universidad torcuato di tella maestría en economía — maestría en econometría 2021

Microeconometría Problem Set 1 Modelo de Probabilidad Lineal, Logit y Probit

• Ejercicio 1. Porcentaje correctamente predicho.

Sea y una variable binaria y considere algún modelo de probabilidad $P(y = 1|x) = F(X\beta)$. Muestre que el porcentaje general predicho correctamente es un promedio ponderado del porcentaje predicho para la variable dependiente igual a 0 (\hat{q}_0) y del porcentaje predicho para la variable dependiente igual a 1 (\hat{q}_1) , donde las ponderaciones son las proporciones de ceros y de unos en la muestra, respectivamente.

• Ejercicio 2. Interpretación del Modelo de Probabilidad Lineal I.

Suponga que se estima el modelo

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

donde x es una variable continua, mientras que y es una variable que sólo puede valer 0 o 1. El tamaño de la muestraes n y sea n_1 la cantidad de elementos que verifican $y_i = 1$. Llame \bar{x}_1 a la media de la variable x tomada solo para aquellos elementos que verifican $y_i = 1$ y \bar{x}_0 a la media de la variable x tomada sobre los valores restantes. Muestre que

$$\hat{\beta}_1 = \frac{p(1-p)(\bar{x}_1 - \bar{x}_0)}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

donde $p = \frac{n_1}{n}$.

• Ejercicio 3. Interpretación del Modelo de Probabilidad Lineal II.

Sea y una resultado binario y sean d_1, d_2, \ldots, d_M variables binarias mutuamente excluyentes y colectivamente exhaustivas, es decir, cada persona de la población cae en una y sólo una categoría.

(a) Muestre que los valores ajustados de la regresión sin intercepto

$$y_i$$
 sobre $d_{1i}, d_{2i}, \ldots, d_{Mi}$

están siempre en el intervalo unitario. En particular, describa qué representa cada coeficiente y el valor ajustado para cada i.

- (b) ¿Qué ocurre si y_i se regresa sobre M combinaciones lineales de d_1, d_2, \ldots, d_M linealmente independientes entre sí? $Ayuda: considere \ 1, d_2, \ldots, d_M$.
- Ejercicio 4. Efectos marginales I.

Sea y un resultado binario y $\mathbf{x}=(x_1,\ldots,x_k)$ un vector de variables explicativas. Sea $G(\cdot)$ la función de distribución acumulada de una variable aletoria continua. Recuerde que si x_j es continua, su efecto marginal se obtiene como

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})\beta_j, \quad \text{donde } g(z) \equiv \frac{dG}{dz}(z)$$

- (a) Muestre que los efectos relativos de dos variables explicativas cualesquiera no dependen de \mathbf{x} .
- (b) Sea x_1 una variable binaria. ¿Cuál es el efecto parcial de cambiar x_1 de 0 a 1? ¿De qué depende? Interprete en el caso en el que y es un indicador de empleo y x_1 es una variable binaria que indica la participación en un programa de capacitación laboral.
- (c) Sea x_2 una variable discreta numérica. ¿Cuál es el efecto parcial de cambiar x_2 de cierto nivel c a c+1? ¿De qué depende? Interprete en el caso en el que y es un indicador de si la persona i fuma y x_2 la cantidad de cigarrillos que fuma por día.

Considere ahora el siguiente modelo

$$P(y = 1 \mid \mathbf{z}) = G(\beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3)$$

- (d) ¿Cuál es el efecto parcial de z_1 sobre $P(y = 1 \mid \mathbf{z})$?
- (e) ¿Cuál es el efecto parcial de z_2 sobre $P(y = 1 \mid \mathbf{z})$?
- (f) ¿Cuál es la elasticidad de z_3 sobre $P(y=1 \mid \mathbf{z})$? ¿Siempre tiene el mismo signo que β_4 ?
- (g) ¿Cuál es la elasticidad de z_1 sobre $P(y = 1 \mid \mathbf{z})$?
- (h) ¿Cómo obtendría errores estándar para todos estos efectos?

universidad torcuato di tella maestría en economía — maestría en econometría 2021

Microeconometría Ejercicios Introductorios a Stata 1 Modelo de Probabilidad Lineal, Logit y Probit

• Ejercicio 1. MPL, Logit y Probit en Stata I.

En este ejercicio usted va a demostrar algunas propiedades de las estimaciones para modelos con variable dependiente discreta.

