Modelado Estocástico

Clase 10: Probit y Logit

1. Método de Máxima Verosimilitud:

Con nuestro supuesto 3 de errores normales con media cero y varianza σ^2, cada error u_i tienen como pdf:

$$f(u_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}{u_i}^2}$$

Si los suponemos independientes, la pdf del conjunto de errores (*joint pdf*, en inglés) es el producto (multiplicación) de las pdf de cada uno de ellos:

$$f(u_1, \dots, u_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} u_i^2}$$

Sea $y_i = \alpha + \beta x_i + u_i$. El valor de la muestra de y_i depende de los valores desconocidos de los parámetros. Su pdf será

$$f(y_1, \dots, y_n | x_1, \dots, x_n; \alpha, \beta) = f(u_1, \dots, u_n; \alpha, \beta)$$

Si tomamos logaritmos tenemos la "*log likelihood function*", que nos quedaría:

$$l = \ln\big(f(y|x; \alpha, \beta)\big) = \ln\big(f(u; \alpha, \beta)\big) =$$

$$l = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} u_i^2$$

Y acá esta todo el truco: reemplazamos $u_i = y_i - \alpha - \beta x_i$ y nos preguntamos cuál debiera ser el valor de los estimadores de α y β que maximizan la verosimilitud de haber observado los valores de la muestra que efectivamente tenemos. Es decir,

$$l(\alpha, \beta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

¿Cuál va a ser el estimador por Máxima Verosimilitud de cada uno de los parámetros?

2. Modelos con Variable Dependiente Dicotómica

Supongamos que la variable dependiente, "y_i", toma los valores 0 ó 1. Por ejemplo, $y_i = 1$ si la persona está empleada, $y_i = 0$ si no está empleada. U otro ejemplo, $y_i = 1$ si la persona realiza una donación, $y_i = 0$ si no. Independientemente de a qué se defina por el evento $y_i = 1$, cuando $y_i = 1$ se lo suele llamar éxito y al caso en que $y_i = 0$ se lo llama fracaso.

Nosotros queremos explicar a la variable y_i en función a una o más variables explicativas x_i. En general, el interés está en la probabilidad de que $y_i = 1$. A esto se lo llama la probabilidad de respuesta:

$$p(\boldsymbol{x}) \equiv P(y = 1|\boldsymbol{x}) = P(y = 1|x_1, x_2, \dots, x_k)$$

Para una variable explicativa x_j continua, el efecto sobre la probabilidad de respuesta de un cambio en x_j es

$$\frac{\partial P(y = 1|\boldsymbol{x})}{\partial x_j} = \frac{\partial p(\boldsymbol{x})}{\partial x_j}$$

2.1. Modelo de Probabilidad Lineal

$$P(y = 1|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Notemos que en este caso, $\beta_1 = \frac{\partial p(\boldsymbol{x})}{\partial x_1}$, es decir, el parámetro β_1 mide el cambio en la probabilidad de éxito de un aumento en una unidad en x_1.

El problema que tiene el modelo de probabilidad lineal es que, dados ciertos valores de las variables explicativas, la probabilidad de respuesta predicha podría caer afuera del intervalo [0,1].

Ejemplo del libro de Wooldridge, Econometric Analysis of Cross Section and Panel Data

Example 15.1 (Married Women's Labor Force Participation): We use the data from MROZ.RAW to estimate a linear probability model for labor force participation (*inlf*) of married women. Of the 753 women in the sample, 428 report working non-zero hours during the year. The variables we use to explain labor force participation are age, education, experience, nonwife income in thousands (*nwifeinc*), number of children less than six years of age (*kidslt6*), and number of kids between 6 and 18

inclusive (*kidsge6*); 606 women report having no young children, while 118 report having exactly one young child. The usual OLS standard errors are in parentheses, while the heteroskedasticity-robust standard errors are in brackets:

$$\widehat{inlf} = \underset{\substack{(.154)\\[.151]}}{.586} - \underset{\substack{(.0014)\\[.0015]}}{.0034}\, nwifeinc + \underset{\substack{(.007)\\[.007]}}{.038}\, educ + \underset{\substack{(.006)\\[.006]}}{.039}\, exper - \underset{\substack{(.00018)\\[.00019]}}{.00060}\, exper^2$$

$$- \underset{\substack{(.002)\\[.002]}}{.016}\, age - \underset{\substack{(.034)\\[.032]}}{.262}\, kidslt6 + \underset{\substack{(.013)\\[.013]}}{.013}\, kidsge6$$

$$N = 753, \qquad R^2 = .264$$

With the exception of *kidsge6*, all coefficients have sensible signs and are statistically significant; *kidsge6* is neither statistically significant nor practically important. The coefficient on *nwifeinc* means that if nonwife income increases by 10 ($10,000), the probability of being in the labor force is predicted to fall by .034. This is a small effect given that an increase in income by $10,000 in 1975 dollars is very large in this sample. (The average of *nwifeinc* is about $20,129 with standard deviation $11,635.) Having one more small child is estimated to reduce the probability of *inlf* = 1 by about .262, which is a fairly large effect.

