Modelos para variables binarias

Randall Romero Aguilar, PhD
randall.romero@ucr.ac.cr

EC4300 - Microeconometría

II Semestre 2023

Última actualización: 19 de septiembre de 2023

UCR
UNIVERSIDAD DE COSTA RICA

ESCUELA de ECONOMÍA
UNIVERSIDAD DE COSTA RICA

Tabla de contenidos

1. Introducción

2. Los modelos para variables binarias

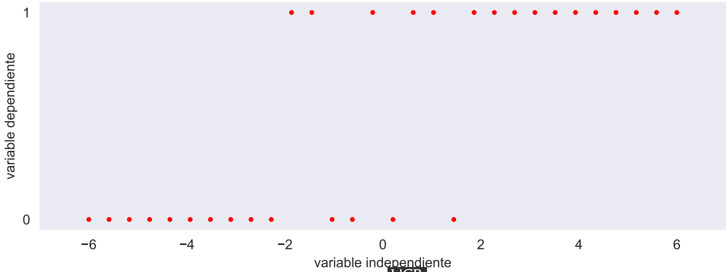
3. Interpretación

4. Discusión sobre modelos de resultados binarios

1. Introducción

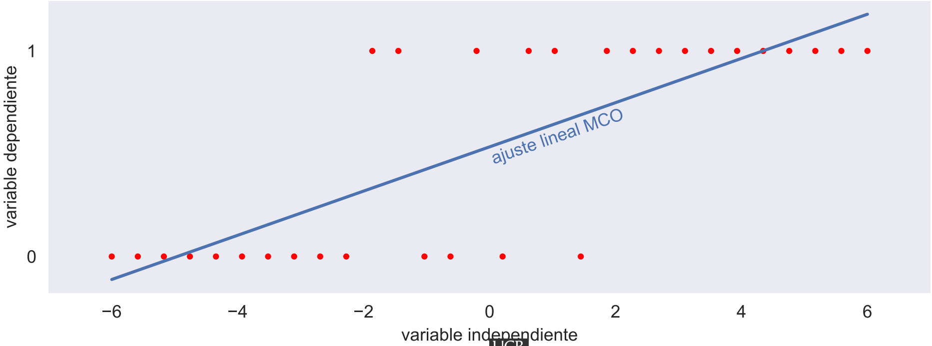
Modelos de regresión con variables dependiente binaria

- En ocasiones, deseamos estimar una regresión en un modelo en que la variable dependiente solo puede tener uno de dos valores posibles. Algunos ejemplos:
  - si un hogar tiene una casa propia $(y = 1)$ o no $(y = 0)$.
  - si un individuo participa en el mercado laboral $(y = 1)$ o no $(y = 0)$.
  - si una persona compra o no un seguro para su vehículo.
  - si una persona asiste o no a un evento (concierto, partido de fútbol).
  - si un votante emite su voto o no.
  - si un deudor (individuo, empresa, país) paga o no su deuda en la fecha establecida.
  - si ocurre o no una crisis macroeconómica (bancaria, balanza de pagos).



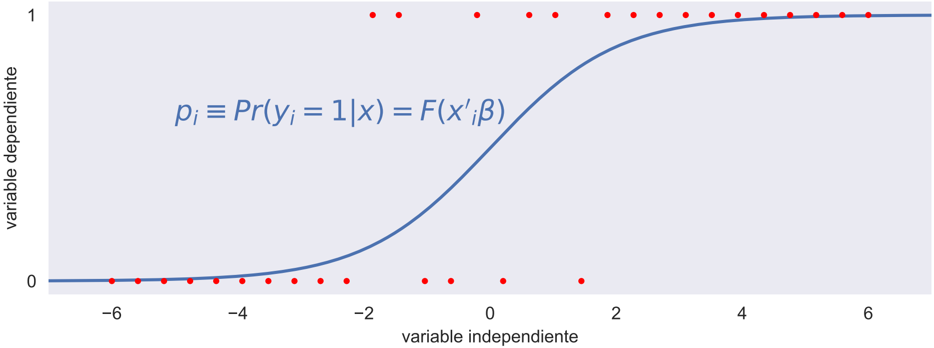
Problemas del modelo de regresión lineal para variables binarias

- Si estimamos un modelo de variable binaria con mínimos cuadrados ordinarios, vemos que el valor ajustado de y no está restringido al intervalo [0, 1], y mucho menos al conjunto $\{0{,}1\}$.
- Bajo los supuestos del MCRL, la variable y tiene distribución condicional $N(X\beta, \sigma^2 I)$, pero en nuestro caso actual es claro que y tiene una distribución Bernoulli.



Enfoque para estimación de variables binarias

- En vez de ajustar una recta a los datos, donde x determina el valor de y, vamos a ajustar una función de distribución acumulada, donde x determina la probabilidad de obtener un valor y dado.



Motivación

- El análisis de regresión de una variable binaria o dicotómica es un problema común en la econometría.
- Los modelos para resultados mutuamente excluyentes se enfocan en determinar la probabilidad p de que ocurra uno de los resultados en vez del otro (que ocurre con probabilidad $1-p$).
- Estos modelos difieren únicamente en la función de distribución que utilizan para estimar la probabilidad p.

2. Los modelos para variables binarias

El modelo básico

- La variable dependiente es una respuesta binaria
- Toma dos valores: 0 y 1.

$$y = \begin{cases} 1, & \text{con probabilidad } p \\ 0, & \text{con probabilidad } 1 - p \end{cases}$$

- No hay pérdida de generalidad en fijar los valores $y \in \{0, 1\}$.
- La distribución de y es de Bernoulli:
  - función de probabilidad: $f(y) = p^y(1-p)^{1-y}$.
  - esperanza $\mathbb{E}(y) = p$.
  - varianza $\mathbb{V}(y) = p(1-p)$.

Una regresión (no lineal) para variables binarias

- Una regresión se forma al parametrizar la probabilidad p como dependiente de un índice $\mathbf{x}'\beta$, donde x es un vector de variables independientes y β un vector de coeficientes desconocidos.
- La probabilidad condicional tiene la forma

$$p_i \equiv \mathbb{P}[y_i = 1 \mid \mathbf{x}] = F\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$$

  donde $F(\boldsymbol{.})$ es una función de distribución.
- Entonces
  - $\mathbb{E}(y_i|\,\mathbf{x}_i) = F\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$.
  - $\mathbb{V}(y_i|\,\mathbf{x}_i) = F\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\left[1 - F\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]$.

Efectos marginales

- Una complicación de modelar la probabilidad como una función no lineal es que el efecto de cambiar un regresor en una unidad (efecto marginal) ya no es simplemente el coeficiente de ese regresor.
- En particular:
$$\frac{\partial p_i}{\partial x_k} = f(\mathbf{x}_i'\boldsymbol{\beta})\beta_k$$
donde $f(\cdot)$ es la función de densidad de probabilidad (derivada de F).
- Es decir, ahora el efecto marginal de x_k depende de los valores del vector $\mathbf{x}_i'$.