- (a) Estime a ins contra retire, age, hstatusg, hhincome, educyear, married, hisp por OLS, Logit y Probit.
- (b) ¿Cuál es el problema de estimar el modelo por OLS?
- (c) Explique analíticamente cuál es la interpretación de un coeficiente β en un modelo de regresión lineal y en un modelo Probit/Logit. ¿Es constante el efecto marginal en los modelos no lineales?
- (d) Para evaluar la eficacia de los modelos Probit y Logit defina como el valor estimado de la variable dependiente y como

$$\hat{y} = \begin{cases} 1 \text{ si } P(\widehat{y=1}|x) > 0.5\\ 0 \text{ si } P(\widehat{y=0}|x) \le 0.5 \end{cases}$$
 (1)

Realice un cuadro de doble entrada con las variables y y \hat{y} . Comente.

- (e) En la literatura se sugiere que $\beta^{logit} \approx 4\beta^{ols}$ y $\beta^{probit} \approx 2.5\beta^{ols}$. Compruébelo para esta muestra.
- (f) Compute la probabilidad esperada que ins=1 cuando las variables están evaluadas en la media.
- (g) Defina el odds ratio como el cociente entre la probabilidad que y = 1 y y = 0. De este modo, un odds ratio de 2 implica que es dos veces más probable que y = 1 a que y = 0. Demuestre que para el caso de un modelo Logit se verifica que

$$\ln\left(\frac{P(y=1|x)}{P(y=0|x)}\right) = X\beta$$

Recuerde que para un modelo Logit

$$P(y = 1|x) = \frac{1}{1 + e^{-X\beta}}$$

• Ejercicio 2. MPL, Logit y Probit en Stata II.

Utilice la base de datos de Mroz, T.A. (1987): "The Sensitiviy of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions", Econometrica, 55, 765-799.. La misma posee datos sobre el desempleo de las mujeres en Estados Unidos en 1975.

- (a) Para comenzar, realize un análisis exploratorio simple de los datos. Para esto, puede ayudarse de los comandos describe, summarize, browse, tab.
- (b) Cree una variable de educación centrada. Recuerde que se le llama variable centrada a una variable transformada como $\tilde{x}_i = x_i \bar{x}$.
- (c) Estudie gráficamente la relación entre el salario y la educación. Puede también desagregar por las variables *inlf, kidslt6*. Para esto, puede ayudarse de los comandos graph, twoway, scatter, lfit y sus opciones.
- (d) ¿Hay valores faltantes o duplicados en la muestra? Intente resolver esto sin el comando browse ni edit.
- (e) Estime un modelo de probabilidad lineal de *inlf* sobre *educ*, *city*, *exper*, *kidslt6*, *expersq*. Además, genere la predicción del modelo.
- (f) ¿Puede realizar inferencia con este modelo? Estime el modelo con errores estándares robustos. ¿Cómo cambian los resultdos?
- (g) ¿Qué ocurre si elimina la constante del modelo?
- (h) ¿Qué ocurre si estima el modelo solo para una ciudad?
- (i) Estime un modelo logit de inlf sobre educ, city, exper, kidslt6, expersq.
- (j) Calcule la predicción del modelo.
- (k) Genere la curva ROC.

- (l) Calcule los efectos marginales en las medias.
- (m) Calcule los efectos marginales en valores particulares de la variable que le resulten de interés.
- (n) Estime un modelo probit con las mismas variables que en (i) y cree una tabla con las estimaciones de todos los modelos.
- Ejercicio 3. Estimar el efecto de la educación sobre la probabilidad de estar desempleado.

Utilice la EPH con datos de individuos del segundo trimestre de 2015, disponible en http://www.indec.gob.ar/bases-de-datos.asp. Use la muestra de jefes de hogar, hombres, 25-65 años, para todos los conglomerados disponibles. Estudie como se define el desempleo de acuerdo al INDEC. Rentrinja la muestra a personas empleadas o desempleadas, es decir excluya aquellos que están fuera de la fuerza laboral (no buscan trabajo, estudian, retirados, etc.). Use las ponderaciones pondera.

- (a) Utilice un modelo de probabilidad lineal para estimar el efecto de la educación sobre la probabilidad de estar desempleado, controlando por ubicación geográfica, edad y estado civil. Construya las probabilidades para cada individuo. ¿Qué proporción de la muestra tiene probabilidades predecidas mayores a 1 o menores a 0 ?
- (b) Estimar el modelo del inciso (a) usando los modelos Probit y logit. ¿Cómo cambian los resultados?
- (c) Estimar la probabilidad de estar desempleado para un hombre universitario (de grado) completo, casado, para cada área metropolitana de la EPH, para todos los aos posibles de edad 25-65. Graficar los efectos marginales de la edad sobre la probabilidad de estar desempleado, junto con los errores estándar de la estimación.