Of the 753 fitted probabilities, 33 are outside the unit interval. Rather than using some adjustment to those 33 fitted values and applying weighted least squares, we just use OLS and report heteroskedasticity-robust standard errors. Interestingly, these differ in practically unimportant ways from the usual OLS standard errors.

2.2 Modelos Probit y Logit

$$P(y = 1|\boldsymbol{x}) = G(\boldsymbol{x\beta}) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) = p(\boldsymbol{x})$$

donde $\boldsymbol{x}$ es de 1 x k y $\boldsymbol{\beta}$ es de k x 1.

Al término $\boldsymbol{x\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ se lo suele llamar "*index*" y a estos modelos se los llama "*index models*".

$p(\boldsymbol{x})$ es una función de $\boldsymbol{x}$ a través de

$$\boldsymbol{x\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

G es una cdf (la normal o la logística). ("cdf" significa función de distribución acumulada)

Sea

$$y^* = \boldsymbol{x\beta} + e$$

donde $y = 1_{[y^*>0]}$.

Aclaración: $1_{[y^*>0]}$ es una función que toma valor igual a 1 si $y^* > 0$, y toma valor igual a cero si no, o sea si $y^* \leq 0$.

A estos modelos se los llama modelos de variables latente.

$$P(y = 1|\boldsymbol{x}) = P(y^* > 0|\boldsymbol{x}) = P(e > -\boldsymbol{x\beta}) = 1 - G(-\boldsymbol{x\beta}) = G(\boldsymbol{x\beta})$$

En el modelo Probit,

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)dv$$

Notar que $g(0) \approx 0{,}4$ donde $g(z) = \frac{d}{dz}G(z)$ es la pdf ("pdf" quiere decir "*probability density function*" o función de densidad) de una variable aleatoria normal estándar. Es decir,

$\phi(v)$ es la pdf de la normal estándar, o sea, $\phi(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}$

En el modelo Logit,

$$G(z) = \Lambda(z) = \frac{e^z}{1 + e^z}$$

Notar que $g(0) = 0{,}25$

Ahora el cambio en la probabilidad de éxito de un aumento en una unidad en x_j no es simplemente el beta como en el modelo de probabilidad lineal, sino que es función de todos los valores de $\boldsymbol{x}$, como puede verse en la expresión a continuación:

$$\frac{\partial p(\boldsymbol{x})}{\partial x_j} = g(\boldsymbol{x\beta})\beta_j$$

Sin embargo, es posible calcular el ratio o cociente del cambio en la probabilidad de éxito de un aumento en una unidad en x_j con respecto al cambio en la probabilidad de éxito de un aumento en una unidad en x_h, y de esta manera nos desentendemos de los valores de $\boldsymbol{x}$, ya que se cancelan y nos queda una expresión que es función de los betas solamente.

$$\frac{\frac{\partial p(\boldsymbol{x})}{\partial x_j}}{\frac{\partial p(\boldsymbol{x})}{\partial x_h}} = \frac{g(\boldsymbol{x\beta})\beta_j}{g(\boldsymbol{x\beta})\beta_h} = \frac{\beta_j}{\beta_h}$$

En resumen, para comparar los beta sombrero de probit y logit con los del modelo de probabilidad lineal general (MPL), hay que dividir el $\hat{\beta}_{i,probit}$ por 2.5 para compararlo con el $\hat{\beta}_{i,MPL}$ y, por otro lado, el $\hat{\beta}_{i,logit}$ hay que dividirlo por 4 para compararlo con el $\hat{\beta}_{i,MPL}$ (y por lo tanto el $\hat{\beta}_{i,logit}$ hay que dividirlo por 1.6 para compararlo con el $\hat{\beta}_{i,probit}$).

