LISBOA SCHOOL OF ECONOMICS & MANAGEMENT
SINCE 1911

Capítulo 5 – Modelos de escolha Binária

Estimação, testes e efeitos parciais

Luís Silveira Santos

lsantos@iseg.ulisboa.pt

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

10 de Abril de 2016

Programa desta aula

1. Revisão
2. Estimação
3. Efeitos parciais
   - Efeito parcial na média
   - Efeito parcial médio
   - Cálculo dos erros-padrão
4. Testes de hipóteses
   - Testes de restrições de exclusão simples ou múltiplas
   - Testes de hipóteses não lineares
5. Interpretação dos resultados
   - Qualidade das previsões
   - R^2 de McFadden
   - Probit vs. Logit vs. MPL

Revisão

- Recordemos que pretendemos estimar:
$$P\left(y=1 \mid \mathbf{X}\right)=G\left(\mathbf{X}\beta\right)\equiv p\left(\mathbf{x}\right)$$
onde se assume que $G\left(\cdot\right)\in\left[0,1\right]$
- Se $\varepsilon \mid \mathbf{X}\sim N\left(0,1\right)$ então,
$$G\left(z\right)\equiv\Phi\left(z\right)=\int_{-\infty}^{z}\phi\left(z\right)\mathrm{d}z$$
- Se $\varepsilon \mid \mathbf{X}\sim Logistic\left(0,1\right)$ então,
$$G\left(z\right)\equiv\Lambda\left(z\right)=\frac{\exp\left(z\right)}{1-\exp\left(z\right)}$$
- Como $y\in\{0,1\}$, sabemos que, em geral, $Y\mid\mathbf{X}\sim Ber\left(G\left(\mathbf{X}\beta\right)\right)$

Estimação

- A estimação de modelos não lineares pode ser feita por duas vias:
  1. Mínimos quadrados não lineares (NLS)
  2. Máxima Verosimilhança (ML)
- Do método NLS iremos obter:
  - Estimadores consistentes e $\sqrt{N}$ assimptoticamente Normais
  - Inferência robusta a heterocedasticidade com forma funcional genérica
- 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 no caso em que a densidade de $Y \mid \mathbf{X}$ esteja bem especificada

Estimação (cont.)

- Como estamos a assumir 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):

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

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

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

  - 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:

$$
\begin{aligned}
\mathbf{s}_i(\boldsymbol{\beta}) = \nabla_{\beta}\ell_i(\boldsymbol{\beta})' = \\
&= y_i\left[\frac{g(\mathbf{x}_i\boldsymbol{\beta})\,\mathbf{x}_i'}{G(\mathbf{x}_i\boldsymbol{\beta})}\right] - (1-y_i)\left[\frac{g(\mathbf{x}_i\boldsymbol{\beta})\,\mathbf{x}_i'}{1-G(\mathbf{x}_i\boldsymbol{\beta})}\right] = \\
&= \frac{g(\mathbf{x}_i\boldsymbol{\beta})\,\mathbf{x}_i'\left\{y_i\left[1-G(\mathbf{x}_i\boldsymbol{\beta})\right]-(1-y_i)\,G(\mathbf{x}_i\boldsymbol{\beta})\right\}}{G(\mathbf{x}_i\boldsymbol{\beta})\left[1-G(\mathbf{x}_i\boldsymbol{\beta})\right]} = \\
&= \frac{g(\mathbf{x}_i\boldsymbol{\beta})\,\mathbf{x}_i'}{G(\mathbf{x}_i\boldsymbol{\beta})\left[1-G(\mathbf{x}_i\boldsymbol{\beta})\right]}\left[y_i - G(\mathbf{x}_i\boldsymbol{\beta})\right]
\end{aligned}
$$

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:

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

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

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

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

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

onde $\nabla_{\boldsymbol{\beta}} g(\cdot)$ 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:

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

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

Ou seja,

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

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

<u>se</u> a densidade de $Y \mid \mathbf{X}$ estiver bem especificada

Resumindo...