Microeconometría I Problem Set 2 Extensiones de Modelos Logit y Probit

- Ejercicio 1. Considere la siguiente afirmación. "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."
- Ejercicio 2. Probit con una variable no observable.

Considere el modelo Probit:

$$P(y = 1 \mid \mathbf{z}, q) = \Phi \left(\mathbf{z}_1 \delta_1 + \gamma_1 z_2 q \right),$$

donde q is independiente de \mathbf{z} y distribuido Normal(0,1); el vector \mathbf{z} es observado pero el escalar q no lo es.

- 1. Encuentre el efecto parcial de z_2 sobre la probabilidad de respueta, a saber, $\frac{\partial P(y=1|\mathbf{z},q)}{\partial z_2}$.
- 2. Muestre que $P(y=1\mid \mathbf{z}) = \Phi\left[\mathbf{z}_1\boldsymbol{\delta}_1/\left(1+\gamma_1^2z_2^2\right)^{1/2}\right]$.
- 3. Defina $\rho_1 \equiv \gamma_1^2.$ ¿Cómo testearía la hipótesis $H_0: \rho_1 = 0$?
- 4. Si tuviera motivos para creer que $\rho_1 > 0$, ¿cómo estimaría δ_1 junto con ρ_1 ?
- Ejercicio 3. Probit con endogeneidad.

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

- 1. Explique el experimento subyacente de interés cuando queremos examinar los efectos del tabaquismo en los días de trabajo perdidos.
- 2. ¿Por qué $cigs_i$ podría estar correlacionada con variables no observables que afectan a $sick_i$?
- 3. Una forma de escribir el modelo de interés es

$$P(\text{ sick } = 1 \mid \mathbf{z}, \text{ cigs }, q_1) = \Phi(\mathbf{z}_1 \boldsymbol{\delta}_1 + \gamma_1 \text{ cigs } + q_1)$$

donde \mathbf{z}_1 es un subconjunto de \mathbf{z} y q_1 es una variable no observable que posiblemente esté correlacionada con cigs. ¿Qué sucede si se ignora q_1 y se estima el probit de sick sobre \mathbf{z}_1 y cigs?

- 4. ¿Puede cigs tener una distribución normal condicional en la población? Explique.
- 5. Explique cómo probar si *cigs* es exógeno. ¿Esta prueba se basa en *cigs* que tienen una distribución normal condicional?
- 6. Suponga 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?

Microeconometría I Problem Set 2 - Stata Extensiones de Modelos Logit y Probit

- Ejercicio 1. Utilice el conjunto de datos BWGHT.RAW para este problema.
 - 1. Defina una variable binaria, smokes, si la mujer fuma durante el embarazo. Estime 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 uno con 12 años de educación?
 - 2. ¿Cree que faminc es exógena en la ecuación de smokes? ¿Qué pasa con motheduc?
 - 3. Suponga que motheduc y white son exógenos en el probit de la parte 1. Suponga también que fatheduc es exógeno a esta ecuación. Estime la forma reducida de $\log(faminc)$ para ver si fatheduc está parcialmente correlacionada $\log(faminc)$.
 - 4. Contraste la hipótesis nula de que log(faminc) es exógenA en el probit del inciso 1.

■ Ejercicio 2. Precios endógenos o exógenos

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).

- 1. ¿Por qué podrían ser endógenos los precios autoinformados?
- 2. Realice 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.
- 3. En función de los resultados, estime la elasticidad de la prevalencia del tabaquismo con respecto a los precios.
- Ejercicio 3. Probit heterocedástico y simulaciones

Buscamos simular el siguiente modelo:

$$Pr(y = 1) = F\{(\beta_0 + \beta_1 x) / \exp(\gamma_1 x_{het})\}\$$

Genere un dataset vacío con 1000 obsevaciones. Genere las siguientes variables:

$$\begin{aligned} x &\sim \mathbb{U}(-1, 1) \\ x_{het} &\sim \mathbb{U}(0, 1) \\ \sigma &\sim e^{1, 5 \cdot x_{het}} \\ p &\sim \mathcal{N}\left(\frac{\beta_0 + \beta_1 \cdot x}{\sigma}\right) \end{aligned}$$

con $\beta_0 = 0.3$ y $\beta_1 = 2$ y defina 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. Estime el modelo probit heterocedástico y compare con las estimaciones del probit usual.