Algunos de los modelos

- Los modelos difieren en la escogencia de F.
- En la práctica, los dos más comunes son los modelos logit y probit, los cuales son simétricos.
- El log-log complementario es asimétrico, se recomienda cuando hay una elevada proporción de 0s o 1s en la muestra.

Modelo	Probabilidad $p = \mathbb{P}(y = 1 \mid \mathbf{x})$	Efecto marginal $\partial p / \partial x_j$
Logit	$\Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right) = e^{\mathbf{x}'\boldsymbol{\beta}} / \left(1 + e^{\mathbf{x}'\boldsymbol{\beta}}\right)$	$\Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right)\left\{1 - \Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right)\right\}\beta_j$
Probit	$\Phi\left(\mathbf{x}'\boldsymbol{\beta}\right) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(z)dz$	$\phi\left(\mathbf{x}'\boldsymbol{\beta}\right)\beta_j$
Log-log complementario	$C\left(\mathbf{x}'\boldsymbol{\beta}\right) = 1 - \exp\left\{-\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)\right\}$	$\exp\left\{-\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)\right\}\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)\beta_j$
Prob. lineal	$F\left(\mathbf{x}'\boldsymbol{\beta}\right) = \mathbf{x}'\boldsymbol{\beta}$	β_j

Modelo logit

- Para el modelo logit, $\mathrm{F}\left(\mathbf{x}'\beta\right)$ es la CDF de la distribución logística.

$$\mathrm{F}\left(\mathbf{x}'\boldsymbol{\beta}\right) = \Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}} = \frac{\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)}$$

- Las probabilidades pronosticadas están limitadas entre 0 y 1.

Modelo probit

- Para el modelo probit, $\mathrm{F}\left(\mathbf{x}'\boldsymbol{\beta}\right)$ es la CDF de la distribución normal estándar.

$$\mathrm{F}\left(\mathbf{x}'\boldsymbol{\beta}\right) = \Phi\left(\mathbf{x}'\boldsymbol{\beta}\right) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(z)\,\mathrm{d}z$$

- De nuevo, las probabilidades pronosticadas están limitadas entre 0 y 1.

Modelo de regresión (modelo de probabilidad lineal)

- En el modelo de probabilidad lineal, $\mathrm{F}\left(\mathbf{x}'\boldsymbol{\beta}\right) = \mathbf{x}'\boldsymbol{\beta}$

$$p = \mathbb{P}[y = 1 \mid \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$$

- Como mencionamos anteriormente, un problema con el modelo de regresión es que las probabilidades predichas no se limitarán entre 0 y 1.
- Por ello, en la práctica rara vez utilizamos el modelo de regresión con datos de respuesta binaria.

Estimación de los modelos

- El estimador de máxima verosimilitud es el más apropiado para estos modelos, porque sabemos con certeza la distribución de la variable y (Bernoulli).
- La función de probabilidad de una observación es $p_i^{y_i}(1-p_i)^{1-y_i}$, donde $p_i = F(\mathbf{x}_i'\boldsymbol{\beta})$.
- Para una muestra de N observaciones independientes, el estimador de máxima verosimilitud maximiza la función

$$
\begin{aligned}
Q(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left[y_i \ln p_i + (1-y_i)\ln\{1-p_i\}\right] \\
&= \sum_{i=1}^{N} \left[y_i \ln F\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + (1-y_i)\ln\left\{1-F\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right\}\right]
\end{aligned}
$$

Score: condiciones de primer orden

- Tomando la derivada respecto al vector $\boldsymbol{\beta}$ tenemos

$$\sum_{i=1}^{N} \frac{f(\mathbf{x}_i' \boldsymbol{\beta})}{F(\mathbf{x}_i' \boldsymbol{\beta})[1 - F(\mathbf{x}_i' \boldsymbol{\beta})]} \left(y_i - F(\mathbf{x}_i' \boldsymbol{\beta})\right) \mathbf{x}_i = 0$$

- Puesto que no hay una solución analítica para esta ecuación, debemos utilizar métodos numéricos (Newton-Raphson, por ejemplo) para encontrar el valor de $\boldsymbol{\beta}$.
- En el caso especial del modelo logit, lo anterior se simplifica un poco (aunque tampoco tiene solución analítica)

$$\sum_{i=1}^{N} \left(y_i - F(\mathbf{x}_i' \boldsymbol{\beta})\right) \mathbf{x}_i = 0$$

3. Interpretación

Modelos de variable latente

- Los modelos de variable binaria pueden interpretarse como de variable latente, aunque no es necesario.
- Distinguimos entre la variable observada y, y una variable continua no observable subyacente (latente) y^*, que satisface $y^* = \mathbf{x}'\boldsymbol{\beta} + u$.
- Aunque no observamos y^*, sí observamos que

$$y = \begin{cases} 1 & \text{si } y^* > 0 \\ 0 & \text{si } y^* \leq 0 \end{cases}$$

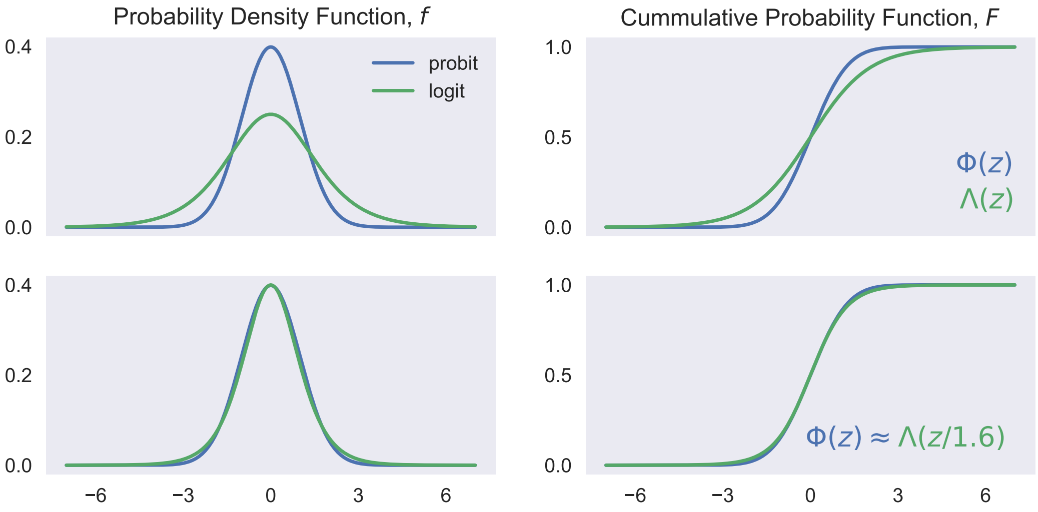
- Entonces tenemos que

$$\begin{aligned} \mathbb{P}(y = 1) &= \mathbb{P}\left(\mathbf{x}'\boldsymbol{\beta} + u > 0\right) \\ &= \mathbb{P}\left(u > -\mathbf{x}'\boldsymbol{\beta}\right) \\ &= 1 - \mathbb{P}\left(u \leq -\mathbf{x}'\boldsymbol{\beta}\right) \\ &= \mathbb{P}\left(u < \mathbf{x}'\boldsymbol{\beta}\right) && \text{(asumiendo simetría)} \\ &= F\left(\mathbf{x}'\boldsymbol{\beta}\right) \end{aligned}$$

