

Model | 140.936735 3 46.9789115 Prob > F = 0.0000

Residual | 467.690904 1,187 .394010871 R-squared = 0.2316

-------------+---------------------------------- Adj R-squared = 0.2296

Total | 608.627639 1,190 .511451797 Root MSE = .6277

------------------------------------------------------------------------------

lfaminc | Coefficient Std. err. t P>|t| [95% conf. interval]

-------------+----------------------------------------------------------------

motheduc | .0709044 .0098338 7.21 0.000 .0516109 .090198

white | .3452115 .050418 6.85 0.000 .2462931 .4441298

fatheduc | .0616625 .008708 7.08 0.000 .0445777 .0787473

\_cons | 1.241413 .1103648 11.25 0.000 1.024881 1.457945

------------------------------------------------------------------------------

**(d)** *Contrastar la hipótesis nula de que log(faminc) es exógena en el Probit del inciso (a).*

Probit regression Number of obs = 1,191

LR chi2(4) = 79.43

Prob > chi2 = 0.0000

Log likelihood = -432.06242 Pseudo R2 = 0.0842

------------------------------------------------------------------------------

smokes | Coefficient Std. err. z P>|z| [95% conf. interval]

-------------+----------------------------------------------------------------

motheduc | -.0826247 .0465204 -1.78 0.076 -.173803 .0085536

white | .4611075 .1965245 2.35 0.019 .0759265 .8462886

lfaminc | -.7622559 .3652949 -2.09 0.037 -1.478221 -.046291

v2hat | .6107298 .3708071 1.65 0.100 -.1160387 1.337498

\_cons | 1.98796 .5996374 3.32 0.001 .8126927 3.163228

------------------------------------------------------------------------------

(1) [smokes]v2hat = 0

chi2( 1) = 2.71

Prob > chi2 = 0.0996

Por lo tanto, con un nivel de significancia del 10%, estos datos aportan evidencia suficiente para indicar que *log(faminc)* es endógena.

**Ejercicio 5.**

*Una preocupación común cuando se utilizan precios autoinformados en la estimación de la prevalencia del tabaquismo con una base de datos de corte transversal (por ejemplo, Global Adult Tobacco Survey o GATS) es la potencial endogeneidad de esta variable. Para abordar este problema potencial, se construyen dos variables de precios diferentes. La primera variable de precio asigna a los fumadores el precio autoinformado pagado por la última compra y utiliza una imputación de regresión aleatoria (random regression imputation, a veces denominada imputación de regresión estocástica) para asignar un precio a los no fumadores de la muestra. La segunda variable de precio asigna a fumadores y no fumadores el promedio del precio autoinformado por unidad primaria de muestreo (UPM, o PSU por Primary Sampling Unit). Siguiendo las recomendaciones en “Economics of Tobacco Toolkit: Economic Analysis of Demand Using Data from the Global Adult Tobacco Survey (GATS)” (John et al., 2019), se puede verificar la endogeneidad del precio autoinformado utilizando el test de Rivers-Vuong (1988).*

**(a)** *¿Por qué podrían ser endógenos los precios autoinformados?*

Los precios autoinformados podrían ser endógenos porque pueden estar correlacionados con variables omitidas en el modelo, que, a su vez, correlacionen con la variable dependiente.

**(b)** *Realizar el test de Rivers-Vuong para los datos provistos en “pricedata.dta” utilizando las variables X en la primera etapa y Z en la segunda etapa.*

Adjusted Wald test

(1) [SmokeCigs]resid1 = 0

F( 1, 5976) = 18.77

Prob > F = 0.0000

Por lo tanto, con un nivel de significancia del 1%, estos datos aportan evidencia suficiente para indicar que los precios autoinformados son endógenos.

**(c)** *En función de los resultados, estimar la elasticidad de la prevalencia del tabaquismo con respecto a los precios.*

Stata.

**Ejercicio 6.**

*Se busca simular el siguiente modelo:*

*Pr (y= 1)= F {}.*

*Generar un dataset vacío con 1000 obsevaciones. Generar las siguientes variables:*

*x U (-1, 1),*

*U (0, 1),*

*,*

*p (),*

*con = 0,3 y = 2 y definir la variable dependiente y como una variable binaria que vale 1 si p es mayor o igual a una variable aleatoria uniforme en el intervalo (0, 1) y 0 en caso contrario. Estimar el modelo Probit heterocedástico y comparar con las estimaciones del Probit usual.*

Probit heterocedástico:

Heteroskedastic probit model Number of obs = 1,000

Zero outcomes = 468

Nonzero outcomes = 532

Wald chi2(1) = 78.21

Log likelihood = -563.0256 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

y | Coefficient Std. err. z P>|z| [95% conf. interval]

-------------+----------------------------------------------------------------

y |

x | 2.484689 .2809507 8.84 0.000 1.934036 3.035342

\_cons | .2876884 .0939283 3.06 0.002 .1035924 .4717845

-------------+----------------------------------------------------------------

lnsigma |

xhet | 1.734142 .2630328 6.59 0.000 1.218608 2.249677

------------------------------------------------------------------------------

LR test of lnsigma=0: chi2(1) = 51.24 Prob > chi2 = 0.0000

Probit:

Probit regression Number of obs = 1,000

LR chi2(1) = 204.90

Prob > chi2 = 0.0000

Log likelihood = -588.64728 Pseudo R2 = 0.1482

------------------------------------------------------------------------------

y | Coefficient Std. err. z P>|z| [95% conf. interval]

-------------+----------------------------------------------------------------

x | 1.054521 .0772801 13.65 0.000 .903055 1.205987

\_cons | .1085126 .0424302 2.56 0.011 .025351 .1916742

------------------------------------------------------------------------------

Tabla comparativa:

--------------------------------------------

(1) (2)

Probit Het~o Probit

--------------------------------------------

y

x 2.485\*\*\* 1.055\*\*\*

(0.281) (0.0773)

\_cons 0.288\*\*\* 0.109\*\*

(0.0939) (0.0424)

--------------------------------------------

lnsigma

xhet 1.734\*\*\*

(0.263)

--------------------------------------------

N 1000 1000

pseudo R-sq 0.148

--------------------------------------------

Standard errors in parentheses

\* p<0.10, \*\* p<0.05, \*\*\* p<0.01