Table 15.1
LPM, Logit, and Probit Estimates of Labor Force Participation

Dependent Variable: *inlf*			
Independent Variable	LPM (OLS)	Logit (MLE)	Probit (MLE)
nwifeinc	−.0034	−.021	−.012
	(.0015)	(.008)	(.005)
educ	.038	.221	.131
	(.007)	(.043)	(.025)
exper	.039	.206	.123
	(.006)	(.032)	(.019)
$exper^2$	−.00060	−.0032	−.0019
	(.00019)	(.0010)	(.0006)
age	−.016	−.088	−.053
	(.002)	(.015)	(.008)
kidslt6	−.262	−1.443	−.868
	(.032)	(0.204)	(.119)
kidsge6	.013	.060	.036
	(.013)	(.075)	(.043)
constant	.586	.425	.270
	(.151)	(.860)	(.509)
Number of observations	753	753	753
Percent correctly predicted	73.4	73.6	73.4
Log-likelihood value	—	−401.77	−401.30
Pseudo *R*-squared	.264	.220	.221

Example 15.2 (Married Women's Labor Force Participation): We now estimate logit and probit models for women's labor force participation. For comparison we report the linear probability estimates. The results, with standard errors in parentheses, are given in Table 15.1 (for the LPM, these are heteroskedasticity-robust).

The estimates from the three models tell a consistent story. The signs of the coefficients are the same across models, and the same variables are statistically significant in each model. The pseudo *R*-squared for the LPM is just the usual *R*-squared reported for OLS; for logit and probit the pseudo *R*-squared is the measure based on the log likelihoods described previously. In terms of overall percent correctly predicted, the models do equally well. For the probit model, it correctly predicts "out of the labor force" about 63.1 percent of the time, and it correctly predicts "in the labor force" about 81.3 percent of the time. The LPM has the same overall percent correctly predicted, but there are slight differences within each outcome.

As we emphasized earlier, the *magnitudes* of the coefficients are not directly comparable across the models. Using the rough rule of thumb discussed earlier, we can

divide the logit estimates by four and the probit estimates by 2.5 to make all estimates comparable to the LPM estimates. For example, for the coefficients on *kidslt6*, the scaled logit estimate is about −.361, and the scaled probit estimate is about −.347. These are larger in magnitude than the LPM estimate (for reasons we will soon discuss). The scaled coefficient on *educ* is .055 for logit and .052 for probit.

If we evaluate the standard normal probability density function, $\phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)$, at the average values of the independent variables in the sample (including the average of $exper^2$), we obtain about .391; this value is close enough to .4 to make the rough rule of thumb for scaling the probit coefficients useful in obtaining the effects on the response probability. In other words, to estimate the change in the response probability given a one-unit increase in any independent variable, we multiply the corresponding probit coefficient by .4.

The biggest difference between the LPM model on one hand, and the logit and probit models on the other, is that the LPM assumes *constant* marginal effects for *educ*, *kidslt6*, and so on, while the logit and probit models imply diminishing marginal magnitudes of the partial effects. In the LPM, one more small child is estimated to reduce the probability of labor force participation by about .262, regardless of how many young children the woman already has (and regardless of the levels of the other dependent variables). We can contrast this finding with the estimated marginal effect from probit. For concreteness, take a woman with $nwifeinc = 20.13$, $educ = 12.3$, $exper = 10.6$, $age = 42.5$—which are roughly the sample averages—and $kidsge6 = 1$. What is the estimated fall in the probability of working in going from zero to one small child? We evaluate the standard normal cdf, $\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)$ with $kidslt6 = 1$ and $kidslt6 = 0$, and the other independent variables set at the values given. We get roughly $.373 - .707 = -.334$, which means that the labor force participation probability is about .334 lower when a woman has one young child. This is not much different from the scaled probit coefficient of −.347. If the woman goes from one to two young children, the probability falls even more, but the marginal effect is not as large: $.117 - .373 = -.256$. Interestingly, the estimate from the linear probability model, which we think can provide a good estimate near the average values of the covariates, is in fact between the probit estimated partial effects starting from zero and one children.

2.3 Estimación por Máxima Verosimilitud

Para obtener el estimador por máxima verosimilitud de los parámetros que multiplican a las variables explicativas es necesario plantear la función de verosimilitud. La función de densidad de y_i dado $\boldsymbol{x_i}$ es

$$f(y_i|\boldsymbol{x_i};\ \boldsymbol{\beta}) = G(\boldsymbol{x_i\beta})^{y_i}(1 - G(\boldsymbol{x_i\beta}))^{1-y_i}, y_i = 0,1,$$

ya que cuando $y_i = 1$, $f(y_i|\boldsymbol{x_i};\ \boldsymbol{\beta}) = G(\boldsymbol{x_i\beta})^1(1 - G(\boldsymbol{x_i\beta}))^0 = p$, y cuando $y_i = 0$, $f(y_i|\boldsymbol{x_i};\ \boldsymbol{\beta}) = G(\boldsymbol{x_i\beta})^0(1 - G(\boldsymbol{x_i\beta}))^1 = 1 - p$.

