Probit y logit

Gabriel V. Montes-Rojas

Modelos de respuesta binaria

Supongamos que la variable dependiente es una variable dummy que toma valores 0 y 1. En general, se usa la denominación de la literatura experimental éxito (success) para 1 y fracaso (failure) para 0.
Ejemplos:

- *Participación en la fuerza laboral*: Supongamos que queremos modelar el efecto de ciertas variables en la participación en la fuerza laboral o sobre el desempleo. Supongamos una variable variable `inlf` (in labor force) que toma 1 si el individuo trabaja y 0 si no lo hace.
- *Bancarrota*: Supongamos un modelo que estima el efecto de ciertas variables sobre la probabilidad de que una firma o un individuo se declare en bancarrota.
- *Default de la deuda soberana*: Supongamos un modelo que estima el efecto de las variables macroeconnómicas (déficit, deuda/PBI) sobre la probabilidad que un país se declare en default.

Modelo de probabilidad lineal

Sea $y = 0, 1$ la variable dependiente. Una opción de primera instancia es usar el **modelo de probabilidad lineal** que consiste en correr una regresión MCO:

$$y = \boldsymbol{x}\boldsymbol{\beta} + u.$$

¿Cuál es la interpretación de β?

$$E[y|\boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta} = P[y = 1|\boldsymbol{x}].$$

Entonces $\beta_1 = \frac{\partial P[y=1|\boldsymbol{x}]}{\partial \boldsymbol{x}}$. En otras palabras: $\boldsymbol{\beta}$ nos da *el efecto marginal sobre la probabilidad de tener un éxito* $(y = 1)$.
Nota: Se puede pensar a y cmouna variable Bernoulli que toma valor 1 con probabilidad p y valor 0 con probabiidad $1 - p$. Entonces, $E[y] = p$, $Var[y] = p(1 - p)$.

Modelo de probabilidad lineal

Sin embargo, este modelo tiene ciertos problemas e inconsistencias:

1. **Valor predicho**: El modelo no garantiza que $0 \leq \hat{y} \leq 1$, lo cual es un problema porque estamos hablando de una probabilidad, ya que $\hat{y} = \widehat{P[y = 1|\boldsymbol{x}]}$. Notar que sin embargo $0 \leq \bar{y} = \hat{y}(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}} \leq 1$.
2. **Heteroscedasticidad**: $Var(y|\boldsymbol{x}) = P[y = 1|\boldsymbol{x}] \times (1 - P[y = 1|\boldsymbol{x}])$. [Ejercicio: Proponer un método para obtener errores estándar correctos usando el modelo de probabilidad lineal.]

Modelos logit y probit

- Supongamos el siguiente modelo de variable latente:

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

donde e es una variable aleatoria contínua independiente de $\boldsymbol{x}$ y la distribución de e es simétrica en cero.

- Pero no observamos y^* (por eso es latente), sino

$$y = \mathbf{1}[y^* > 0] = \mathbf{1}[e > -(\boldsymbol{x}\boldsymbol{\beta})],$$

donde $\mathbf{1}[.]$ es la función indicador que toma valor 1 si el argumento en $[.]$ es verdadero, 0 si no lo es.

- Entonces,

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

donde F es la función de distribución acumulada (cdf) de e.

- Diferentes supuestos sobre F determinan distintos modelos:
  - Si F es normal, entonces tenemos el modelo **probit**;
  - Si F es logística, entonces tenemos el modelo **logit**.