Interpretación de coeficientes

- Un aumento en x_k aumenta/disminuye la probabilidad de que $\mathbf{y} = 1$ (hace que ese resultado sea más/menos probable). En otras palabras, un aumento en x_k hace que el resultado de 1 sea más o menos probable.
- Interpretamos el signo del coeficiente pero no la magnitud.
- La magnitud no se puede interpretar utilizando el coeficiente porque diferentes modelos tienen diferentes escalas de coeficientes.

Logit vs probit



Comparación de coeficientes

- Los coeficientes difieren entre los modelos debido a la forma funcional de la función F.

$$\hat{\beta}^{\text{logit}} \simeq 4.0\hat{\beta}^{\text{OLS}}$$
$$\hat{\beta}^{\text{probit}} \simeq 2.5\hat{\beta}^{\text{OLS}}$$
$$\hat{\beta}^{\text{logit}} \simeq 1.6\hat{\beta}^{\text{probit}}$$

- No debemos comparar la magnitud de los coeficientes entre diferentes modelos.
- A pesar de lo anterior, los efectos marginales y los valores pronosticados por los modelos logit y probit sí son muy similares.

Efectos marginales

- Al estimar los modelos probit y logit, es común informar los efectos marginales después de informar los coeficientes.
- Los efectos marginales reflejan el cambio en la probabilidad de $y = 1$ dado un cambio de 1 unidad en una variable independiente $\mathbf{x}$.
- Para los modelos logit y probit, los efectos marginales se calculan como:

$$\frac{\partial p}{\partial \mathbf{x}_j} = \mathrm{f}\left(\mathbf{x}'\boldsymbol{\beta}\right)\beta_j = \begin{cases} \Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right)\left[1 - \Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right)\right]\beta_j & \text{(logit)} \\ \phi\left(\mathbf{x}'\boldsymbol{\beta}\right)\beta_j & \text{(probit)} \end{cases}$$

- Los coeficientes y los efectos marginales tienen los mismos signos porque $\mathrm{F}'\left(\mathbf{x}'\beta\right) > 0$.
- Los efectos marginales dependen de $\mathbf{x}$, por lo que necesitamos estimar los efectos marginales para un valor específico de $\mathbf{x}$ (típicamente la media).

Estimación de efectos marginales 1: Efectos marginales en la media

- Los efectos marginales se estiman para la observación promedio en la muestra: $\bar{\mathbf{x}}$.

$$\frac{\partial p}{\partial\, \mathbf{x}_j} = \mathrm{f}\left(\bar{\mathbf{x}}'\boldsymbol{\beta}\right) \beta_j$$

- La mayoría de los artículos publicados informan efectos marginales en la media.
- Un problema es que puede que no haya una "persona" (observación) así en la muestra.

Estimación de efectos marginales 2: Promedio de efectos marginales

- Los efectos marginales se estiman como el promedio de los efectos marginales individuales.
$$\frac{\partial p}{\partial \mathbf{x}_j} = \frac{\sum_{i=1}^{N} \mathrm{f}\left(\mathbf{x}_i' \boldsymbol{\beta}\right)}{N} \beta_j$$
- Este es un mejor enfoque para estimar los efectos marginales.
- En la práctica, las dos formas de estimar los efectos marginales producen resultados casi idénticos la mayor parte del tiempo.

Efectos parciales para variables discretas

- Para variables explicativas discretas d, no tiene sentido calcular el efecto marginal con una derivada.
- Separemos $\mathbf{x} = [\mathbf{z} \quad d]$ y sea $\mathbb{P}(y = 1|\mathbf{z}, d) = F(\mathbf{z}'\boldsymbol{\beta}_1 + \beta_2 d)$
- Calculamos el efecto marginal de la variable discreta con diferencia finita:

$$
\begin{aligned}
\mathrm{ME}_j &= \Pr(y = 1 \mid \mathbf{z}, d = 1) - \Pr(y = 1 \mid \mathbf{z}, d = 0) \\
&= F\left(\mathbf{z}'\boldsymbol{\beta}_1 + \beta_2\right) - F\left(\mathbf{z}'\boldsymbol{\beta}_1\right)
\end{aligned}
$$

Interpretación de los efectos marginales

- Un aumento en x_k aumenta (disminuye) la probabilidad de que $\mathbf{y} = 1$ por el efecto marginal expresado como un porcentaje.
  - Para las variables independientes discretas, el efecto marginal se expresa en comparación con la categoría base $(\mathbf{x}_k = 0)$.
  - Para las variables independientes continuas, el efecto marginal se expresa para un cambio de una unidad en $\mathbf{x}_k$.
- Interpretamos tanto el signo como la magnitud de los efectos marginales.
- Los modelos probit y logit producen efectos marginales casi idénticos.

Razón de probabilidades (*odds ratio*) para el modelo logit

- La razón de probabilidad o el riesgo relativo es $p/(1-p)$ y mide la probabilidad de que $y = 1$ en relación con la probabilidad de que $\mathbf{y} = 0$.

$$
\begin{aligned}
p &= \frac{\exp\left(\mathbf{x}'\,\beta\right)}{1+\exp\left(\mathbf{x}'\,\beta\right)} \\
\frac{p}{1-p} &= \exp\left(\mathbf{x}'\,\beta\right) \\
\ln\frac{p}{1-p} &= \mathbf{x}'\,\beta
\end{aligned}
$$

- Una razón de probabilidad de 2 significa que el resultado $y = 1$ es el doble de probable que el resultado $y = 0$.
- Los *odds ratios* se estiman con el modelo logístico.
- En economía, es más usual reportar efectos marginales que *odds ratios*.

Probabilidades pronosticadas y medidas de bondad de ajuste

- Después de estimar los modelos, podemos pronosticar la probabilidad de que $\mathbf{y} = 1$ para cada observación.

$$\hat{p} = \mathbb{P}[y = 1 \mid \mathbf{x}] = \mathrm{F}\left(\mathbf{x}'\,\hat{\beta}\right)$$

- Para el modelo de regresión lineal, las probabilidades pronosticadas no se limitan al intervalo [0, 1].
- Para los modelos logit y probit, las probabilidades pronosticadas sí están entre 0 y 1.
- Usualmente, si la probabilidad pronosticada es mayor que 0.5, pronosticamos que $y = 1$, de lo contrario $y = 0$.