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

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

Efeitos parciais

- Recordemos da aula anterior que os efeitos parciais obtidos por via destes modelos não lineares são dados por:
$$\frac{\partial P\left(Y=1 \mid \mathbf{X}\right)}{\partial x_j}=\frac{\partial p\left(\mathbf{x}\right)}{\partial x_j}=\beta_j \times g\left(\mathbf{X}\beta\right)$$
- No entanto a função $g\left(\mathbf{X}\beta\right)$ depende dos valores que cada elemento da matriz $\mathbf{X}$ pode assumir
- A solução passa por calcular dois tipos de efeitos parciais:
    1. Efeito parcial na média (PEA)
    2. Efeito parcial médio (APE)

Efeito parcial na média

- Neste contexto, vamos calcular a função $g(\mathbf{X}\beta)$ para valores "representativos" de cada elemento da matriz $\mathbf{X}$:
  1. Médias
  2. Máximos e Mínimos

- Desta forma podemos:
  1. Analisar os efeitos parciais no indivíduo "médio"
  2. Proceder a uma análise de sensibilidade dos efeitos parciais para valores extremos de x_j, $j = 1, 2, \ldots, K$

Efeitos parciais na média (cont.)

- Se x_j é variável contínua:

$$\widehat{PEA}_j = \frac{\partial \widehat{P\left(Y = 1 \mid \mathbf{X} = \overline{\mathbf{X}}\right)}}{\partial x_j} = \hat{\beta}_j \times g\left(\overline{\mathbf{X}}\hat{\beta}\right)$$

- Se x_j é variável discreta:

$$\widehat{PEA}_j = G\left[\hat{\beta}_0 + \hat{\beta}_1\overline{x}_1 + \hat{\beta}_2\overline{x}_2 + \ldots + \hat{\beta}_{j-1}\overline{x}_{j-1} + \hat{\beta}_j\left(c_j + 1\right)\right] - \\ - G\left(\hat{\beta}_0 + \hat{\beta}_1\overline{x}_1 + \hat{\beta}_2\overline{x}_2 + \ldots + \hat{\beta}_{j-1}\overline{x}_{j-1} + \hat{\beta}_j c_j\right)$$

- Quando x_j é variável binária, podemos utilizar o PEA_j anterior, fixando $c_j = 0$

Efeito parcial médio

- Uma alternativa passa por calcular a média da função $g\left(\mathbf{X}\beta\right)$:

$$APE_j = \beta_j \times E\left[g\left(\mathbf{X}\beta\right)\right]$$

- Se x_j é variável contínua, um estimador consistente será:

$$\widehat{APE}_j = \hat{\beta}_j \left[\frac{1}{N}\sum_{i=1}^{N} g\left(\mathbf{x}_i\hat{\beta}\right)\right]$$

- Se x_j é variável discreta, um estimador consistente será:

$$\widehat{APE}_j = \frac{1}{N}\sum_{i=1}^{N}\left\{G\left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_{j-1} x_{i,j-1} + \hat{\beta}_j\left(c_j + 1\right)\right] - \right.$$
$$\left. -G\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_{j-1} x_{i,j-1} + \hat{\beta}_j c_j\right)\right\}$$

- Quando x_j é variável binária, fixamos $c_j = 0$ no APE_j anterior

Cálculo dos erros-padrão

- A maioria dos *softwares* econométricos permite-nos realizar inferência sobre para as estimativas dos *PEA*s e dos *APE*s, através do cálculo dos erros-padrão
- Estes erros-padrão são habitualmente calculados numericamente através de duas metodologias:
  - Método Delta
  - *Bootstrap*

Testes de hipóteses

- Tal como acontece no caso dos modelos lineares, estamos interessados em realizar testes de hipóteses sobre os parâmetros do modelo estimado
- Iremos dar enfoque a dois tipos de testes:
  1. Testes de restrições de exclusão simples ou múltiplas
  2. Testes de hipóteses não lineares sobre os parâmetros β
- As estatísticas de teste que seguidamente se irão apresentar são assimptoticamente equivalentes, porém algumas delas poderão ser computacionalmente difíceis de obter

