Probit y Logit Binomial

Fundamentos de Econometría

Juan Palomino[1]

[1]Magister en Economía Aplicada con Mención Estudios Regionales
juan.palominoh@pucp.pe

Departamento de Economía

Índice

Modelos de Variable Dependiente Binaria

- La variable dependiente es la siguiente:

$$y_i = \begin{cases} 1 & si\ algún\ evento\ ocurre \\ 0 & si\ el\ evento\ no\ ocurre \end{cases}$$

- Algunos ejemplos son:
  - ¿Ha migrado usted en los últimos cinco años?
  - ¿Es un consumidor más probable a comprar la misma marca o intentar con una nueva?
  - ¿Usted se ha enfermedad en las últimas dos semanas?

Variable Dependiente Binaria

- Entonces, la pregunta es: **¿Cómo estimar un modelo con variable dependiente binaria?**

El primer enfoque es aplicar MCO como si fuera la **variable dependiente continua**.

¿Qué pasa si estimamos por MCO... ?

- El **modelo de probabilidad lineal** es el modelo de regresión lineal para una variable dependiente binaria. El modelo estructural es:

$$y_i = x_i^{'}\beta + \varepsilon_i$$

donde:

$$y_i = \begin{cases} 1 & si\ algún\ evento\ ocurre \\ 0 & si\ el\ evento\ no\ ocurre \end{cases}$$

- Cuando y es una variable aleatoria binaria, entonces:

$$E(y_i|x_i) = [1 \times Pr(y_i = 1|x_i)] + [0 \times Pr(y_i = 0|x_i)] = Pr(y_i = 1|x_i)] = x_i^{'}\beta$$

Modelos de Probabilidad Lineal (MPL)