Medidas de bondad de ajuste: % de valores pronosticados correctamente

- Sea n_{ij} el número de observaciones en la muestra en los que $y = j$ y se pronostica que $\hat{y} = i$.
- Podemos crear la siguiente matriz de confusión:

	$\mathbf{y} = 1$	$\mathbf{y} = 0$	Total
$\hat{y} = 1$	n_{11}	n_{10}	$n_{11} + n_{10}$
$\hat{y} = 0$	n_{01}	n_{00}	$n_{01} + n_{00}$
Total	$n_{11} + n_{01}$	$n_{10} + n_{00}$	

- Tenemos cuatro casos de $0/1$: dos de ellos son predicciones correctas y dos de ellos son predicciones incorrectas.
- El porcentaje de valores pronosticados correctamente es la proporción de pronósticos verdaderos a pronósticos totales:

$$\frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}} \times 100\%$$

Medidas de bondad de ajuste: Sensibilidad y especificidad

	$\mathbf{y} = 1$	$\mathbf{y} = 0$
$\hat{y} = 1$	n_{11}	n_{10}
$\hat{y} = 0$	n_{01}	n_{00}

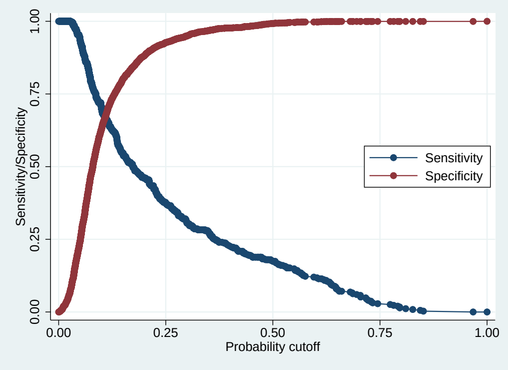
Sensibilidad

Porcentaje de eventos $y = 1$ pronosticados correctamente$= \frac{n_{11}}{n_{11}+n_{01}} \times 100\%$

Especificidad

Porcentaje de eventos $y = 0$ pronosticados correctamente$= \frac{n_{00}}{n_{00}+n_{10}} \times 100\%$

Ambas dependen del valor de probabilidad a partir del cual pronosticamos que $y = 1$.



Ejemplo de matriz de confusión

$x_i'\hat{\beta}$	$\hat{p}_i = F\left(x_i'\hat{\beta}\right)$	Pronosticado $\hat{y}_i$	Observado y_i	
-0.55	0.63	1	0	✗
0.43	0.39	0	0	✓
1.47	0.19	0	0	✓
-2.24	0.90	1	1	✓
-0.55	0.63	1	0	✗
-2.31	0.91	1	1	✓
0.09	0.48	0	0	✓
0.39	0.40	0	0	✓
0.99	0.27	0	1	✗
2.35	0.09	0	1	✗
-0.47	0.62	1	1	✓
1.82	0.14	0	0	✓
-0.46	0.61	1	1	✓
-1.38	0.80	1	0	✗
0.76	0.32	0	0	✓
0.94	0.28	0	0	✓
-1.91	0.87	1	0	✗
-1.67	0.84	1	1	✓
-1.96	0.88	1	1	✓
0.93	0.28	0	1	✗

y_i

$\hat{y}_i$		1	0	Total
	1	6	4	10
	0	3	7	10
	Total	9	11	20

Porcentaje de valores pronosticados correctamente

(6 + 7) / 20 = 65%

Sensibilidad = % de casos y=1 pronosticados correctamente

6 / 9 = 67%

Especificidad = % de casos y=0 pronosticados correctamente

7 / 11 = 64%

$\hat{p}_i = F(x_i'\hat{\beta})$	$\hat{y}_i$ Pronosticado 0.25	0.5	0.75	Observado y_i
0.63	1	1	0	0
0.39	1	0	0	0
0.19	0	0	0	0
0.90	1	1	1	1
0.63	1	1	0	0
0.91	1	1	1	1
0.48	1	0	0	0
0.40	1	0	0	0
0.27	1	0	0	1
0.09	0	0	0	1
0.62	1	1	0	1
0.14	0	0	0	0
0.61	1	1	0	1
0.80	1	1	1	0
0.32	1	0	0	0
0.28	1	0	0	0
0.87	1	1	1	0
0.84	1	1	1	1
0.88	1	1	1	1
0.28	1	0	0	1
# de 1s =	17	10	6	9

	0.25	y_i 1	0	Total
$\hat{y}_i$	1	8	9	17
	0	1	2	3
	Total	9	11	20

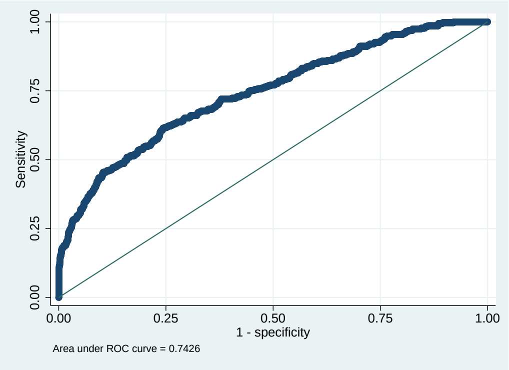
	0.50	y_i 1	0	Total
$\hat{y}_i$	1	6	4	10
	0	3	7	10
	Total	9	11	20

	0.75	y_i 1	0	Total
$\hat{y}_i$	1	4	2	6
	0	5	9	14
	Total	9	11	20

p crítico	0.25	0.5	0.75
Sensibilidad	8 / 9 = 89%	6 / 9 = 67%	4 / 9 = 44%
Especificidad	2 / 11 = 18%	7 / 11 = 64%	9 / 11 = 82%

La curva ROC

- La curva *receiver operating characteristic* describe las combinaciones de sensibilidad vs 1 - especificidad conforme cambia el punto de quiebre c: es decir, $\mathbb{P}(y=1) > c \Rightarrow \hat{y} = 1$.
- Un modelo que no tiene capacidad predictiva describe una línea de 45°.
- Entre más se curve hacia arriba la ROC, más poder explicativo tiene el modelo.
- Así, el área bajo la curva es una medida de bondad del ajuste: si es 0.5 no tiene poder explicativo, si es 1.0 pronostica perfectamente.



Pseudo-R cuadrado

- El pseudo r-cuadrado se calcula como:

$$\mathrm{R}^2 = 1 - \frac{L_{ur}}{L_r}$$

- Compara la probabilidad log-verosimilitud sin restricciones L_{ur} para el modelo que estamos estimando y la log-verosimilitud restringida L_r con solo un intercepto.
- Si las variables independientes no tienen potencia explicativa, el modelo restringido será el mismo que el modelo sin restricciones y R-cuadrado será 0.

