LISBOA
SCHOOL OF
ECONOMICS &
MANAGEMENT

Capítulo 5 – Modelos de escolha Binária

Introdução; modelos lineares e não lineares; Estimação

Luís Silveira Santos

lsantos@iseg.ulisboa.pt

Mestrado em Econometria Aplicada e Previsão,
Instituto Superior de Economia e Gestão – Universidade de Lisboa

Abril de 2017

Programa desta aula

1. Introdução
2. Modelo Probabilístico Linear
   - Especificação
   - Problemas
   - Observações
3. Modelos não lineares – Probit e Logit
   - Especificação
   - Motivação económica
   - Especificação da função $G(\cdot)$
4. Estimação
5. Observações

Variáveis dependentes binárias

- Pretendemos estudar variáveis que assumem apenas dois conjuntos de resultados possíveis (tem ou não um dado atributo).
- Exemplos:
  1. $y_i = \begin{cases} 1 & \text{se o indivíduo } i \text{ está empregado} \\ 0 & \text{o.c.} \end{cases}$
  2. $y_i = \begin{cases} 1 & \text{se a empresa } i \text{ tem fundo de pensões próprio} \\ 0 & \text{o.c.} \end{cases}$
  3. $y_i = \begin{cases} 1 & \text{se a Família } i \text{ tem seguro de saúde} \\ 0 & \text{o.c.} \end{cases}$
  4. ...

Quantidades de interesse

- Neste contexto, pretendemos explicar a probabilidade de y exibir o atributo, em função de um conjunto de regressores $\mathbf{X}$:
$$E(Y \mid \mathbf{X}) = 1 \times P(Y = 1 \mid \mathbf{X}) + 0 \times P(Y = 0 \mid \mathbf{X}) = P(Y = 1 \mid \mathbf{X})$$
- Por sua vez, estamos também interessados em obter os efeitos sobre a probabilidade de y exibir o atributo, quando um determinado regressor x_j ($j = 1, \ldots, K$) varia uma unidade:
  - Se x_j é variável contínua:
$$\frac{\partial P(Y = 1 \mid \mathbf{X})}{\partial x_j} = \frac{\partial p(\mathbf{x})}{\partial x_j}, \quad j = 1, \ldots, K$$
  - Se x_j é variável discreta:
$$p(x_1, x_2, \ldots, x_{j-1}, c_j + 1) - p(x_1, x_2, \ldots, x_{j-1}, c_j)$$
onde c_j é um atributo da variável x_j (ex.: determinar o efeito na probabilidade de "sucesso" quando se tem mais um filho)

Características das V.A.s com distribuição de Bernoulli

- Uma vez que $y \in \{0, 1\}$, sabemos que $Y \mid \mathbf{X} \sim Ber(\theta)$
- Neste caso em concreto teremos $\theta = p(\mathbf{x})$
- Os seus momentos são:
  - $E(Y \mid \mathbf{X}) = p(\mathbf{x})$
  - $Var(Y \mid \mathbf{X}) = p(\mathbf{x})[1 - p(\mathbf{x})]$
- Pela natureza de y, sabemos ainda que:
  - $P(Y = 0 \mid \mathbf{X}) = 1 - P(Y = 1 \mid \mathbf{X})$
  - Exibe heterocedasticidade natural, excepto no caso em que $p(\mathbf{x})$ não dependa de $\mathbf{X}$, i.e., quando $P(Y = 1 \mid \mathbf{X}) = P(Y = 1)$

Que modelo ajustar?

- A principal questão relativamente ao estudo das variáveis binárias diz respeito à determinação do tipo de modelo a ajustar aos dados.