Modelo probit

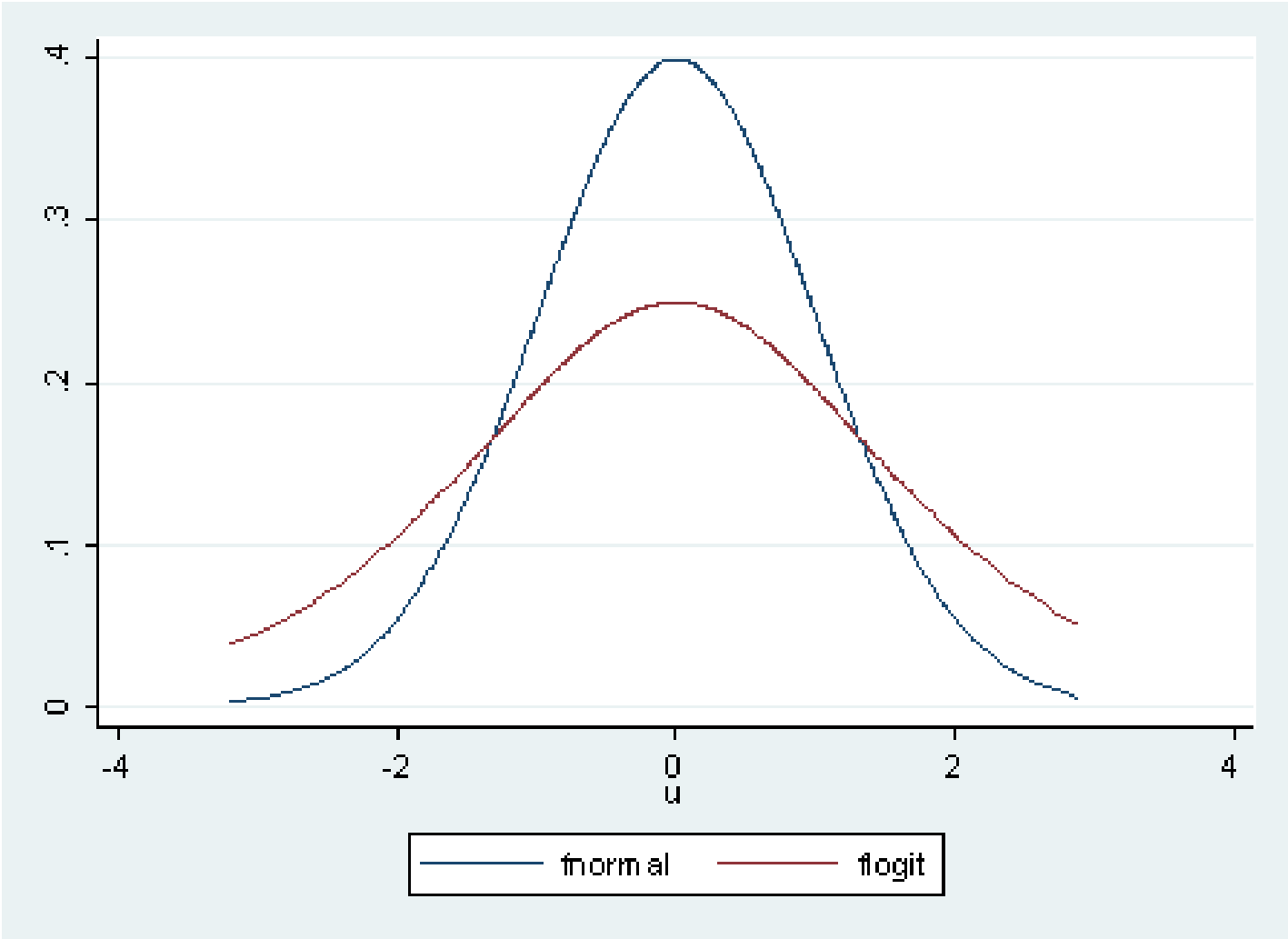
- Supongamos $e \sim N(0, \sigma_e^2)$, modelo **probit**.
- En este caso $F(z) = P[e \leq z] = \int_{-\infty}^{z} \phi(v)dv = \Phi(z)$ donde ϕ es la función de densidad normal, $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, y Φ la *cdf* normal.
- $P[y = 1|\boldsymbol{x}] = P[e > -(\boldsymbol{x}\boldsymbol{\beta})] = 1 - F(-\boldsymbol{x}\boldsymbol{\beta}) = F(\boldsymbol{x}\boldsymbol{\beta}) = \Phi(\boldsymbol{x}\boldsymbol{\beta})$
- $P[y = 0|\boldsymbol{x}] = P[e \leq (-\boldsymbol{x}\boldsymbol{\beta})] = F(-\boldsymbol{x}\boldsymbol{\beta}) = 1 - F(\boldsymbol{x}\boldsymbol{\beta}) = 1 - \Phi(\boldsymbol{x}\boldsymbol{\beta})$