4. Discusión sobre modelos de resultados binarios

Elección entre el modelo logit y probit

- La elección depende del proceso de generación de datos, el cual es desconocido.
- Los modelos producen resultados casi idénticos (diferentes coeficientes pero efectos marginales similares).

Codificación de la variable dependiente

- Si revertimos las categorías 0 y 1, los signos de los coeficientes se invierten (positivos se vuelven negativos y viceversa) pero las magnitudes son las mismas.

Ejemplo 1:

Visita al médico

- `doctor-visit.ipynb`
- `mus10data.dta`

- Muestra de 2002 de 4412 personas en EE.UU. (Medical Expenditure Panel Survey).
- Variables:

visit	Variable dependiente: ¿visitó la persona a un médico durante el año?
private	¿tiene seguro médico?
chronic	¿tiene una enfermedad crónica?
female	¿es mujer?
income	ingreso en miles de dólares.

```
/* Definir variables */
global exogenas private chronic female income
global endogena visit

/* Preparar datos */
use mus10data
quietly keep if year02 == 1
generate visit = docvis > 0
label variable visit "=  1 si visitó al médico"

keep $endogena $exogenas
summarize
```

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      income |      4,412     34.3402     29.0399   -49.9990   280.7770
      female |      4,412      0.4719      0.4993     0.0000     1.0000
     private |      4,412      0.7854      0.4106     0.0000     1.0000
     chronic |      4,412      0.3264      0.4689     0.0000     1.0000
       visit |      4,412      0.6360      0.4812     0.0000     1.0000
```

```
/* Modelo probit */
probit $endogena $exogenas, vce(robust)
```

```
Iteration 0:   log pseudolikelihood =    -2892.9
Iteration 1:   log pseudolikelihood = -2337.6553
Iteration 2:   log pseudolikelihood = -2331.8213
Iteration 3:   log pseudolikelihood = -2331.8084
Iteration 4:   log pseudolikelihood = -2331.8084

Probit regression                                       Number of obs =  4,412
                                                        Wald chi2(4)  = 910.77
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -2331.8084                       Pseudo R2     = 0.1940

-----------------------------------------------------
             |               Robust
       visit | Coefficient  std. err.      z    P>|z|
-------------+---------------------------------------
     private |     0.7663     0.0528   14.51   0.000
     chronic |     1.0645     0.0511   20.82   0.000
      female |     0.5530     0.0434   12.74   0.000
      income |     0.0056     0.0009    6.48   0.000
       _cons |    -0.9594     0.0526  -18.24   0.000
-----------------------------------------------------
```

Parámetros estimados de modelos LOGIT y PROBIT para visitas al médico

	visit_probit		visit_logit		visit_ols	
$=1$ si tiene seguro	0.766	***	1.273	***	0.265	***
$=1$ si tiene enfermedad crónica	1.064	***	1.832	***	0.306	***
$=1$ si es mujer	0.553	***	0.928	***	0.170	***
ingreso en miles de US$	0.006	***	0.010	***	0.002	***
Number of observations	4412		4412		4412	

*** p<.01, ** p<.05, * p<.1

$$\hat{\beta}^{\text{logit}} \simeq 4.0\hat{\beta}^{\text{OLS}}$$
$$\hat{\beta}^{\text{probit}} \simeq 2.5\hat{\beta}^{\text{OLS}}$$
$$\hat{\beta}^{\text{logit}} \simeq 1.6\hat{\beta}^{\text{probit}}$$

```
/* Derivadas en la media */
margins, dydx(*) atmeans
```

```
Conditional marginal effects
Model VCE: Robust

Expression: Pr(visit), predict()
dy/dx wrt:  private chronic female income
At: private = .7853581 (mean)
    chronic = .3263826 (mean)
    female  = .4718948 (mean)
    income  = 34.34018 (mean)

-----------------------------------------------
         |           Delta-method
         |   dy/dx std. err.       z    P>|z|
---------+-------------------------------------
 private |  0.2771    0.0192    14.40   0.000
 chronic |  0.3849    0.0179    21.45   0.000
  female |  0.1999    0.0156    12.79   0.000
  income |  0.0020    0.0003     6.49   0.000
-----------------------------------------------
```

```
/* Promedio de derivadas */
margins, dydx(*)
```

```
Average marginal effects
Model VCE: Robust

Expression: Pr(visit), predict()
dy/dx wrt:  private chronic female income

-----------------------------------------------
         |           Delta-method
         |   dy/dx std. err.       z    P>|z|
---------+-------------------------------------
 private |  0.2291    0.0147    15.62   0.000
 chronic |  0.3183    0.0132    24.12   0.000
  female |  0.1653    0.0123    13.46   0.000
  income |  0.0017    0.0003     6.57   0.000
-----------------------------------------------
```

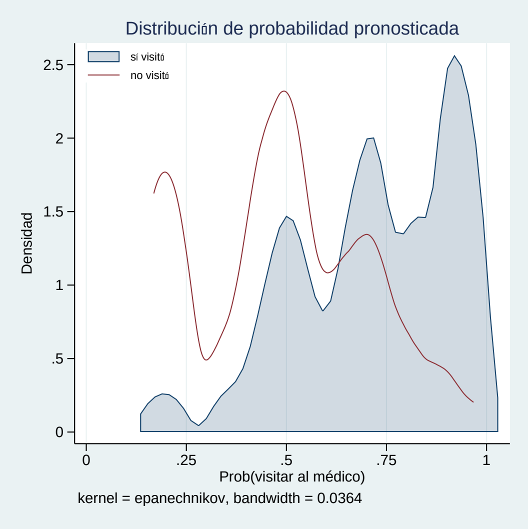
Efectos marginales

	probit_atmeans	probit	logit_atmeans	logit	ols
$=1$ si tiene seguro	0.277	0.229	0.277	0.226	0.265
$=1$ si tiene enfermedad crónica	0.385	0.318	0.398	0.325	0.306
$=1$ si es mujer	0.200	0.165	0.202	0.165	0.170
ingreso en miles de US$	0.002	0.002	0.002	0.002	0.002
Number of observations	4412	4412	4412	4412	4412

```
/* Pronósticos del modelo */
predict phatprobit, pr
summarize $endogena phatprobit
```

```
    Variable |    Mean  Std. dev.     Min      Max
-------------+------------------------------------
       visit | 0.6360    0.4812  0.0000  1.0000
  phatprobit | 0.6357    0.2317  0.1687  0.9928
   phatlogit | 0.6360    0.2325  0.1669  0.9849
     phatols | 0.6360    0.2290  0.1922  1.2168
```