$$\Downarrow$$

- Modelos lineares vs. modelos não lineares

MPL – Especificação

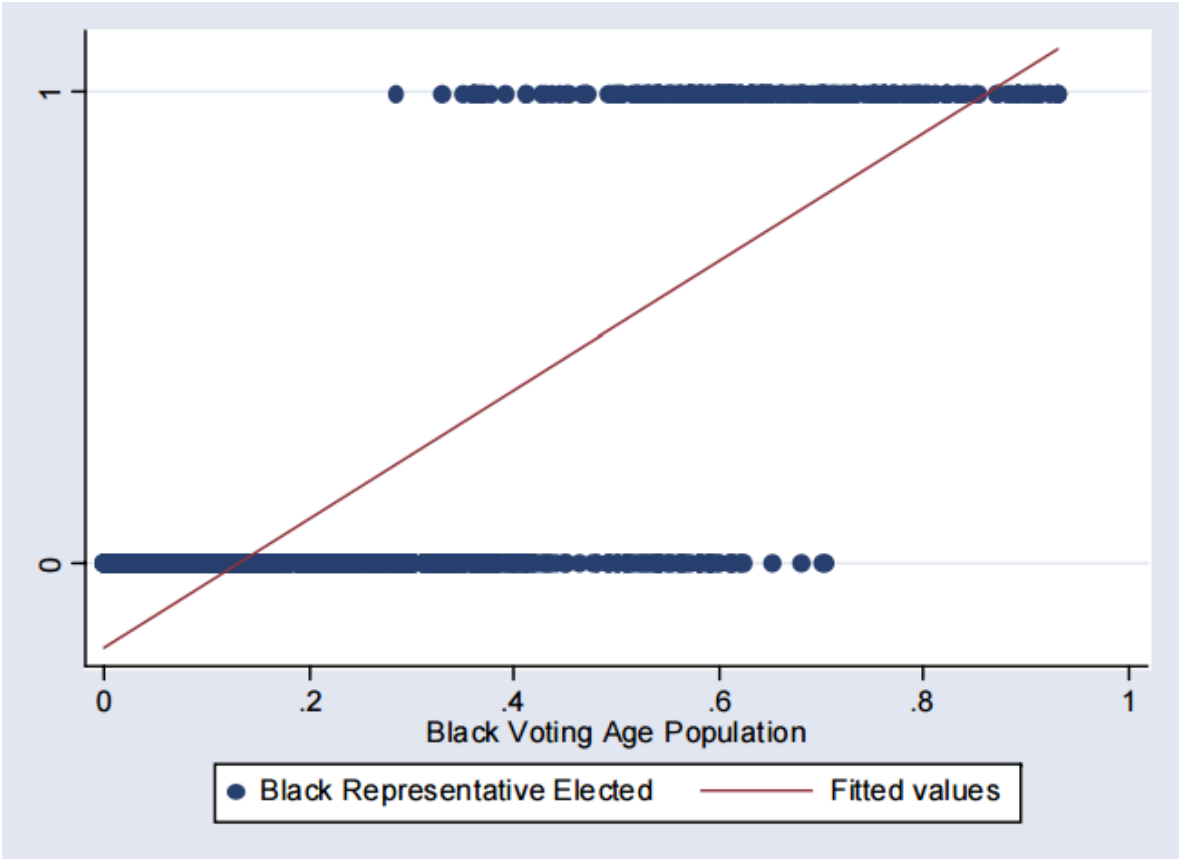
- O MPL estabelece que a probabilidade de "sucesso" é função linear de um conjunto de K regressores:
$$P(Y = 1 \mid \mathbf{X}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K = \mathbf{X}\boldsymbol{\beta}$$
em que $\mathbf{X}$ é matriz $N \times K$ e $\boldsymbol{\beta}$ é vector $K \times 1$.

- Sabemos que $P(Y = 1 \mid \mathbf{X}) = E(Y \mid \mathbf{X})$. Assim sendo, dentro do contexto das hipóteses clássicas (Wooldridge, 2013), as estimativas OLS do vector de parâmetros $\boldsymbol{\beta}$ são centradas e consistentes

- No contexto linear, os efeitos parciais são dados pela fórmula habitual:
$$\frac{\partial P(Y = 1 \mid \mathbf{X})}{\partial x_j} = \beta_j, \quad j = 1, \ldots, K$$

MPL – Problemas

- Observemos graficamente a seguinte relação entre uma variável binária e uma variável contínua:



MPL – Problemas (cont.)

1. Os valores ajustados, $\widehat{P(Y=1 \mid \mathbf{X})} = \mathbf{X}\hat{\boldsymbol{\beta}} \notin [0,1]$
2. Variações sucessivas em x_j, *ceteris paribus*, implicam variações sucessivas de $P(Y=1 \mid \mathbf{X})$, nos valores de β_j, conduzindo a situações em que $P(Y=1 \mid \mathbf{X}) \notin [0,1]$
3. Heterocedasticidade natural,
$$Var(Y \mid \mathbf{X}) = P(Y=1 \mid \mathbf{X})[1 - P(Y=1 \mid \mathbf{X})] = \\ = \mathbf{X}\boldsymbol{\beta}(1 - \mathbf{X}\boldsymbol{\beta})$$
4. Uma vez que hipótese de normalidade não se verifica, teremos de nos basear na teoria assimptótica para dedução das distribuições assimptóticas dos estimadores e das estatísticas de teste

MPL – Problemas (cont.)