Microeconometría I Problem Set 2 - Stata Modelos para Variables Categóricas No Ordenadas

- Ejercicio 1. Alternativas de pesca. La variable dependiente y toma el valor 1, 2, 3 o 4 dependiendo de cuál de los cuatro modos alternativos de pesca, respectivamente, playa, muelle, barco privado y barco chárter, se elija. En la base de datos, estos son beach, pier, private o charter. Los datos provienen de J. A. Herriges and C. L. Kling, "Nonlinear Income Effects in Random Utility Models", Review of Economics and Statistics, 81(1999): 62-72.
 - Abra la base y describa las categorías.
 - Estime un modelo logit multinomial.
 - Estime un modelo logit condicional.
- Ejercicio 2. Predicción de calificaciones de clientes

Net Promoter Score®, o NPS®, mide la experiencia del cliente y predice el crecimiento del negocio. Es utilizada por empresas que brindan servicios al consumidor final (bancos, telefónicas, etc). EL NPS se calcula usando la respuesta a una pregunta usando una escala de 0 a 10: ¿Qué tan probable es que recomiende a un amigo o colega? Los encuestados se agrupan de la siguiente manera:

- Los promotores (puntuación 9-10) son entusiastas leales que seguirán comprando y recomendarán a otros, lo que impulsará el crecimiento.
- Los neutrales (puntuación 7-8) son clientes satisfechos pero poco entusiastas que son vulnerables a las ofertas de la competencia.
- Los detractores (puntuación 0-6) son clientes insatisfechos que pueden dañar su marca e impedir el crecimiento a través del boca a boca negativo.

Al restar el porcentaje de detractores del porcentaje de promotores, se obtiene el puntaje neto del promotor, que puede oscilar entre un mínimo de -100 (si todos los clientes son detractores) y un máximo de 100 (si todos los clientes son promotores). Estas encuestas se utilizan para generar estrategias de originacion (nuevos clientes) y de reducción de *churn* (fuga de clientes). La base con la que se va a hacer la primera parte de la practica consiste en la encuesta de NPS que se le hace a los clientes de un Banco luego de efectuar una transacción en caja. En base a esto, utilizando la base NPS.dta responda las siguientes preguntas. L

- 1. Abra y describa la base.
- 2. Genere una variable que clasifique a los clintes en función de si son promotores, detractores o neutrales.
- 3. Analice cómo cambia la variable de espera en función de la clasificación de los clientes.
- 4. Tome una muestra del 10 % de los datos. Estime un logit multinomial para predecir cómo cambian las clasificaciones en función de la espera, condicionando en explicativas que considere relevantes.
- 5. Calcule los efectos marginales.
- 6. Repita el análisis con un probit multinomial y compare.
- 7. Realice un test de la significatividad de las variables.
- Ejercicio 3. Alternativas de pesca.
 - 1. Utilizando la EPH del cuarto trimestre de 2016, estime un modelo multinomial que le permita predecir la condición de actividad de una persona, entre inactivo, ocupado o desocupado.

La tabla 1 muestra los comandos de Stata utilizados para estos modelos.

Tabla 1: Modelos y comandos en Stata

Modelo	Comando
Logit multinomial	mlogit
Logit condicional	asclogit
Probit multinomial	mprobit
Logit ordenado	ologit
Probit ordenado	oprobit

Microeconometría I Problem Set 4 - Stata Modelos para Variables Categóricas Ordenadas

- Ejercicio 1. Predicción de calificaciones de clientes Considere el ejercicio del Problem Set anterior con el mismo título que este. Repita el análisis utilizando un modelo ordenado.
- Ejercicio 2. Modelo secuencial
 - (a) Considere la base de datos nlsw88.dta. En la misma hay datos de un grupo de mujeres de entre 30 y 40 años para estudiar los patrones de la fuerza laboral. Estime un logit secuencial con la decisión de educación utilizando el comando seqlogit y muestre que puede obtener los mismos resultados estimando varios modelos logit por separado.
 - (b) Considere la base de datos gss.dta. La misma posee datos de la encuesta GSS (General Social Survey). Esta encuesta realiza investigaciones científicas básicas sobre la estructura y el desarrollo de la sociedad estadounidense con un programa de recopilación de datos diseñado tanto para monitorear el cambio social dentro de los Estados Unidos como para comparar a los Estados Unidos con otras naciones. Iniciado en 1972, el GSS contiene un "núcleo. estándar de preguntas demográficas, de comportamiento y de actitud, además de temas de especial interés. Muchas de las preguntas centrales se han mantenido sin cambios desde 1972 para facilitar los estudios de tendencias temporales, así como la replicación de hallazgos anteriores. En este ejercicio utilizamos datos de educación similares a los de la pregunta anterior.