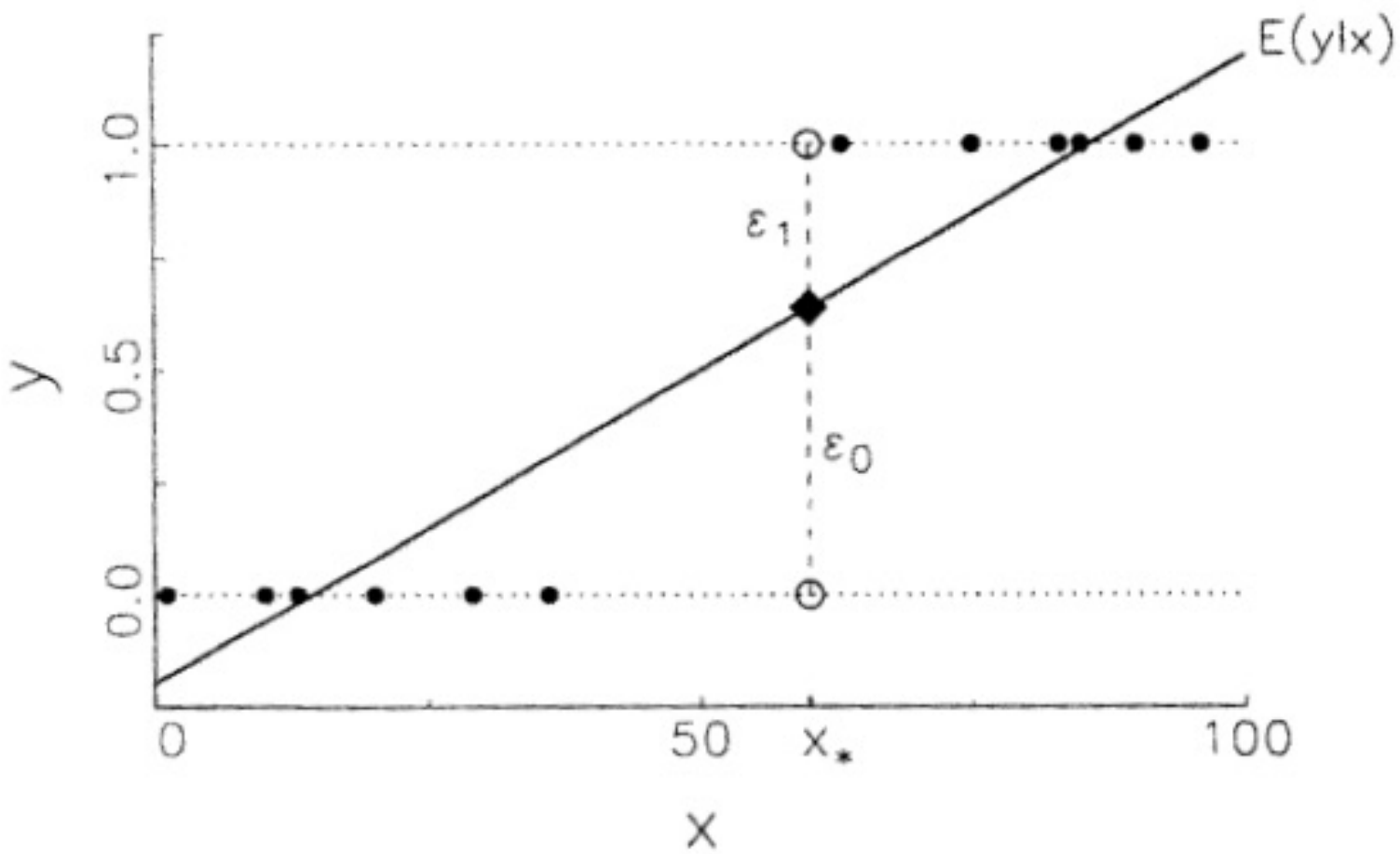

Figura: Modelo de Probabilidad Lineal

Problemas con MPL

- **Heterocedasticidad:** Si una variable aleatoria binaria tiene media μ, entonces su varianza es $\mu(1-\mu)$. Luego:

$$Var(y_i|x_i) = Pr(y_i = 1|x_i)[1 - Pr(y_i = 1|x_i)] = x_i'\beta(1 - x_i'\beta)$$

  lo que implica que la varianza de los errores depende de las x's y no es constante.
- **Predictores sin sentido:** Los predictores del MLP predicen valores de y que son negativos o mayores que 1.
- **Forma Funcional:** Ya que el modelo es lineal, un incremento de una unidad en x_k resulta un cambio de β_k en la probabilidad de un evento. El aumento es el mismo independientemente del valor actual de x.

Problemas con MPL

- Considerar las siguientes dos especificaciones:

$$y_i = \alpha + \beta x_i + \delta d_i + \varepsilon_i$$
$$y_i = F(\alpha + \beta x_i + \delta d_i)$$

donde d_i es una variable dummy.

- El cambio discreto en y cuando d cambia de 0 a 1, manteniendo x constante como:

$$\frac{\Delta y}{\Delta d} = (\alpha + \beta x_i + \delta, 1) - (\alpha + \beta x_i + \delta, 0) = \delta$$

- Para nuestra segunda función el cambio discreto es...?

Problemas con MPL

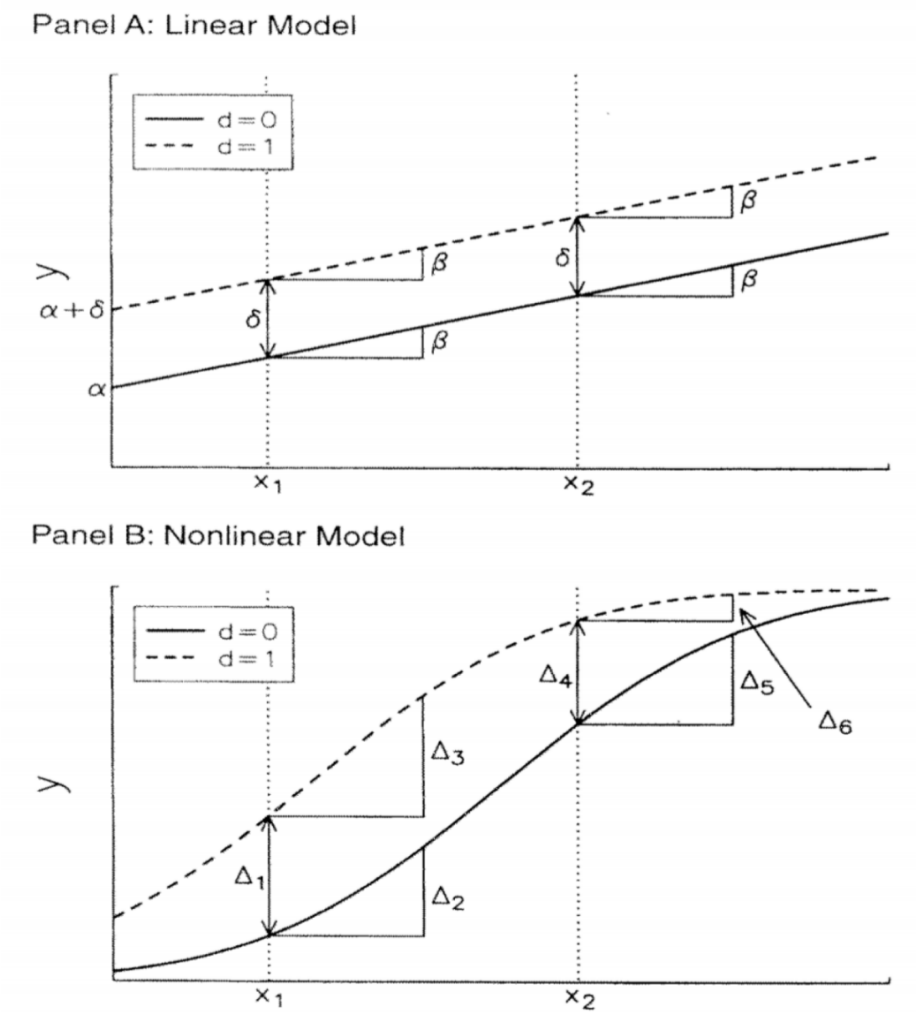

Figura: Efectos en Modelo Lineal y No Lineal

Índice

Enfoque de Variable Latente

- La variable latente y^* es asumida a ser linealmente relacionada a las variables observadas x a través del modelo estructural:

$$y_i^* = x_i^{'}\beta + \varepsilon_i$$

- La variable latente y^* está vinculada a la variable binaria observada y por la ecuación de medida:

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

Enfoque de Variable Latente

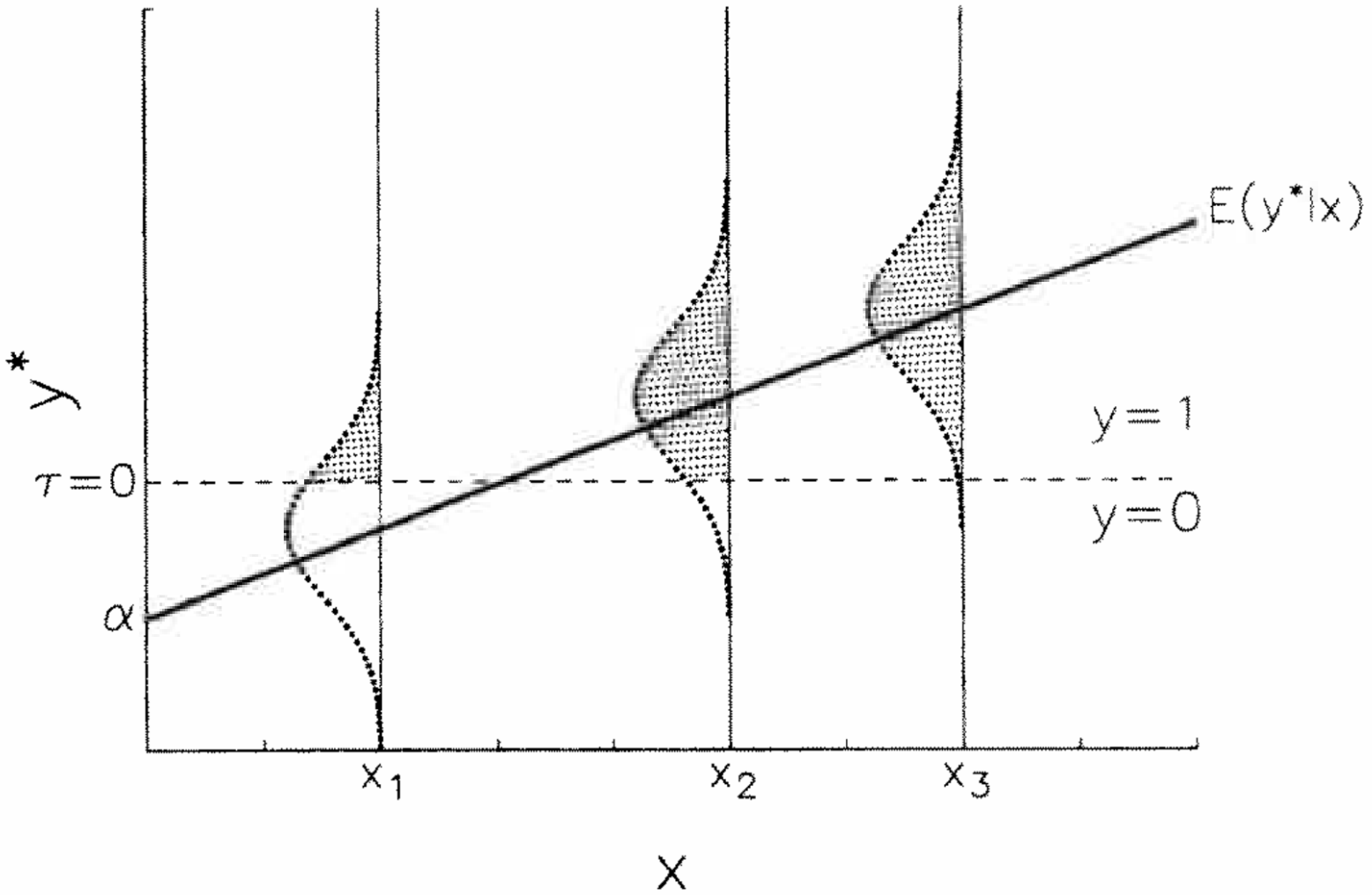

Figura: Distribución de y^* dado x en el modelo binario

Enfoque de Variable Latente

$$
\begin{aligned}
Pr(y_i = 1|x_i) =& Pr(y_i^* > 0|x_i) \\
=& Pr(x_i^{'}\beta + \varepsilon_i > 0|x_i) \\
=& Pr(\varepsilon_i > -x_i^{'}\beta|x_i) \\
=& 1 - Pr(\varepsilon_i \leq -x_i^{'}\beta|x_i) \quad \because Pr(X > x = 1 - Pr(X \leq x) \\
=& Pr(\varepsilon_i \leq x_i^{'}\beta|x_i) \ por\ simetría \\
=& F(x_i^{'}\beta) \\
=& \int_{-\infty}^{x_i^{'}\beta} f(\varepsilon_i) d\varepsilon_i
\end{aligned}
$$