```
kdensity phatprobit if visit==1,
     recast(area) addplot((
     kdensity phatprobit if visit
     ==0))
```



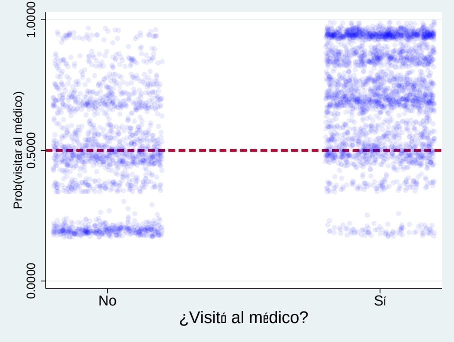
```
estat classification
```

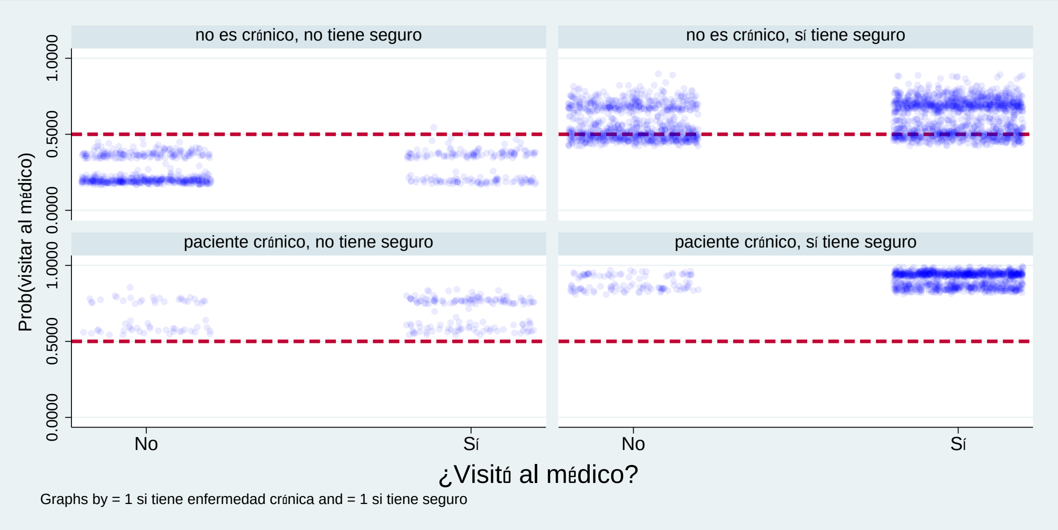
```
              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      2327           729  |       3056
     -     |       479           877  |       1356
-----------+--------------------------+-----------
   Total   |      2806          1606  |       4412

Classified + if predicted Pr(D) >= .5
True D defined as visit != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   82.93%
Specificity                     Pr( -|~D)   54.61%
Positive predictive value       Pr( D| +)   76.15%
Negative predictive value       Pr(~D| -)   64.68%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   45.39%
False - rate for true D         Pr( -| D)   17.07%
False + rate for classified +   Pr(~D| +)   23.85%
False - rate for classified -   Pr( D| -)   35.32%
--------------------------------------------------
Correctly classified                        72.62%
--------------------------------------------------
```

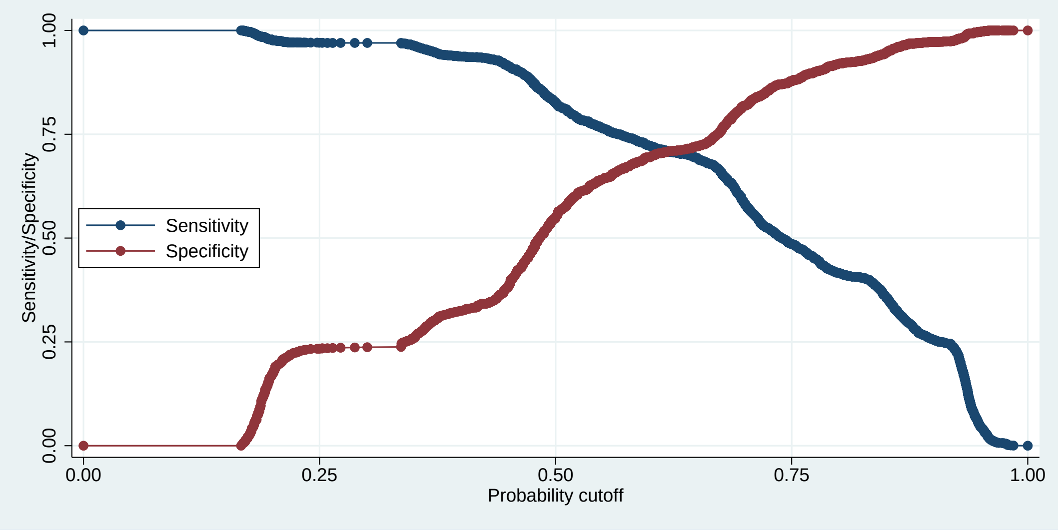
```
generate visitnoise = visit +
    runiform(-0.1,0.1)