Modelo logit

- Supongamos $e \sim$logística, modelo **logit**.
- En este caso: $F(z) = P[e \leq z] = \frac{\exp(z)}{1+exp(z)} = \Lambda(z)$, donde Λ es la *cdf* de una variable aleatoria logística. Notar que la *pdf* es $\lambda(z) = \frac{\exp(z)}{(1+exp(z))^2}$.
- $P[y = 1|\boldsymbol{x}] = P[e > -(\boldsymbol{x}\boldsymbol{\beta})] = 1 - F(-\boldsymbol{x}\boldsymbol{\beta}) = F(\boldsymbol{x}\boldsymbol{\beta}) = \Lambda(\boldsymbol{x}\boldsymbol{\beta})$
- $P[y = 0|\boldsymbol{x}] = P[e \leq -(\boldsymbol{x}\boldsymbol{\beta})] = F(-\boldsymbol{x}\boldsymbol{\beta}) = 1 - F(\boldsymbol{x}\boldsymbol{\beta}) = 1 - \Lambda(\boldsymbol{x}\boldsymbol{\beta})$

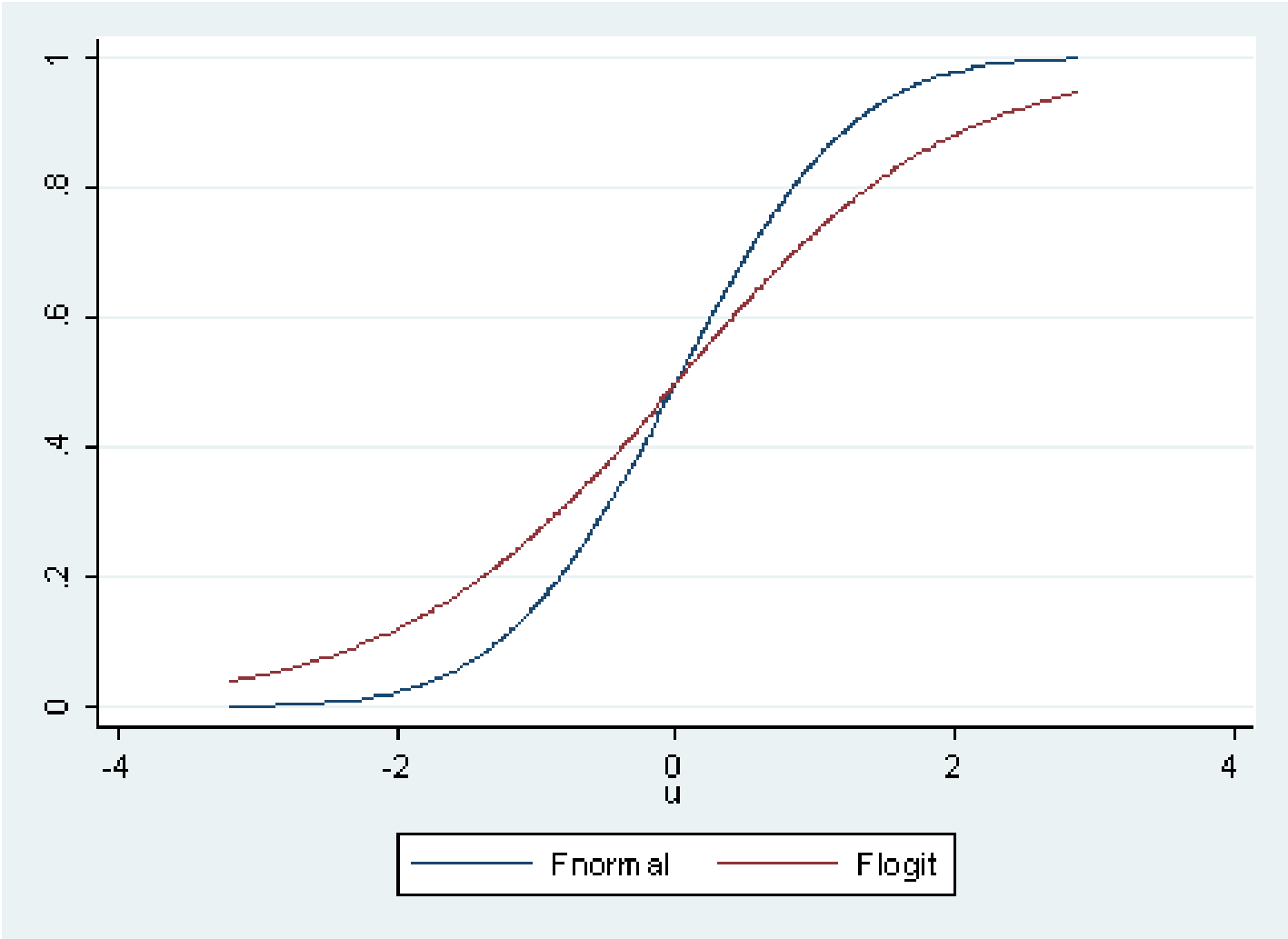
Probit vs. Logit

Funciones de densidad



Probit vs. Logit

Funciones de distribución



Modelos logit y probit

¿Cómo interpretar los coeficientes?

- Si $\boldsymbol{x}$ es una variable continua,

$$\frac{\partial P[y=1|\boldsymbol{x}]}{\partial \boldsymbol{x}} = \frac{\partial F(\boldsymbol{x}\boldsymbol{\beta})}{\partial \boldsymbol{x}} = f(\boldsymbol{x}\boldsymbol{\beta})\boldsymbol{\beta}.$$

  Entonces $\boldsymbol{\beta} \neq \frac{\partial P[y=1|\boldsymbol{x}]}{\partial \boldsymbol{x}}$ y no se puenden interpretar los coeficientes de los modelos probit y logit directamente.