- Na prática, apenas o problema 2 é verdadeiramente importante, no entanto este é consequência de se assumir um modelo linear para este tipo de dados
- O problema 3 (sobre heterocedasticidade natural) pode ser resolvido de duas formas:
  - Erros-padrão robustos de White
  - Estimação por GLS (uma vez que a forma teórica da matriz de covariância é conhecida)

MPL – Observações

- Se estimarmos o OLS sem corrigir os erros-padrão, existe apenas uma situação em que a estatística F é válida:

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

uma vez que, sob H_0, $Var(Y \mid \mathbf{X}) = \beta_0 (1 - \beta_0) =$ constante

- Podemos estimar $P(Y = 1 \mid \mathbf{X})$ através do WLS, utilizando como ponderador:

$$\left[\mathbf{x}_i \hat{\boldsymbol{\beta}} \left(1 - \mathbf{x}_i \hat{\boldsymbol{\beta}}\right)\right]^{-1/2}$$

Note-se que este ponderador é específico para cada indíviduo i

- O MPL pode ser interpretado como a melhor aproximação linear (em erro quadrático médio), à verdadeira probabilidade $p(\mathbf{x})$

Modelos não lineares – Especificação

- Pretende-se agora estudar a probabilidade de "sucesso" com recurso a modelos não lineares:

  $$P(Y = 1 \mid \mathbf{X}) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K) = G(\mathbf{X}\boldsymbol{\beta})$$

  onde se assume que $G : \mathbb{R} \rightarrow [0, 1]$

- Neste caso, os efeitos parciais já refletem a natureza não linear do modelo:

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

  onde $g(\cdot)$ é a primeira derivada da função $G(\cdot)$. Estas funções dependem de todos os valores da matriz $\mathbf{X}$.

- Na maior parte das aplicações, assume-se que $G(\cdot)$ é uma função de distribuição, cuja expressão depende do problema económico em questão

Modelo de variável latente

- Considere-se o seguinte modelo de variável latente:

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, N$$

onde,

  - y_i^* é variável dependente latente escalar
  - $\mathbf{x}_i$ é vector de regressores, $1 \times K$
  - $\boldsymbol{\beta}$ é vector de parâmetros, $K \times 1$
  - $\varepsilon_i \perp \mathbf{x}_i, \forall i$

Exemplo: compra de carro

- Vamos supor que pretendemos explicar a utilidade de um determinado indivíduo adquirir um veículo automóvel:

$$U_{\text{Comprar},i} = \mathbf{x}_i\boldsymbol{\gamma} + \varepsilon_{1i}$$

e

$$U_{\text{Não comprar},i} = \mathbf{x}_i\boldsymbol{\delta} + \varepsilon_{2i}$$

Exemplo: compra de carro (cont.)

- Na prática não observamos **quantitativamente** a utilidade que um dado indivíduo i retira das suas decisões, mas sim o resultado da sua decisão:

$$\begin{aligned} y_i^* &= U_{\text{Comprar}} - U_{\text{Não comprar}} = \\ &= \mathbf{x}_i \underbrace{(\gamma - \delta)}_{\beta} + \underbrace{(\varepsilon_{1i} - \varepsilon_{2i})}_{\varepsilon_i} \end{aligned}$$

- Ou seja:

$$y_i = \begin{cases} 1 & \text{se } y_i^* > 0 \\ 0 & \text{o.c.} \end{cases}$$

Exemplo: compra de carro (cont.)

- Neste contexto, estamos interessados em calcular a probabilidade de um dado indivíduo adquirir o veículo automóvel, i.e.:

$$\begin{aligned} P(Y = 1 \mid \mathbf{X}) &= P(Y^* > 0 \mid \mathbf{X}) \\ &= P(\mathbf{X}\boldsymbol{\beta} + \varepsilon > 0 \mid \mathbf{X}) \\ &= P(\varepsilon > -\mathbf{X}\boldsymbol{\beta} \mid \mathbf{X}) \quad \text{(HIP.: distrib. de } \varepsilon \text{ é simétrica)} \\ &= P(\varepsilon < \mathbf{X}\boldsymbol{\beta} \mid \mathbf{X}) \\ &= G(\mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

- A estimação desta probabilidade depende da distribuição assumida para o termo de erro (condicionado pelos regressores), $\varepsilon \mid \mathbf{X}$.