Cuando planteamos la densidad anterior como función de $\boldsymbol{\beta}$, dados y_i, $\boldsymbol{x_i}$ es lo que llamamos la función de verosimilitud y si tomamos el logaritmo de la función de verosimilitud de y_i dado $\boldsymbol{x_i}$ obtenemos la función de "*log likelihood*" que sería:

$$l_i(\boldsymbol{\beta}) = y_i \log\big(G(\boldsymbol{x_i\beta})\big) + (1 - y_i) \log\big(1 - G(\boldsymbol{x_i\beta})\big)$$

La densidad conjunta de $y_1, \ldots, y_n$ es lo que llamamos la función de verosimilitud y es el producto de las densidades individuales. Mientras que la función de log likelihood en definitiva va a ser la sumatoria de éstas, es decir, $\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta})$. El estimador $\widehat{\boldsymbol{\beta}}$ por el método de máxima verosimilitud de $\boldsymbol{\beta}$ es el estimador que maximiza la función de verosimilitud (o equivalentemente, la de *log likelihood*). Si $G(\cdot)$ es la función logística estándar, entonces $\widehat{\boldsymbol{\beta}}$ es el estimador logit, mientras que si $G(\cdot)$ es la función normal estándar, entonces $\widehat{\boldsymbol{\beta}}$ es el estimador probit. La naturaleza no lineal del problema de maximización nos impide obtener una fórmula para $\widehat{\boldsymbol{\beta}}$ en ambos casos. Pero por las propiedades del método de máxima verosimilitud, los estimadores van a ser consistentes, asintóticamente normales y asintóticamente eficientes.

2.4 Test de hipótesis

Para testar una hipótesis nula que involucre a varios beta, por ejemplo,

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_A: al\ menos\ uno\ distinto\ de\ cero$$

utilizamos un test de "*likelihood ratio*" (LR), que consiste en plantear que el estadístico es $LR = 2(\mathcal{L}_U - \mathcal{L}_R)$, donde $\mathcal{L}_U$ y $\mathcal{L}_R$ son las funciones de *log likelihood* de los modelos no restringido (*unrestricted*) y restringido, respectivamente. El modelo restringido es el modelo cuando imponemos la hipótesis nula. LR tiene una distribución asintótica chi-cuadrado con una cantidad de grados de libertad igual a la cantidad de restricciones bajo la hipótesis nula. El valor crítico, por lo tanto, será el percentil 95 (99) de la distribución chi-cuadrado con q grados de libertad si usamos un nivel de significancia de 5% (1%), es decir, rechazamos H_0 en la cola derecha.

2.5 Pseudo R^2

Como medida de bondad de ajuste del modelo suele usarse el estadístico

$$1 - \mathcal{L}_U / \mathcal{L}_0$$

donde $\mathcal{L}_U$ y $\mathcal{L}_R$ es la función de *log likelihood* del modelo estimado (*unrestricted*) y $\mathcal{L}_0$ es la función de *log likelihood* de un modelo solamente con una constante (es decir, todos los beta son simultáneamente iguales a cero).

2.6 Stata

Con el comando `probit` o `logit` Stata presenta los coeficientes de los estimadores, es decir, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$, y no el cambio en la probabilidad de respuesta, $\frac{\partial p(\boldsymbol{x})}{\partial x_j} = g(\boldsymbol{x}\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$, es decir, los efectos marginales. Con el comando `dprobit`

Stata presenta los efectos marginales evaluados en las medias de las variables explicativas.

Nota al pie: recordemos que $p(\boldsymbol{x}) = P(y = 1|\boldsymbol{x}) = G(\boldsymbol{x\beta})$.

Si x_j es una variable continua, el efecto parcial sobre $p(\boldsymbol{x})$ de un cambio en una unidad de x_j es, como vimos, $\frac{\partial p(\boldsymbol{x})}{\partial x_j} = g(\boldsymbol{x\beta})\beta_j$.

En cambio, si x_1 es una variable *dummy*, el efecto sobre $p(\boldsymbol{x})$ de un cambio de x_1 de 0 a 1 es

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \cdots + \beta_k x_k)$$