- Necesitamos $f(z) = \frac{\mathrm{d}F}{\mathrm{d}z}(z)$, la densidad asumida de e en el modelo de variable latente. Notar que cada indiviuo i tiene un valor marginal potencialmente diferente: $\frac{\partial P[y=1|\boldsymbol{x}_i]}{\partial \boldsymbol{x}} = f(\boldsymbol{x}_i\boldsymbol{\beta})\boldsymbol{\beta}$.
- Sí se puede interpretar el signo: $sign(\boldsymbol{\beta}) = sign\left(\frac{\partial P[y=1|\boldsymbol{x}]}{\partial \boldsymbol{x}}\right)$.
- Pero, ¿qué valor de $\boldsymbol{x}$ usar para computar $f(\boldsymbol{x}\boldsymbol{\beta})$? En general se usa $\boldsymbol{x} = \bar{\boldsymbol{x}}$.
- También podemos computar el efecto marginal promedio como $E_{\boldsymbol{x}}[f(\boldsymbol{x}\boldsymbol{\beta})\boldsymbol{\beta}]$.
- Si x_K es una variable dummy, entonces se mide el efecto de un cambio de 0 a 1:

$$F(\beta_1 + \beta_2\bar{x}_2 + ... + \beta_{K-1}\bar{x}_{K-1} + \beta_K) - F(\beta_1 + \beta_2\bar{x}_2 + ... + \beta_{K-1}\bar{x}_{K-1}).$$

Comparación de modelos

- Modelo de probabilidad lineal: $F(x) = x$
- Modelo probit: $F(x) = \int_{-\infty}^{x} \frac{1}{2\pi} e^{-\frac{u^2}{2}} du = \Phi(x)$
- Modelo logit: $F(x) = \frac{u^x}{1+u^x} = \Lambda(x)$
- $\hat{\beta}_\Lambda \simeq 1.6\hat{\beta}_\Phi$.
- $\hat{\beta}_{pl} \simeq 0.4\hat{\beta}_\Phi$ excepto para la constante.
- $\hat{\beta}_{pl} \simeq 0.4\hat{\beta}_\Phi + 0.5$ para la constante.