Estime un logit secuencial, interprete los resultados y muestre el efecto de la educación del padre en las decisiones de educación en cada transición.

La tabla 1 muestra los comandos de Stata utilizados para estos modelos.

Tabla 1: Modelos y comandos en Stata

Modelo	Comando
Logit multinomial	mlogit
Logit condicional	asclogit
Probit multinomial	mprobit
Logit ordenado	ologit
Probit ordenado	oprobit

Microeconometría I Problem Set 5 - Stata Modelos para Variables Dependientes Limitadas - Tobit

■ Ejercicio 1. Variables Censuradas: Modelo Tobit I

El modelo Tobit es relevante cuando la variable dependiente y de una regresión lineal se observa solo en algún intervalo de su soporte, porque en este caso los estimadores de MCC no son consistentes.

(a) Considere la base auto.dta. Estime el modelo

$$\mathsf{mpg} = \alpha + \beta \mathsf{wgt} + u$$

donde wgt=weight/1000. Luego estime el modelo generando una variable censurada suponiendo que no se observan autos con $mpg \le 17$. Estime por MCC y utilizando un modelo Tobit. Compare.

- (b) Repita el inciso anterior suponiendo que ahora no se observan autos con $mpg \geq 24$.
- (c) ¿Cómo se interpretan los coeficientes del modelo? Compute los efectos marginales.

■ Ejercicio 2. Variables Censuradas: Modelo Tobit II

El siguiente ejercicio está tomado de Cameron & Trivedi. La variable dependiente para el gasto ambulatorio (ambulatory expenditure, ambexp) y los regresores (age, female, educ, blhisp, totchr, and ins) se obtienen de la encuesta Medical Expenditure Panel Survey de 2001

- (a) Abra y describa la base mus16datav2. ¿Qué puede decir sobre el cumplimiento de las condiciones que requiere un Tobit?
- (b) Compute los efectos marginales.
- (c) Compute los efectos marginales haciendo las cuentas con los comandos de escalares y matrices de Stata.
- (d) Considere la variable dependiente en logaritmos. ¿Qué interpretación tiene esto sobre la variable dependiente? ¿Qué complicaciones introduce en el análisis? Estime un tobit para el logaritmo de ambexp.
- Ejercicio 3. Variables Censuradas: Modelo Tobit III Considere la base de datos mroz.dta, que posee datos que permiten estudiar la oferta laboral anual de mujeres casadas. Considere las horas trabajadas, hours, y las explicativas nwifeinc, educ, exper, expersq, age, kidslt6, kidsge6. Estime un modelo lineal y un modelo Tobit. Compare. Compute los efectos marginales.

Microeconometría I Problem Set 6 - Stata Modelos para Variables Dependientes Limitadas - Heckman

■ Ejercicio 1. Gastos ambulatorios

Retome la base de datos del Ejercicio 2 del Problem Set 5. Ahora estimaremos un modelo de dos partes de Heckman. Estos modelos sirven para muestras autoseleccionadas. Se modela explícitamente la ecuación que determina la selección y la ecuación de interés. En este ejercicio se pide estimar un modelo de Heckman para los gastos ambulatorios y comparar con las predicciones de un modelo Tobit.

■ Ejercicio 2. Ecuación salarial para las mujeres I

Considere la base de datos womenwk. Describa la base. Estime una ecuación salarial en función de la educación y la edad por Mínimos Cuadrados Clásicos. Repita utilizando un modelo de Heckman, utilizando las variables married, children, educ, age para la ecuación de selección. Utilice el comando heckman.

■ Ejercicio 2. Ecuación salarial para las mujeres I

Conceptualmente, vamos a repetir el ejercicio anterior utilizando la base de datos mroz.dta que ya hemos utilizado. Ahora se pide modelar explícitamente la ecuación de selección con un probit y la ecuación estructural con un modelo lineal aumentada por la inversa del ratio del Mills. Reporte el efecto marginal sobre las horas trabajadas correctamente estimado.