twoway (scatter phatprobit
    visitnoise, mcolor(blue%8)),
     yline(0.5)
```



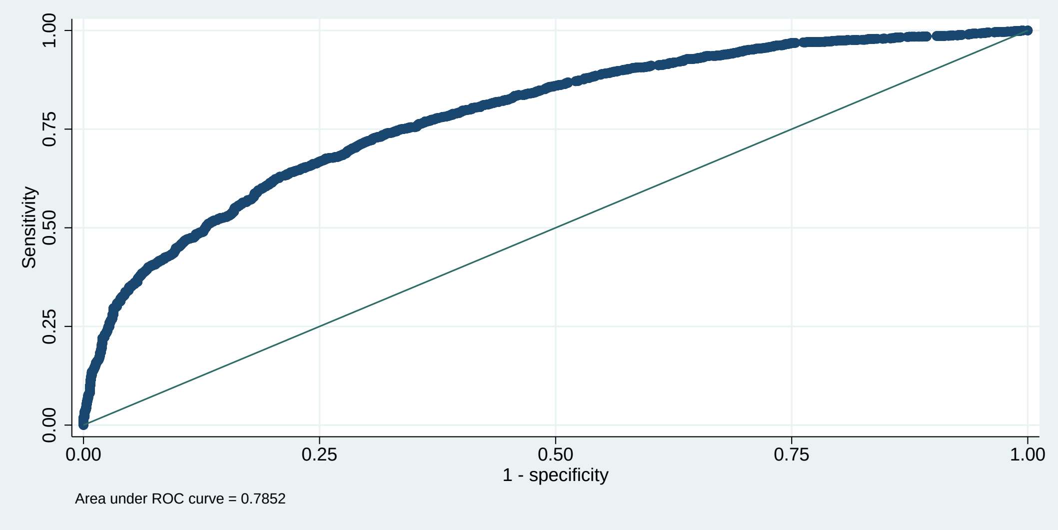

no es crónico, no tiene seguro
no es crónico, sí tiene seguro
paciente crónico, no tiene seguro
paciente crónico, sí tiene seguro
Prob(visitar al médico)
0.0000
0.5000
1.0000
No
Sí
¿Visitó al médico?
Graphs by = 1 si tiene enfermedad crónica and = 1 si tiene seguro

lsens



Sensitivity/Specificity
1.00
0.75
0.50
0.25
0.00
Sensitivity
Specificity
0.00
0.25
0.50
0.75
1.00
Probability cutoff

`lroc`


Sensitivity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1 - specificity
Area under ROC curve = 0.7852

Ejemplo 2:
Crisis bancarias y de balanza de pagos

- `ejemplo-crisis.ipynb`
- `crisis-binary.xlsx`

Variable	Obs	Mean	Std. Dev.	Min	Max
crisisbank	3024	0.1157	0.3200	0.0000	1.0000
crisisbop	3024	0.2189	0.4136	0.0000	1.0000
Dolarización (nivel)	3023	0.0852	0.0900	0.0000	0.7979
Dolarización (Δ%)	2939	0.0181	0.0815	-0.6645	0.6328
Crédito doméstico / PIB	2785	0.0263	0.2516	-1.0000	2.3342
Exportaciones (Δ%)	2940	0.1391	0.3336	-0.8529	3.4702
Importaciones (Δ%)	2940	0.1472	0.3390	-0.7616	2.3421
Inflación (nivel)	2898	0.0413	0.1144	-0.0968	3.9696
Multiplicador dinero (Δ%)	2929	0.0411	0.2840	-0.8667	3.3476
M2 / Reservas (Δ%)	2940	0.0657	0.5398	-0.9478	7.9688
Déficit comercial (nivel)	2947	-0.1138	0.4279	-2.0877	0.6573
Depósitos domésticos (Δ%)	2821	0.0812	0.3036	-0.8503	3.3052
Tipo de cambio real (desv. tendencia)	2905	0.0000	31.3367	-53.6713	343.6609
Reservas (Δ%)	2940	0.2654	0.8081	-0.9285	10.8824
Interés EEUU (nivel)	3017	6.2596	3.1325	0.8608	17.2877

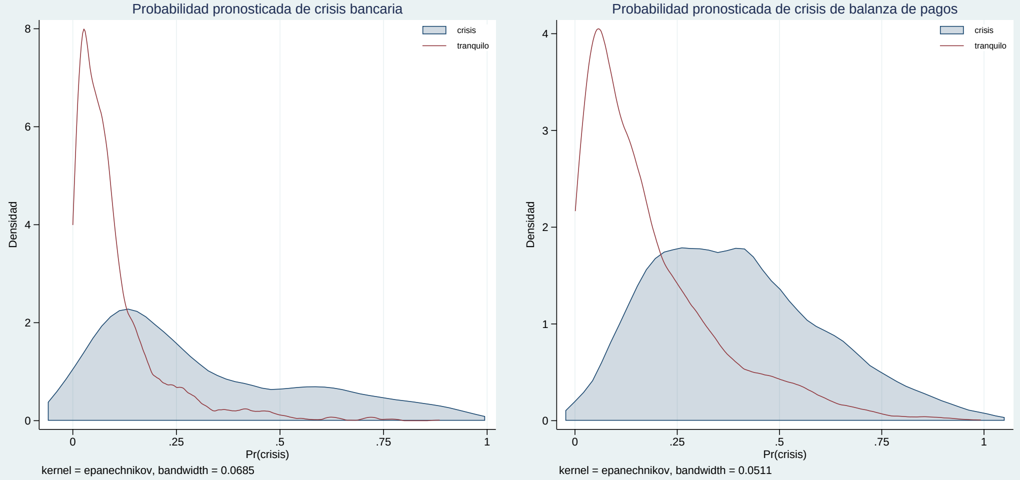
Efectos marginales

	bank_atmeans	bank	bop_atmeans	bop
Dolarización (nivel)	0.388 ***	0.487 ***	-1.344 ***	-1.332 ***
Dolarización (Δ%)	-0.035	-0.043	0.241 **	0.239 **
Crédito doméstico / PIB	0.141 ***	0.177 ***	0.238 ***	0.236 ***
Exportaciones (Δ%)	-0.012	-0.014	-0.047 *	-0.047 *
Importaciones (Δ%)	-0.105 ***	-0.132 ***	-0.070 ***	-0.070 ***
Inflación (nivel)	-0.126 **	-0.158 **	0.284	0.281
Multiplicador dinero (Δ%)	0.009	0.012	0.049 **	0.049 **
M2 / Reservas (Δ%)	0.039 ***	0.049 ***	0.138 ***	0.137 ***
Déficit comercial (nivel)	-0.080 ***	-0.101 ***	0.029	0.029
Depósitos domésticos (Δ%)	-0.068 ***	-0.085 ***	-0.169 ***	-0.168 ***
Tipo de cambio real (desv.)	-0.000	-0.000	-0.000	-0.000
Reservas (Δ%)	-0.005	-0.007	0.029 **	0.029 **
Interés EEUU (nivel)	0.007 ***	0.009 ***	0.018 ***	0.018 ***

*** p<.01, ** p<.05, * p<.1

Probabilidad pronosticada de crisis

Variable	Obs	Mean	Std. dev.	Min	Max
crisisbank	3,024	0.1157	0.3200	0.0000	1.0000
phatbank	2,773	0.1262	0.1516	0.0000	0.9266
crisisbop	3,024	0.2189	0.4136	0.0000	1.0000
phatbop	2,773	0.2322	0.1985	0.0008	0.9993


Probabilidad pronosticada de crisis bancaria
Densidad
Pr(crisis)
crisis
tranquilo
0
2
4
6
8
0
.25
.5
.75
1
kernel = epanechnikov, bandwidth = 0.0685
Probabilidad pronosticada de crisis de balanza de pagos
Densidad
Pr(crisis)
crisis
tranquilo
0
1
2
3
4
0
.25
.5
.75
1
kernel = epanechnikov, bandwidth = 0.0511

Logistic model for banking crisis

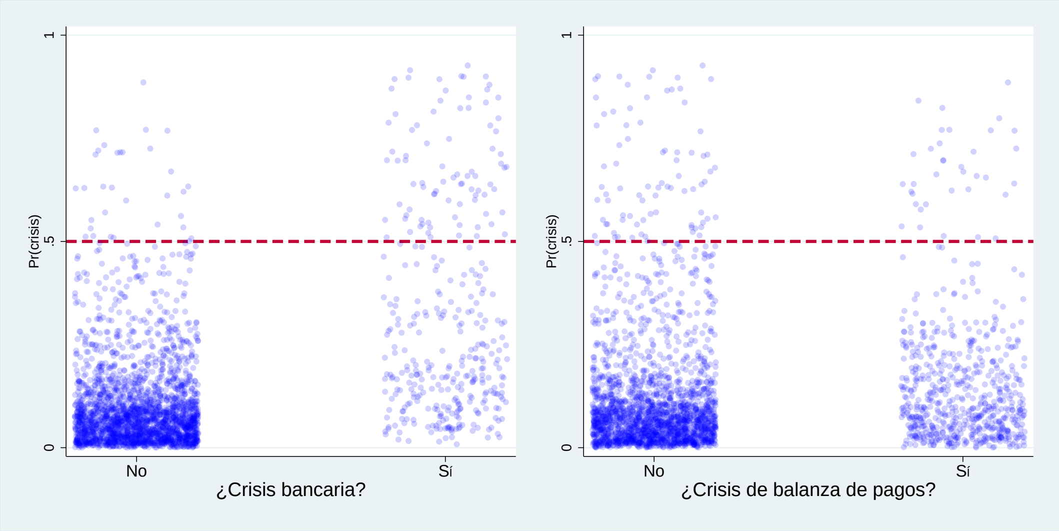
```
           -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        85            32  |        117
     -     |       265          2391  |       2656
-----------+--------------------------+-----------
   Total   |       350          2423  |       2773