Exemplo: compra de carro (cont.)

- Habitualmente consideramos $\varepsilon \mid \mathbf{X} \sim N\left(0, \sigma_\varepsilon^2\right)$, no entanto esta decisão pode conduzir a problemas adicionais:

$$
\begin{aligned}
P\left(\varepsilon < \mathbf{X}\boldsymbol{\beta} \mid \mathbf{X}\right) &= P\left(\frac{\varepsilon}{\sigma_\varepsilon} < \frac{\mathbf{X}\boldsymbol{\beta}}{\sigma_\varepsilon} \,\middle|\, \mathbf{X}\right) \\
&= P\left(\frac{\varepsilon}{\sigma_\varepsilon} < \frac{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K}{\sigma_\varepsilon} \,\middle|\, \mathbf{X}\right) \\
&= P\left(\frac{\varepsilon}{\sigma_\varepsilon} < \frac{\beta_0}{\sigma_\varepsilon} + \frac{\beta_1}{\sigma_\varepsilon} x_1 + \frac{\beta_2}{\sigma_\varepsilon} x_2 + \ldots + \frac{\beta_K}{\sigma_\varepsilon} x_K \,\middle|\, \mathbf{X}\right)
\end{aligned}
$$

- O vector de parâmetros a ser estimado é:

$$\boldsymbol{\theta} = \left(\sigma_\varepsilon^2, \beta_0, \beta_1, \beta_2, \ldots, \beta_K\right)$$

Exemplo: compra de carro (cont.)

- No entanto, pela especificação do modelo e da distribuição condicionada $\varepsilon \mid \mathbf{X}$, não é possível estimar separadamente σ_ε^2 dos restantes βs

- Como em geral se desconhece σ_ε^2, porque não observamos ε, o que estamos a estimar?

$$(\beta_0, \beta_1, \beta_2, \ldots, \beta_K) \text{ OU } \left(\frac{\beta_0}{\sigma_\varepsilon}, \frac{\beta_1}{\sigma_\varepsilon}, \frac{\beta_2}{\sigma_\varepsilon}, \ldots, \frac{\beta_K}{\sigma_\varepsilon}\right)$$

$$\Downarrow$$

PROBLEMA DE IDENTIFICAÇÃO

- A única quantidade que sabemos **de certeza** são as *odds*:

$$\frac{\partial p(\mathbf{x})/\partial x_j}{\partial p(\mathbf{x})/\partial x_h} = \frac{\beta_j \phi(\cdot)}{\beta_h \phi(\cdot)} = \frac{\beta_j}{\beta_h}, \; h \neq j; \; h, j = 1, 2, \ldots, K$$

Especificação da função $G(\cdot)$

- A especificação da função G depende da hipótese assumida sobre a distribuição do termo de erro
- No entanto, ao contrário dos modelos lineares, não basta assumir hipóteses genéricas sobre os parâmetros dessa distribuição, uma vez que podem conduzir a problemas de identificação
- Na literatura surgiram duas especificações para a distribuição $\varepsilon \mid \mathbf{X}$, que devido à sua simplicidade e por acomodar a maioria dos problemas económicos são bastante utilizadas

Especificação da função $G(\cdot)$ (cont.)

- $\varepsilon \mid \mathbf{X} \sim N(0,1)$:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(z)\,\mathrm{d}z$$