Testes de restrições de exclusão simples ou múltiplas

- Considere-se o seguinte modelo:

$$P(Y = 1 \mid \mathbf{X}, \mathbf{Z}) = G(\mathbf{X}\beta + \mathbf{Z}\gamma)$$

  onde $\mathbf{Z}$ é matriz $N \times Q$ e γ é vector de dimensão $Q \times 1$

- Pretendemos testar a hipótese:

$$H_0 : \gamma = 0$$

- Existem três estatísticas de teste em opção:
  1. Estatística de Wald
  2. Rácio de verosimilhanças (ou teste LR)
  3. Teste de Score (ou teste LM)

Testes de restrições de exclusão simples ou múltiplas (cont.)

1. **ESTATÍSTICA DE WALD:**

$$W = (R\hat{\gamma} - r)' \left(R\hat{V}R'\right)^{-1} (R\hat{\gamma} - r) \xrightarrow{d} \chi^2_{(Q)}$$

onde $\hat{V}$ é a matriz de variância assimptótica do modelo com $G(\mathbf{X}\beta + \mathbf{Z}\gamma)$

2. **RÁCIO DE VEROSIMILHANÇAS:**

$$LR = 2(\mathscr{L}_{ur} - \mathscr{L}_r) \xrightarrow{d} \chi^2_{(Q)}$$

onde $\mathscr{L}_{ur}$ é a função log-verosimilhança para o modelo com $G(\mathbf{X}\beta + \mathbf{Z}\gamma)$ e $\mathscr{L}_r$ é a função log-verosimilhança para o modelo com $G(\mathbf{X}\beta)$

Testes de restrições de exclusão simples ou múltiplas (cont.)

3. **TESTE DE SCORE:**

$$LM = NR_u^2 \xrightarrow{d} \chi^2_{(Q)}$$

onde R_u^2 é o R^2 não centrado da regressão auxiliar, estimada por OLS, de:

$$\frac{\hat{u}_i}{\sqrt{\hat{G}_i\left(1-\hat{G}_i\right)}} \text{ sobre } \frac{\hat{g}_i}{\sqrt{\hat{G}_i\left(1-\hat{G}_i\right)}}\mathbf{x}_i \text{ e } \frac{\hat{g}_i}{\sqrt{\hat{G}_i\left(1-\hat{G}_i\right)}}\mathbf{z}_i$$

com $\hat{u}_i = y_i - G\left(\mathbf{x}_i\hat{\beta}\right)$, $\hat{G} = G\left(\mathbf{x}_i\hat{\beta}\right)$ e $\hat{g} = g\left(\mathbf{x}_i\hat{\beta}\right)$

$\Rightarrow$ Note-se que se pondera por $\left[\hat{G}_i\left(1-\hat{G}_i\right)\right]^{-1/2}$ pois, sob H_0,

$$Var\left(u_i \mid \mathbf{X},\mathbf{Z}\right) = G\left(\mathbf{x}_i\beta\right)\left[1 - G\left(\mathbf{x}_i\beta\right)\right]$$

Testes de hipóteses não lineares

- Relativamente aos testes de hipóteses não lineares, apesar de podermos considerar novamente as estatísticas anteriores, existem problemas computacionais relativamente ao seu cálculo
- Em concreto, impor-se restrições não lineares na estimação dos modelos Probit e Logit eleva substancialmente o nível de complexidade computacional de maximização da função de verosimilhança, com implicações directas no cálculo das estatísticas de teste LM e LR
- Resta-nos, portanto, a estatística de Wald. No entanto, esta estatística não é invariante a reparameterizações (ao contrário das estatísticas de teste LM e LR), traduzindo-se em fracas propriedades de amostras finitas

Testes de hipóteses não lineares (cont.)