Modelos logit

- Para los modelos logit tenemos una interpretación de los coeficientes.
- Notar que $P[y = 1|\boldsymbol{x}]/P[y = 0|\boldsymbol{x}] = e^{\boldsymbol{x}\boldsymbol{\beta}}$. Esto se conoce como el ratio de odds (odds ratio) y mide cuan probable es un evento en proporción a cuan probable es otro.
- Aplicando logs, $log\left(P[y = 1|\boldsymbol{x}]/P[y = 0|\boldsymbol{x}]\right) = \boldsymbol{x}\boldsymbol{\beta}$, entonces

$$\frac{\partial log\left(P[y = 1|\boldsymbol{x}]/P[y = 0|\boldsymbol{x}]\right)}{\partial x_j} = \beta_j.$$

Los coeficientes se interpretan como el efecto porcentual sobre el ratio de odds (semi-elasticidad).

Una introducción a máxima verosimilitud

- La variable dependiente muestral es $\{y_i\}_{i=1}^N$, que son 0 o 1 para cada observación.
- Si observamos un 1, $y_i = 1$, ¿cuál es la probabilidad asociada a tener este valor en particular en dicha observación i? Usando el modelo de variable latente, $y_i^* = \boldsymbol{x}\boldsymbol{\beta}_i + e_i > 0$, y como e se asumió logit/probit $P[y_i = 1|\boldsymbol{x}_i] = F(\boldsymbol{x}_i\boldsymbol{\beta})$.
- Si observamos un 1, $y_i = 0$, ¿cuál es la probabilidad asociada a tener este valor en particular en dicha observación i? Usando el modelo de variable latente, $y_i^* = \boldsymbol{x}\boldsymbol{\beta}_i + e_i > 0$, y como e se asumió logit/probit $P[y_i = 0|\boldsymbol{x}_i] = 1 - F(\boldsymbol{x}_i\boldsymbol{\beta})$.
- En general tenemos (en forma condensada),

$$P[y|\boldsymbol{x}] = [F(\boldsymbol{x}\boldsymbol{\beta})]^y [1 - F(\boldsymbol{x}\boldsymbol{\beta})]^{1-y}.$$

Una introducción a máxima verosimilitud

- Si tomamos TODAS las observaciones juntas, $\{y_i\}_{i=1}^N$, asumiendo independencia entre las observaciones,

$$P[y_1, y_2, ..., y_N|\boldsymbol{x}] = \prod_{i=1}^{N} P[y_i|\boldsymbol{x}_i] = \prod_{i=1}^{N} [F(\boldsymbol{x}_i\boldsymbol{\beta})]^{y_i} [1 - F(\boldsymbol{x}_i\boldsymbol{\beta})]^{1-y_i}$$

Esta es la función de verosimilitud (likelihood function).

Dos eventos A y B son independientes si $P[A\&B] = P[A] \times P[B]$.

- En general es más fácil trabajar con el log de la función de verosimilitud (log-likelihood function).

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\beta})$$

$$\ell_i(\boldsymbol{\beta}) = \log\ P[y_i|\boldsymbol{x}_i] = y_i \log F(\boldsymbol{x}_i\boldsymbol{\beta}) + (1 - y_i) \log[1 - F(\boldsymbol{x}_i\boldsymbol{\beta})]$$

- Entonces, el estimador de máxima verosimilitud (maximum likelihood estimator, MLE) es $\hat{\boldsymbol{\beta}}$ que maximiza $\mathcal{L}(\boldsymbol{\beta})$. En otras palabras, para cada valor posible de $\boldsymbol{\beta}$, $\mathcal{L}(\hat{\boldsymbol{\beta}}) \geq \mathcal{L}(\boldsymbol{\beta})$.

Una introducción a máxima verosimilitud

- La función score de este modelo es

$$\boldsymbol{s}_i(\beta) = \frac{f(\boldsymbol{x}_i\beta)\boldsymbol{x}_i'[y_i - F(\boldsymbol{x}_i\beta)]}{F(\boldsymbol{x}_i\beta)[1 - F(\boldsymbol{x}_i\beta)]}$$