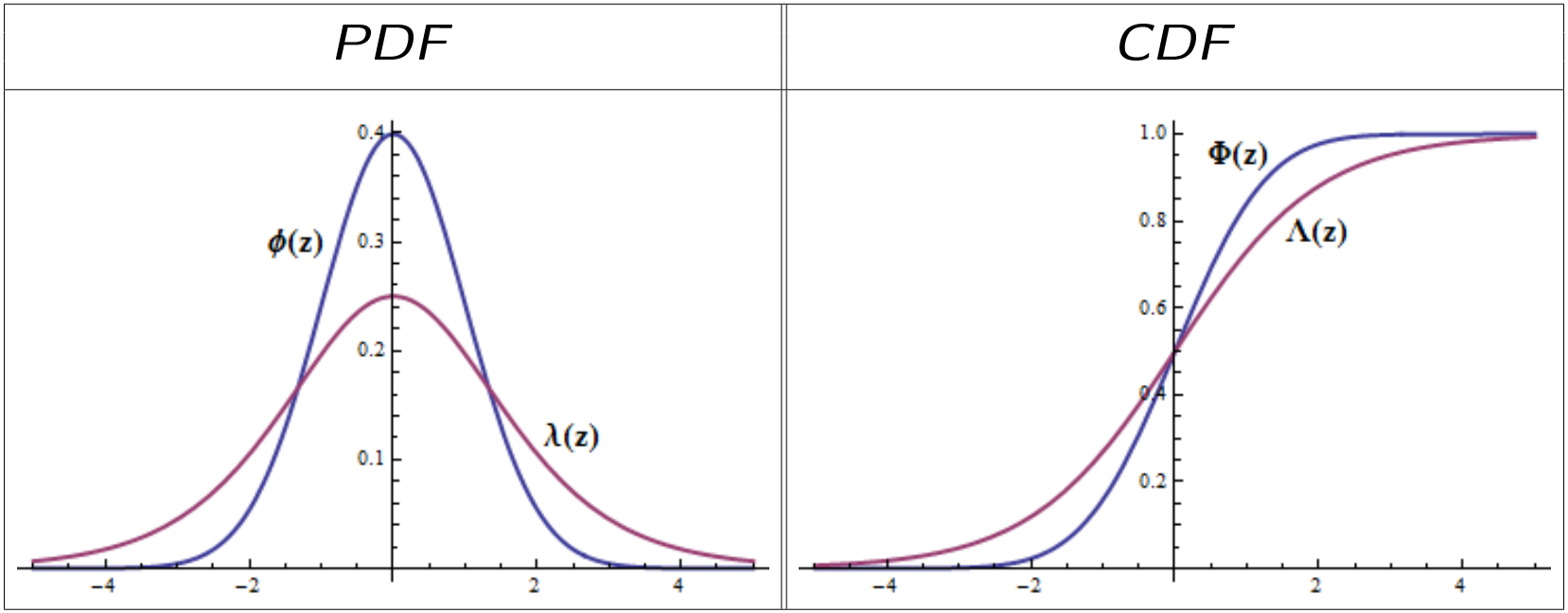
onde $\phi(z) = (2\pi)^{-1/2} \exp\left(-z^2/2\right)$, a função densidade da distribuição Normal estandardizada, $Var(\varepsilon \mid \mathbf{X}) = 1$

- $\varepsilon \mid \mathbf{X} \sim Logistic(0,1)$:

$$G(z) = \Lambda(z) = \frac{\exp(z)}{1 - \exp(z)}$$

onde $\Lambda(z)$ é a função densidade da distribuição Logística estandardizada, $Var(\varepsilon \mid \mathbf{X}) = \pi^2/3$

Especificação da função $G(\cdot)$ (cont.)



- Na prática, a especificação de $G(\cdot)$ depende da densidade associada aos valores extremos da variável de interesse.

Estimação

- A estimação de modelos não lineares pode ser feita com recurso a dois métodos: Mínimos quadrados não lineares (NLS) e Máxima Verosimilhança (ML)
  1. Do método NLS iremos obter:
     - Estimadores consistentes e $\sqrt{N}$ assimptoticamente Normais
     - Inferência robusta a heterocedasticidade com forma funcional genérica
  2. Do método ML iremos obter:
     - Estimadores consistentes e $\sqrt{N}$ assimptoticamente Normais
     - Variância assimptótica que atinge o limite inferior de Fréchet-Cramer-Rao $\Rightarrow$ estimador assimptoticamente mais eficiente
     - **No entanto**, estes resultados são válidos apenas se a função $G(\cdot)$ corresponder à verdadeira densidade de $Y \mid \mathbf{X}$

Estimação (cont.)