Classified + if predicted Pr(D) >= .5
True D defined as crisis != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   24.29%
Specificity                     Pr( -|~D)   98.68%
Positive predictive value       Pr( D| +)   72.65%
Negative predictive value       Pr(~D| -)   90.02%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    1.32%
False - rate for true D         Pr( -| D)   75.71%
False + rate for classified +   Pr(~D| +)   27.35%
False - rate for classified -   Pr( D| -)    9.98%
--------------------------------------------------
Correctly classified                        89.29%
--------------------------------------------------
```

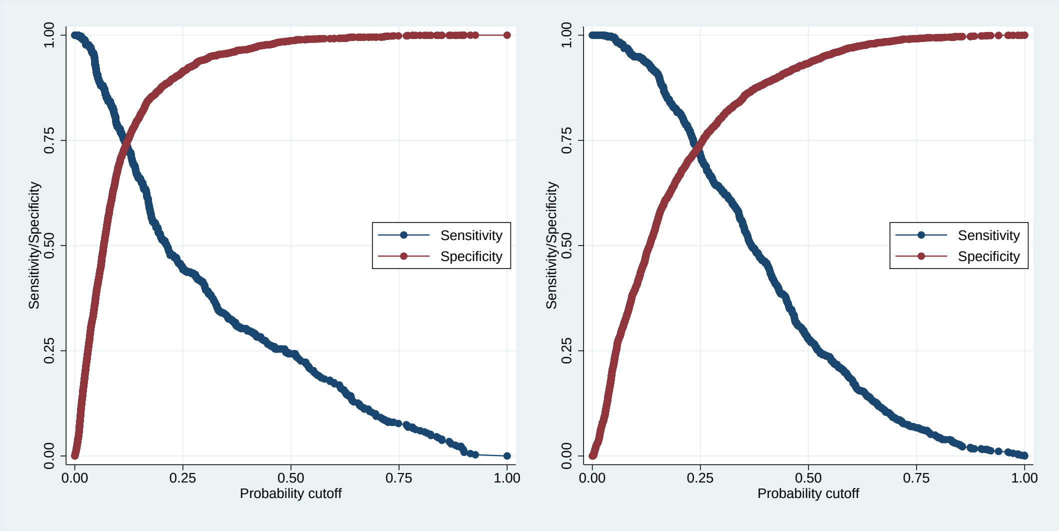
Logistic model for balance of payments crisis

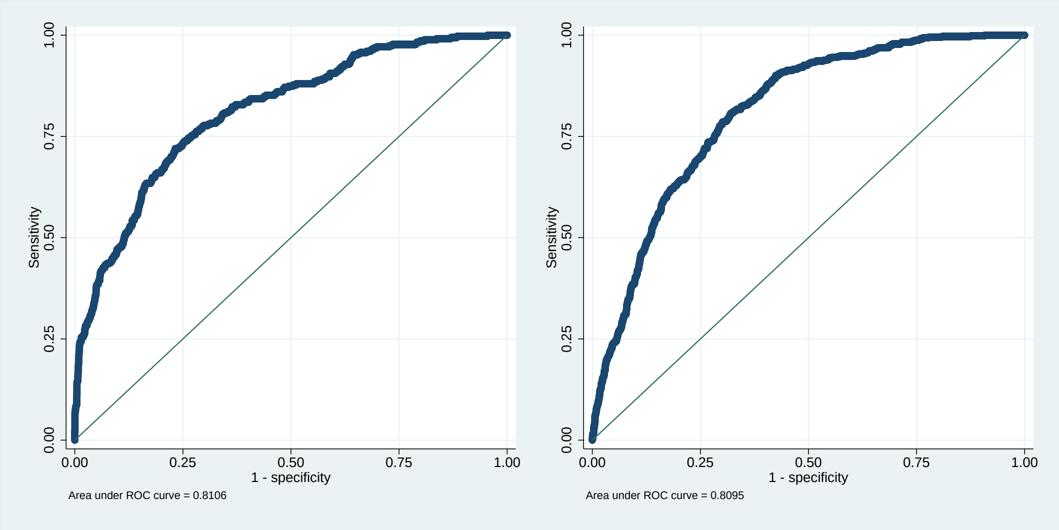
```
           -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       179           142  |        321
     -     |       465          1987  |       2452
-----------+--------------------------+-----------
   Total   |       644          2129  |       2773

Classified + if predicted Pr(D) >= .5
True D defined as crisis != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   27.80%
Specificity                     Pr( -|~D)   93.33%
Positive predictive value       Pr( D| +)   55.76%
Negative predictive value       Pr(~D| -)   81.04%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    6.67%
False - rate for true D         Pr( -| D)   72.20%
False + rate for classified +   Pr(~D| +)   44.24%
False - rate for classified -   Pr( D| -)   18.96%
--------------------------------------------------
Correctly classified                        78.11%
--------------------------------------------------
```



Pr(crisis)
1
.5
0
No
Sí
¿Crisis bancaria?
Pr(crisis)
1
.5
0
No
Sí
¿Crisis de balanza de pagos?


Sensitivity/Specificity
0.00
0.25
0.50
0.75
1.00
Probability cutoff
Sensitivity
Specificity
Sensitivity/Specificity
0.00
0.25
0.50
0.75
1.00
Probability cutoff
Sensitivity
Specificity


Sensitivity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1 - specificity
Area under ROC curve = 0.8106
Sensitivity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1 - specificity
Area under ROC curve = 0.8095

Referencias I

- Cameron, A. Colin y Pravin K. Trivedi (2022). *Microeconometrics Using Stata*. 2ª ed. Stata Press.
- Hansen, Bruce E. (2022). *Econometrics*. Princeton University Press.
- Wooldridge, Jeffrey M. (2010). *Econometric Analysis of Cross Section and Panel Data*. 2ª ed.