- Ejercicio: Probar que $E[\boldsymbol{s}_i(\beta_0)|\boldsymbol{x}_i] = 0$, donde β_0 es el valor verdadero de β.
- El hessiano es

$$\begin{aligned}\boldsymbol{H}_i(\beta) = & \left\{ \frac{f'(\boldsymbol{x}_i\beta)[y_i - F(\boldsymbol{x}_i\beta)]}{F(\boldsymbol{x}_i\beta)[1 - F(\boldsymbol{x}_i\beta)]} - \frac{f^2(\boldsymbol{x}_i\beta)}{F(\boldsymbol{x}_i\beta)[1 - F(\boldsymbol{x}_i\beta)]} \right. \\ & \left. - \frac{f(\boldsymbol{x}_i\beta)[y_i - F(\boldsymbol{x}_i\beta)]}{(F(\boldsymbol{x}_i\beta)[1 - F(\boldsymbol{x}_i\beta)])^2}(f(\boldsymbol{x}_i\beta) - 2F(\boldsymbol{x}_i\beta)f(\boldsymbol{x}_i\beta)) \right\}(\boldsymbol{x}_i'\boldsymbol{x}_i).\end{aligned}$$

$$-E[\boldsymbol{H}_i(\beta)|\boldsymbol{x}_i] = \boldsymbol{A}_i(\beta) = \frac{f^2(\boldsymbol{x}_i\beta)(\boldsymbol{x}_i'\boldsymbol{x}_i)}{F(\boldsymbol{x}_i\beta)[1 - F(\boldsymbol{x}_i\beta)]}$$

Una introducción a máxima verosimilitud

- Por normalidad asintótica de máxima verosimilitud

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{A}_0^{-1}),$$

  donde $\boldsymbol{A}_0 = -E[\boldsymbol{H}(\beta_0)]$.
- Con los elementos de la diapositiva anterior tenemos todo lo que necesitamos para armar la matriz de varianzas y covarianzas.
- Probar que $-\boldsymbol{A}_0 = -E[\boldsymbol{H}(\beta_0)] = \boldsymbol{B}_0 = E[\boldsymbol{s}(\beta_0)\boldsymbol{s}(\beta_0)']$. ¿Por qué?
- ¿Cómo estimar la varianza en forma consistente?

$$AVar(\hat{\beta}) = \left(\sum_{i=1}^{N} \frac{f^2(\boldsymbol{x}_i\hat{\beta})(\boldsymbol{x}_i'\boldsymbol{x}_i)}{F(\boldsymbol{x}_i\hat{\beta})[1 - F(\boldsymbol{x}_i\hat{\beta})]} \right)^{-1}$$

Pseudo R cuadrado

- En los modelos de regresión, el R-cuadrado tiene una interpretación intuitiva y simple.
- Eso no ocurre por fuera de los modelos OLS...
- Una primera aproximación puede ser obtener el resultado predecido como $\hat{y}_i = max\{j = 0, 1 : p_j(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})\}$, es decir la probabilidad que sea mayor a 0.5.

  Entonces, $peudo - R^2 = \frac{\sum_{i=1}^{N} 1[\hat{y}_i = y_i]}{N}$.
- McFadden (1974) propone usar el log likelihood ratio index:

$$LRI = 1 - \frac{\mathcal{L}(\hat{\boldsymbol{\beta}})}{\mathcal{L}(\boldsymbol{\beta} = 0)}$$

  En general LRI no se acerca a 1 cuando se incrementan la cantiad de variables. Y si es 1, en realidad es un problema ya que dice que hay una variable que predice perfectamente el resultado, ej. $x > 0$ entonces $y = 1$.

STATA

- Es útil seguir el manual de STATA (v.13): `http://www.stata.com/manuals13/rprobit.pdf` para probit, `http://www.stata.com/manuals13/rlogit.pdf` para logit.
- `probit y x1 x2`
- `logit y x1 x2`
- Dado que los coeficientes no se pueden interpretar directamente (sólo el signo), necesitamos computar los efectos marginales:
  - `dprobit y x1 x2`
  - `logit y x1 x2`
  - `margins, dydx(*)` (calculado como el promedio de las densidades)
  - `margins, dydx(*) atmeans` (calculado con la densidad en el promedio de las X)

Ejemplos

- Considere los ejemplos de `http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge17.html`
- Descripción de la base de datos de Mroz: `http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.des`
- Ejemplo de programación en STATA: `https://stats.idre.ucla.edu/stata/code/imple-linear-and-nonlinear-models-using-statas-ml-command/`