- Assumindo que a densidade de $Y \mid \mathbf{X}$ está bem especificada, iremos optar pelo estimador assimptoticamente mais eficiente: o estimador ML
  - Assim sendo, em primeiro lugar, será necessário definir a densidade de $Y \mid \mathbf{X}$ (em termos genéricos):

$$f\left(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}\right) = \left[G\left(\mathbf{x}_i\boldsymbol{\beta}\right)\right]^{y} \left[1 - G\left(\mathbf{x}_i\boldsymbol{\beta}\right)\right]^{1-y}$$

  - Em seguida, iremos obter a função log-verosimilhança para o indivíduo i:

$$\ell_i\left(\boldsymbol{\beta}\right) = y_i \log\left[G\left(\mathbf{x}_i\boldsymbol{\beta}\right)\right] + \left(1 - y_i\right) \log\left[1 - G\left(\mathbf{x}_i\boldsymbol{\beta}\right)\right]$$

  - Note-se que a função log-verosimilhança para uma amostra i.i.d. de dimensão N é imediatamente obtida por via do somatório da função log-verosimilhança individual

Funções Score e Hessiana

- Se derivarmos uma vez a função log-verosimilhança individual iremos obter a função score para o indivíduo i:

$$\mathbf{s}_i(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta})' =$$

$$= y_i \left[\frac{g(\mathbf{x}_i\boldsymbol{\beta})\mathbf{x}_i'}{G(\mathbf{x}_i\boldsymbol{\beta})} \right] - (1 - y_i) \left[\frac{g(\mathbf{x}_i\boldsymbol{\beta})\mathbf{x}_i'}{1 - G(\mathbf{x}_i\boldsymbol{\beta})} \right] =$$

$$= \frac{g(\mathbf{x}_i\boldsymbol{\beta})\mathbf{x}_i' \{y_i [1 - G(\mathbf{x}_i\boldsymbol{\beta})] - (1 - y_i) G(\mathbf{x}_i\boldsymbol{\beta})\}}{G(\mathbf{x}_i\boldsymbol{\beta}) [1 - G(\mathbf{x}_i\boldsymbol{\beta})]} =$$

$$= \frac{g(\mathbf{x}_i\boldsymbol{\beta})\mathbf{x}_i'}{G(\mathbf{x}_i\boldsymbol{\beta}) [1 - G(\mathbf{x}_i\boldsymbol{\beta})]} [y_i - G(\mathbf{x}_i\boldsymbol{\beta})]$$

onde $g(\cdot)$ é a primeira derivada da função $G(\cdot)$

Funções Score e Hessiana (cont.)

- Derivando 2ª vez, iremos obter a hessiana para o indivíduo i:

$$\mathbf{H}_i(\beta) = \nabla^2_\beta \ell_i(\beta) =$$

$$= \frac{g(\mathbf{x}_i\beta)\mathbf{x}_i' \times [-g(\mathbf{x}_i\beta)\mathbf{x}_i]}{G(\mathbf{x}_i\beta)[1-G(\mathbf{x}_i\beta)]} + [y_i - G(\mathbf{x}_i\beta)]\left(\frac{\nabla_\beta g(\mathbf{x}_i\beta)\mathbf{x}_i'\mathbf{x}_i G(\mathbf{x}_i\beta)[1-G(\mathbf{x}_i\beta)]}{\{G(\mathbf{x}_i\beta)[1-G(\mathbf{x}_i\beta)]\}^2} - \right.$$

$$\left. - \frac{g(\mathbf{x}_i\beta)\mathbf{x}_i'\{g(\mathbf{x}_i\beta)\mathbf{x}_i[1-G(\mathbf{x}_i\beta)] - g(\mathbf{x}_i\beta)\mathbf{x}_i G(\mathbf{x}_i\beta)\}}{\{G(\mathbf{x}_i\beta)[1-G(\mathbf{x}_i\beta)]\}^2}\right) =$$

$$= -\frac{[g(\mathbf{x}_i\beta)]^2\mathbf{x}_i'\mathbf{x}_i}{G(\mathbf{x}_i\beta)[1-G(\mathbf{x}_i\beta)]} +$$

$$+ [y_i - G(\mathbf{x}_i\beta)]\left(\frac{\nabla_\beta g(\mathbf{x}_i\beta)\mathbf{x}_i'\mathbf{x}_i}{G(\mathbf{x}_i\beta)[1-G(\mathbf{x}_i\beta)]} - \frac{[g(\mathbf{x}_i\beta)]^2\mathbf{x}_i'\mathbf{x}_i[1-2G(\mathbf{x}_i\beta)]}{\{G(\mathbf{x}_i\beta)[1-G(\mathbf{x}_i\beta)]\}^2}\right)$$

onde $\nabla_\beta g(\cdot)$ é a segunda derivada da função $G(\cdot)$

Funções Score e Hessiana (cont.)