Como siempre asumimos que $E(\varepsilon \mid X) = 0$

Enfoque de Variable Latente

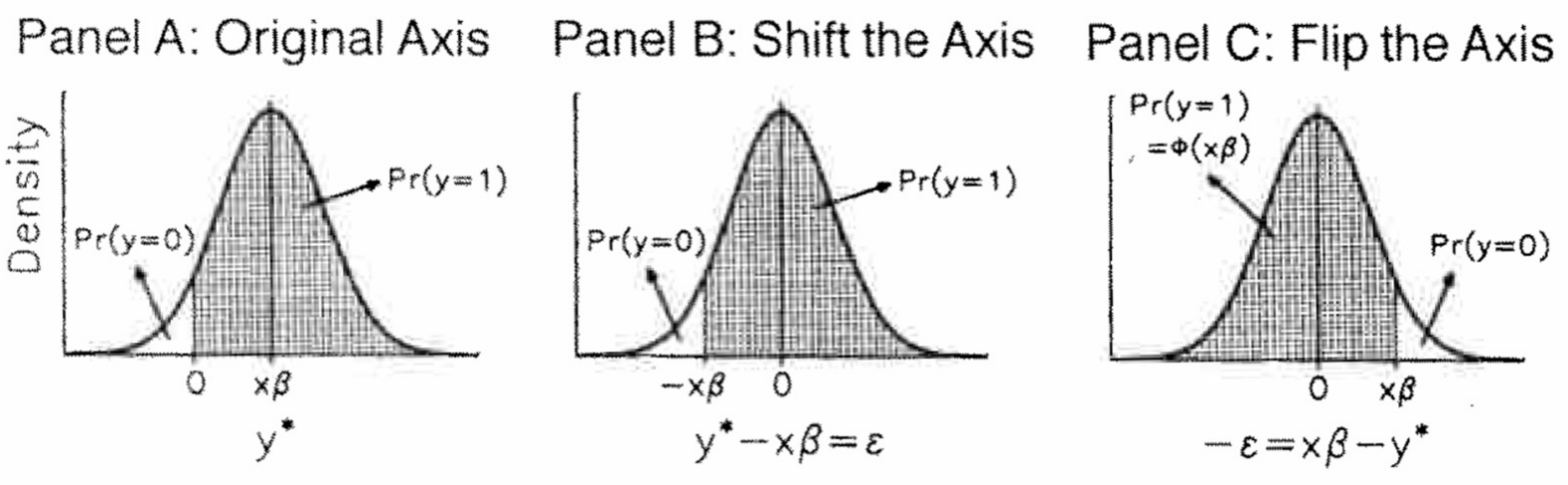

Figura: Computar $Pr(y = 1 \mid x)$ en el modelo binario

Enfoque de Variable Latente

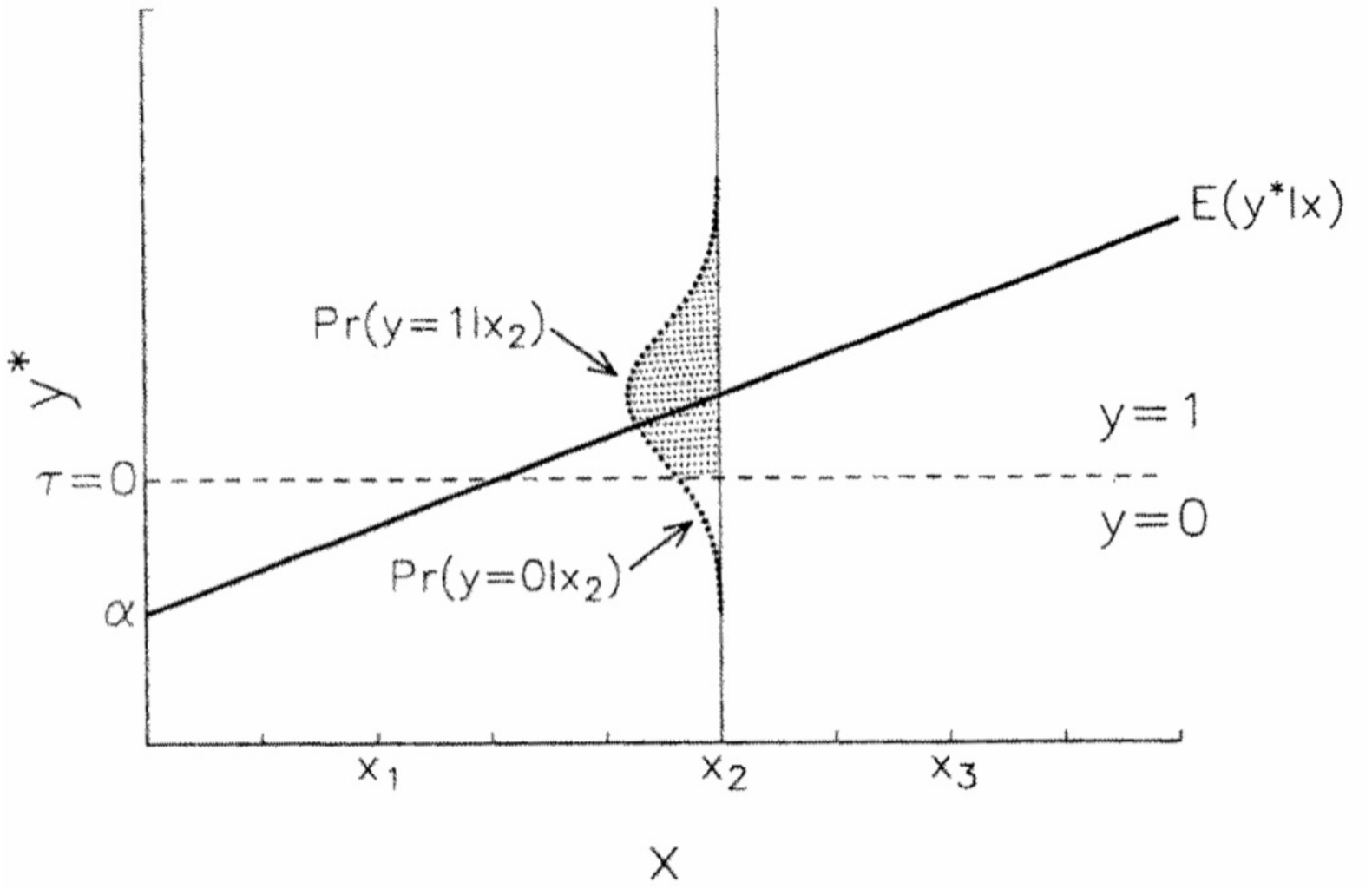

Figura: Probabilidad de valores observados en el modelo binario

Índice

Probit

- En el modelo probit, el término de error ε se distribuye como normal con $E(\varepsilon|X) = 0$ y $Var(\varepsilon|X) = 1$.
- Entonces, la pdf es:

$$\phi(\varepsilon) = \frac{1}{\sqrt{2\pi}} exp(-\frac{\varepsilon^2}{2})$$

- y la función de distribución acumulada (cdf) es:

$$\Phi(\varepsilon) = \int_{-\infty}^{\varepsilon} \frac{1}{\sqrt{2\pi}} exp(-\frac{t^2}{2}) dt$$

Logit

- En el modelo logistico, los errores son asumidos a tener una distribución logistica standard con media 0 y varianza $\pi^2/3$.
- Se elige esta variación inusual porque da como resultado una ecuación particularmente simple para el pdf

$$\lambda(\varepsilon) = \frac{exp(\varepsilon)}{[1+exp(\varepsilon)]^2}$$

- y una ecuación más simple para el cdf:

$$\Lambda(\varepsilon) = \frac{exp(\varepsilon)}{1+exp(\varepsilon)}$$

Distribución Logistica y Normal

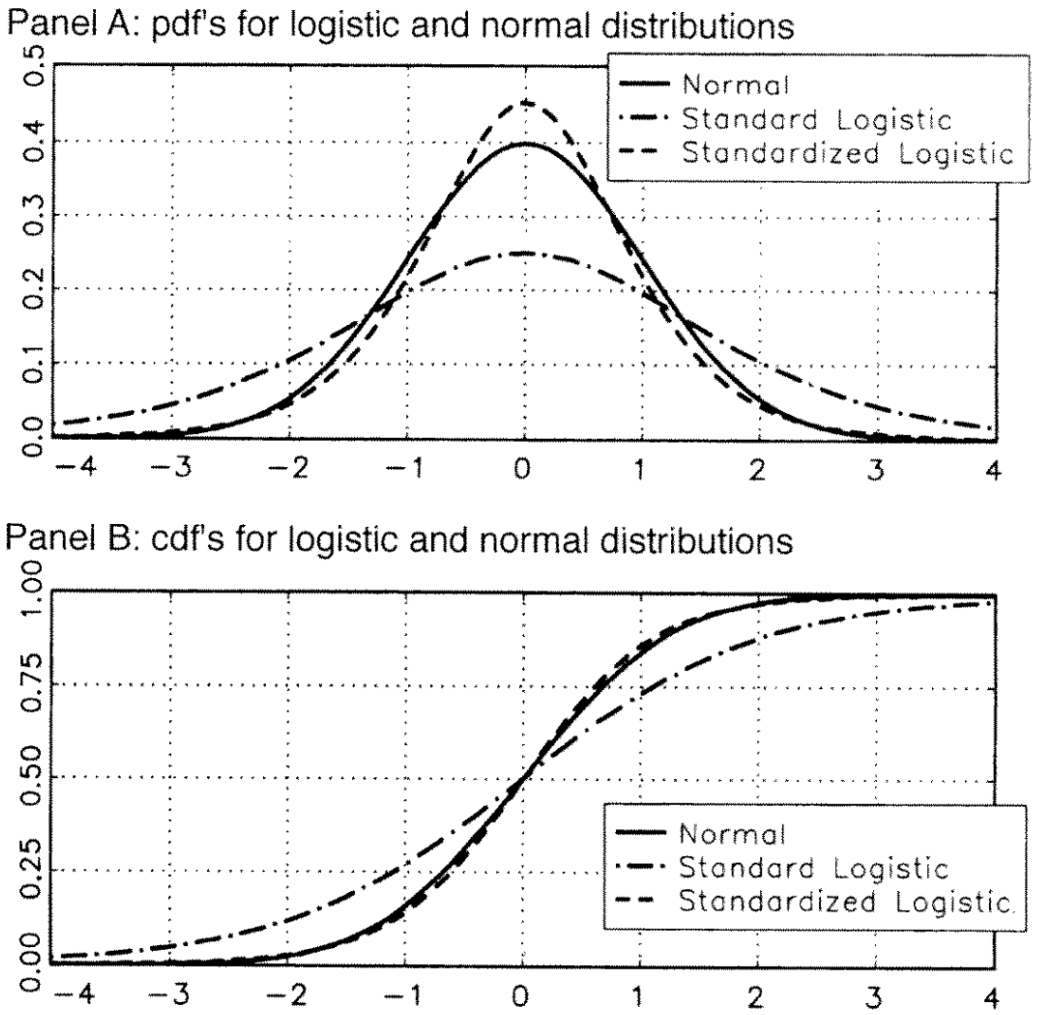

Figura: Distribución Logistica y Normal

Modelos de Variable Dependiente Binaria

- Si los ε_i's son independientemente y normalmente distribuido, $\varepsilon_i \sim N(0, \sigma^2)$. Entonces:

$$\begin{aligned} Pr(y_i = 1|x_i) =& Pr(\frac{\varepsilon_i}{\sigma} > -\frac{x_i'\beta}{\sigma} \mid x_i) \\ =& 1 - \Phi(-\frac{x_i'\beta}{\sigma}) \\ =& \Phi(\frac{x_i'\beta}{\sigma}) \ \textit{por simetría de distribución normal standard} \end{aligned}$$

- La probabilidad depende de β y σ, pero solo la fracción β/σ es identificada, pero no los parametros β y σ
- Por ejemplo, si β y σ son multiplicados por una constance c, entonces la probabilidad permanece sin cambios.
- Tipicamente, sea $\sigma = 1$ como normalización

Índice

Estimación Máxima Verosimilitud

- El resultado es distribuido Bernoulli, la distribución binomial con solo una prueba. Una notación compacta muy conveniente para la densidad de y_i, o más formalmente su función de probabilidad, es:

$$f(y_i|x_i) = P_i^{y_i}(1-P_i)^{1-y_i}, \quad y_i = 1,0$$

donde $P_i = F(x_i'\beta)$. Este produce probabilidades P_i y $(1-P_i)$ ya que $f(1) = P^1(1-P)^0 = P$ y $f(0) = P^0(1-P)^1 = 1-P$. Suponiendo que cada probabilidad es independiente de la otra, la función de probabilidad conjunta (o función de verosimilitud) es:

$$Pr(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n|X) = \prod_{i=1}^{N} f(y_i|x_i)$$

- La función de verosimilitud para una muestra de n observaciones pueden ser escritos como:

$$L(\beta|\underbrace{y, X}_{data}) = \prod_{i=1}^{N}[F(x_i'\beta)]^{y_i}[1-F(x_i'\beta)]^{1-y_i}$$

Función Log-Verosimilitud

- Tomando logs, obtenemos la función de Log-Likelihood, que debe ser maximizado:

$$\begin{aligned} logL(\beta|data) =& log(\prod_{i=1}^{N}[F(x_i'\beta)]^{y_i}[1-F(x_i'\beta)]^{1-y_i}) \\ =& \sum_{i=1}^{n}\{log([F(x_i'\beta)]^{y_i}) + log([1-F(x_i'\beta)]^{1-y_i})\} \\ =& \sum_{i=1}^{n}\{y_i log([F(x_i'\beta)]) + (1-y_i) log([1-F(x_i'\beta)])\} \end{aligned}$$

Truco útil

Si la distribución es simétrica, como la normal y logistica, entonces $1-F(x_i'\beta) = F(-x_i'\beta)$. Sea $q_i = 2y_i - 1$. Entonces:

$$logL(\beta|data) = \sum_{i=1}^{n} logF(q_i x_i'\beta)$$

Gradiente

- Las condiciones de primer orden:

$$\underbrace{\frac{\partial logL(\beta|data)}{\partial \beta}}_{(K\times 1)} = \frac{\partial}{\partial \beta}\sum_{i=1}^{N}\{y_i log[F(x_i'\beta)] + (1-y_i)log[1-F(x_i'\beta)]\}$$

$$= \sum_{i=1}^{n}\{y_i\frac{f(x_i'\beta)x_i}{F(x_i'\beta)} + (1-y_i)\frac{-f(x_i'\beta)x_i}{1-F(x_i'\beta)}\} \qquad \frac{\partial F(x_i'\beta)}{\partial \beta} = f(x_i'\beta)x_i$$