- Ainda assim, iremos optar por calcular a estatística de Wald, uma vez que os resultados assimptóticos permanecem válidos

- Tendo por base o modelo com $G(\mathbf{X}\beta)$, vamos considerar que pretendemos testar uma restrição não linear genérica, i.e.,

$$H_0 : \mathbf{c}(\beta) = 0$$

onde $\mathbf{c}(\beta)$ é vector de dimensão $K_1 \times 1$, com $K_1 \leq K$

- Neste caso, a estatística de Wald é dada por:

$$W = \mathbf{c}\left(\hat{\beta}\right)' \left(\nabla_\beta \mathbf{c}\left(\hat{\beta}\right) \hat{V} \; \nabla_\beta \mathbf{c}\left(\hat{\beta}\right)'\right)^{-1} \mathbf{c}\left(\hat{\beta}\right) \xrightarrow{d} \chi^2_{(K_1)}$$

onde $\nabla_\beta \mathbf{c}\left(\hat{\beta}\right)$ é o jacobiano do vector $\mathbf{c}(\beta)$ avaliado em $\hat{\beta}$

Qualidade do ajustamento

- Um outro aspecto importante da nossa análise será avaliar o potencial do modelo estimado
- Para tal, temos duas medidas da qualidade do ajustamento:
  1. Qualidade das previsões
  2. R^2 de McFadden (ou pseudo-R^2)
- Note-se, porém, que não devemos esperar valores convincentes destas medidas

Qualidade das previsões

- Sabe-se que:

$$\hat{y}_i = \begin{cases} 1 & \text{se } G\left(\mathbf{x}_i\hat{\beta}\right) \geq 0.5 \\ 0 & \text{se } G\left(\mathbf{x}_i\hat{\beta}\right) < 0.5 \end{cases}$$

- Vamos então definir:
  - $N_0 = \#\left(y_0\right)$, número de "insucessos" na amostra
  - $N_1 = \#\left(y_1\right)$, número de "sucessos" na amostra
  - $N_{00} = \#\left(\hat{y}_0\right)$, número de "insucessos" bem previstos
  - $N_{11} = \#\left(\hat{y}_1\right)$, número de "sucessos" bem previstos

Qualidade das previsões (cont.)

- Uma medida da qualidade das previsões envolve a relação entre a proporção dos "sucessos" e dos "insucessos" bem previstos:

$$q_0 = \frac{N_{00}}{N_0} \text{ e } q_1 = \frac{N_{11}}{N_1}$$

- Se alguma das proporções acima apresentar um valor demasiado baixo, poderemos ajustar o *threshold* das probabilidades estimadas de 0.5 para $\bar{y}$, uma vez que este é um estimador consistente para a probabilidade de sucesso não condicionada

- A proporção total de previsões correctas é dada por:

$$q = \frac{N_{00} + N_{11}}{N} = \frac{N_0}{N} q_0 + \frac{N_1}{N} q_1$$

R^2 de McFadden

- O R^2 de McFadden é dado por

$$R^2_{\text{Pseudo}} = 1 - \frac{\mathscr{L}_{ur}}{\mathscr{L}_0}$$

onde $\mathscr{L}_{ur}$ é a função log-verosimilhança para o modelo estimado e $\mathscr{L}_0$ é a função log-verosimilhança para o modelo apenas com a constante

Probit vs. Logit vs. MPL

- A escolha entre o Probit e o Logit é, em geral, difícil
- Uma vez que ambos têm implícita uma função de distribuição, a decisão recai sobre a caracterização dos valores extremos dos nossos dados
- Nos casos em que exista uma grande densidade de valores extremos, deveremos optar pelo Logit, uma vez que a distribuição Logística estandardizada tem caudas mais pesadas

Probit vs. Logit vs. MPL (cont.)

- No entanto, é possível comparar as magnitudes das estimativas entre os três modelos através de factores de escala:
  - MPL $\rightarrow g(0) = 1$
  - Probit $\rightarrow g(0) = \frac{1}{\sqrt{2\pi}} \approx 0.4$
  - Logit $\rightarrow g(0) = \frac{\exp(0)}{1+\exp(0)} \times \left[1 - \frac{\exp(0)}{1+\exp(0)}\right] = 0.25$
- **EXEMPLO:** dividir as estimativas Logit por $0.4/0.25 = 1.6$ para torná-las comparáveis com as estimativas Probit