Se aos resultados obtidos anteriormente aplicarmos o valor esperado condicionado verificamos que:

$$\rightarrow \; E\left[\mathbf{s}_i\left(\beta\right) \mid \mathbf{X}_i\right] = \frac{g\left(\mathbf{x}_i\beta\right)\mathbf{x}_i'}{G\left(\mathbf{x}_i\beta\right)\left[1-G\left(\mathbf{x}_i\beta\right)\right]} \underbrace{\left[E\left(Y_i \mid \mathbf{X}_i\right) - G\left(\mathbf{x}_i\beta\right)\right]}_{=0} = 0$$

$$\rightarrow \; E\left[\mathbf{H}_i\left(\beta\right) \mid \mathbf{X}_i\right] = -\frac{\left[g\left(\mathbf{x}_i\beta\right)\right]^2 \mathbf{x}_i'\mathbf{x}_i}{G\left(\mathbf{x}_i\beta\right)\left[1-G\left(\mathbf{x}_i\beta\right)\right]} + \underbrace{\left[E\left(Y_i \mid \mathbf{X}_i\right) - G\left(\mathbf{x}_i\beta\right)\right]}_{=0} \Upsilon\left(\mathbf{x}_i\beta\right)$$

Ou seja,

- A propriedade do valor esperado condicionado da função Score verifica-se, para uma qualquer função $G\left(\cdot\right)$
- A variância assimptótica do estimador ML,

$$\widehat{Avar\left(\hat{\beta}\right)} = \left\{\sum_{i=1}^{N} -E\left[\mathbf{H}_i\left(\beta\right) \mid \mathbf{X}_i\right]\right\}^{-1}$$

$\Rightarrow$ Apenas no caso em que a densidade de $Y \mid \mathbf{X}$ esteja bem especificada

Resumindo...

	Probit	Logit
$\mathbf{s}_i(\beta)$	$\frac{\phi(\mathbf{x}_i\beta)\mathbf{x}_i'[y_i-\Phi(\mathbf{x}_i\beta)]}{\Phi(\mathbf{x}_i\beta)[1-\Phi(\mathbf{x}_i\beta)]}$	$[y_i-\Lambda(\mathbf{x}_i\beta)]\mathbf{x}_i$
$\mathbf{H}_i(\beta)$	$-\frac{[\phi(\mathbf{x}_i\beta)]^2\mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\beta)[1-\Phi(\mathbf{x}_i\beta)]}+[y_i-G(\mathbf{x}_i\beta)]\varphi(\mathbf{x}_i\beta)$	$-\lambda(\mathbf{x}_i\beta)\mathbf{x}_i'\mathbf{x}_i$
$E[\mathbf{s}_i(\beta)\mid\mathbf{X}_i]$	0	0
$E[\mathbf{H}_i(\beta)\mid\mathbf{X}_i]$	$-\frac{[\phi(\mathbf{x}_i\beta)]^2\mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\beta)[1-\Phi(\mathbf{x}_i\beta)]}$	$-\lambda(\mathbf{x}_i\beta)\mathbf{x}_i'\mathbf{x}_i$
$\widehat{Avar}\left(\hat{\beta}\right)$	$\left\{\sum_{i=1}^{N}\frac{[\phi(\mathbf{x}_i\beta)]^2\mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\beta)[1-\Phi(\mathbf{x}_i\beta)]}\right\}^{-1}$	$\left\{\sum_{i=1}^{N}\lambda(\mathbf{x}_i\beta)\mathbf{x}_i'\mathbf{x}_i\right\}^{-1}$

NOTA: $\lambda(\cdot)=\Lambda(\cdot)[1-\Lambda(\cdot)]$

Observações

- Atente-se ao facto de não estarmos interessados nos efeitos parciais de x_j sobre y_i^* mas sim nos efeitos parciais de x_j sobre y_i, dado que:
  - Não observamos y_i^* mas sim y_i
  - Mesmo que observássemos y_i^*, a sua unidade de medida seria pouco clara, ou até inexistente (em especial, o exemplo da utilidade)

- Adicionalmente, destaca-se a utilidade da estimação por via de um modelo linear, na medida em que a **direcção** dos efeitos parciais é igual à versão dos efeitos parciais num modelo não linear:

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = \beta_j \quad \text{vs.} \quad \frac{\partial p(\mathbf{x})}{\partial x_j} = \beta_j \times \underbrace{g(\mathbf{X}\boldsymbol{\beta})}_{>0}$$