$$= \sum_{i=1}^{n}\{\frac{y_i}{F(x_i'\beta)} - \frac{(1-y_i)}{1-F(x_i'\beta)}\}f(x_i'\beta)x_i$$

$$= \sum_{i=1}^{n}\{\frac{y_i(1-F(x_i'\beta)) - (1-y_i)F(x_i'\beta)}{F(x_i'\beta)[1-F(x_i'\beta)]}\}f(x_i'\beta)x_i$$

$$= \sum_{i=1}^{n}\{\frac{y_i - F(x_i'\beta)}{F(x_i'\beta)[1-F(x_i'\beta)]}\}f(x_i'\beta)x_i$$

Gradiente

- Asimismo, el estimador MV $\hat{\beta}_{MV}$ es dado por la solución de:

$$\underbrace{\sum_{i=1}^{n}\{\frac{y_i - F(x_i'\beta)}{F(x_i'\beta)[1-F(x_i'\beta)]}\}}_{(1\times 1)}\underbrace{f(x_i'\beta)}_{(1\times 1)}\underbrace{x_i}_{(K\times 1)} = \underbrace{0}_{K\times 1}$$

Esta ecuación no tiene una solución analítica de $\hat{\beta}_{MV}$ y tenemos que resolverlo por métodos computacionales.

- Usando nuestro truco previo, tenemos:

$$\frac{\partial logL(\beta)}{\partial \beta} = \sum_{i=1}^{n}[\frac{q_i f(q_i x_i'\beta)}{F(x_i'\beta)}]x_i = \sum_{i=1}^{n}\lambda_i x_i = 0$$

Hesiana

- La Hesiana es:

$$H = \frac{\partial^2 logL}{\partial\beta\partial\beta'} = \sum_{i=1}^{N}[\frac{q_i^2 f'(q_i x_i'\beta)}{F(q_i x_i'\beta)} - \lambda_i^2]x_i x_i'$$

- Nuevamente, tenga en cuenta que $F(\cdot)$ depende de la distribución del término de error. Si asumimos un modelo probit, entonces:

$$F(\varepsilon) = \Phi(\varepsilon) = \int_{-\infty}^{\varepsilon} \frac{1}{\sqrt{2\pi}} exp(-\frac{t^2}{2})dt$$

$$f(\varepsilon) = \phi(\varepsilon) = \frac{1}{\sqrt{2\pi}} exp(-\frac{\varepsilon^2}{2})$$

$$f'(\varepsilon) = -\varepsilon\phi(\varepsilon)$$

- y para el modelo logit, tenemos:

$$F(\varepsilon) = \Lambda(\varepsilon)$$

$$f(\varepsilon) = \Lambda(\varepsilon)[1-\Lambda(\varepsilon)]$$

$$f'(\varepsilon) = \Lambda(\varepsilon)[1-\Lambda(\varepsilon)][1-2\Lambda(\varepsilon)]$$

Comentarios

- El hessiano siempre es definido negativo, por lo que la probabilidad logarítmica es globalmente cóncava.
- El método de Newton generalmente convergerá al máximo de la probabilidad de registro en solo unas pocas iteraciones, a menos que los datos estén especialmente mal condicionados.

Índice

Identificación

- Asumir que el término de error tiene media diferente de cero θ. Entonces:

$$\begin{aligned} Pr(y_i \mid x_i) &= Pr(\varepsilon_i \leq \alpha + x_i'\beta) \\ &= Pr(\varepsilon_i - \theta \leq (\alpha - \theta) + x_i'\beta) \\ &= Pr(\varepsilon_i^* \leq \alpha^* + x_i' + x_i'\beta) \end{aligned}$$

- Mientras el modelo de elección binaria contenga un término constante, no hay pérdida de generalidad, suponiendo que la media del término aleatorio sea cero.
- Una media distinta de cero desaparecería en el término constante de la función de utilidad.

Índice

Efectos Marginales

- Recordar que:

$$E(y_i|x_i) = [1 \times Pr(y_i = 1|x_i)] + [0 \times Pr(y_i = 0|x_i)] = Pr(y_i = 1|x_i) = F(x_i'\beta)$$

El efecto marginal es dado por:

$$\underbrace{\frac{\partial E(y_i|x_i)}{\partial x_i}}_{(K\times 1)} = [\frac{dF(x_i'\beta)}{d(x_i'\beta)}]\beta = \underbrace{f(x_i'\beta)}_{(1\times 1)}\underbrace{\beta}_{(K\times 1)}$$

donde $f(.)$ es la función de densidad de probabilidad.

- Entonces:

$$Probit \Rightarrow \frac{\partial E(y_i|x_i)}{\partial x_i} = \phi(x_i'\beta_i)\beta$$

$$Logit \Rightarrow \frac{\partial E(y_i|x_i)}{\partial x_i} = \Lambda(x_i'\beta_i)[1 - \Lambda(x_i'\beta_i)]\beta$$

Efectos Marginales

- Algunos aspectos son importantes para comentar:
  - Efectos marginales pueden variar con los valores de x
  - Practicas comunes:
    - ★ Calcular efectos marginales en el promedio de las variables
    - ★ Calcular efectos marginales en valores específicos
    - ★ Evaluar efectos marginales en cada observación y usar el promedio muestral de los efectos marginales individuales.

Efectos Marginales: Variable Dummy

- El efecto marginal apropiado para una variable independiente binaria, es decir, d, sería:

$$ME = [Pr(y_i = 1|\bar{x}_{(d)}, d_i = 1)] - [Pr(y_i = 1|\bar{x}_{(d)}, d_i = 0)]$$

donde $\bar{x}_{(d)}$ denota los promedios de todas las otras variables en el modelo.

Efectos Marginales: Elasticidades

- Es común reportar elasticidades de probabilidades, en vez de derivadas. Estos son computados como:

$$\begin{aligned}\varepsilon_{i,k} &= \frac{\partial lnPr(y_i = 1|x)}{\partial lnx_{i,k}} \\ &= \frac{\partial lnPr(y_i = 1|x)}{\partial x_{i,k}} \frac{x_{i,k}}{Pr(y_i = 1|x)}\end{aligned}$$

- Como es una relación de cambios porcentuales, la elasticidad no es útil para las variables dummies.

Bondades de Ajuste

- Se han realizado varias sugerencias sobre cómo evaluar la calidad general de un modelo de respuesta binaria.
  - Primer enfoque: imitar la medida R^2
  - Evaluar el rendimiento predictivo del modelo

- R^2 no son directamente aplicables en modelos no lineales, como los modelos de respuesta binaria, ya que no tenemos un resultado de descomposición de la varianza adecuado
- Se sugiere la medida del pseudo R^2

Índice

Bondades de Ajuste

- Sea
  - $logL(\hat{\beta}_r)$ el valor de la función de log-verosimilitud maximizada en el modelo de solo constante.
  - $logL(\hat{\beta}_u)$: es el valor de log-verosimilitud maximizada en el modelo completo.
- Notar que el valor de las funciones de log-verosimilitud es siempre negativo, entonces:

$$logL(\hat{\beta}_u) \geq logL(\hat{\beta}_r) \Rightarrow \left|logL(\hat{\beta}_u)\right| \leq \left|logL(\hat{\beta}_r)\right|$$

- De tal manera que:

$$0 \leq 1 - \frac{logL(\hat{\beta}_u)}{logL(\hat{\beta}_r)} = R^2_{McFadden} \leq 1$$

- El R^2 McFadden podría ser cero si el modelo completo no tiene poder explicativo.

Bondades de Ajuste

- Otra medida es el McKelvey y Zavoina (1975). Se base en el modelo lineal latente $y_i^* = x_i'\beta + \varepsilon_i$. En particular, si $\hat{y}_i^* = x_i'\hat{\beta}$, entonces escribimos:

$$R^2_{MZ} = \frac{SSE^*}{SSR^* + SSE^*} = \frac{\sum_{i=1}(\hat{y}_i^* - \bar{\hat{y}}^*)^2}{n\sigma^2 + \sum_{i=1}(\hat{y}_i^* - \bar{\hat{y}}^*)^2}$$

- donde SSE^* denota la suma de cuadrados explicada, y SSR^* denota la suma de residuos al cuadrado del modelo latente.
  - Para probit $\sigma^2 = 1$
  - Para logit $\sigma^2 = \pi^2/3$

Índice

Medidas de Información

Criterio de Información de Akaike (AIC)

El criterio de Información de Akaike es definido como:

$$AIC = \frac{-2log\hat{L} + 2K}{n}$$

donde $log\hat{L}$ es la probabilidad del modelo y K es la cantidad de parámetros en el modelo.

- Valores grandes de $log\hat{L}$ indica un mejor ajuste.
- $-2log\hat{L}$ oscila entre 0 y $+\infty$ con valores más pequeños que indican un mejor ajuste.
- A medida que K aumenta, $-2log\hat{L}$ se convierte más pequeño ya que **más parámetros hacen que lo que se observa sea más probable**.
- $2K$ es agregado como un penal.
- Todo lo demás es igual, los valores más pequeños sugieren un modelo de mejor ajuste.
- Usar para comparar modelos en diferentes muestras o para comparar modelos no anidados.

Índice

Medidas de Información

Criterio de Información de Bayes (BIC)

El criterio de Información BIC es definido como:

$$BIC = \frac{-2log\hat{L} + Klog(n)}{n}$$

donde $log\hat{L}$ es la probabilidad del modelo y K es la número de parámetros en el modelo y n es el número de individuos.

- Es posible aumentar la probabilidad agregando parámetros, pero hacerlo puede resultar en un ajuste excesivo. Tanto BIC como AIC resuelven este problema al introducir un término de penalización por la cantidad de parámetros en el modelo; el término de penalización es mayor en el BIC que el AIC.