# MODELOS DE ESCOLHA BINÁRIA

## Introdução

Em muitas aplicações do modelo de regressão linear, pressupõe-se que a variável dependente é uma variável aleatória contínua e que tem por domínio o conjunto dos números reais.

Dois aspectos em que essa perspectiva se manifesta de modo óbvio ocorrem na interpretação dos coeficientes de regressão e nas hipóteses quanto à distribuição de probabilidade da variável explicada. Quando se escreve, a propósito de um coeficiente de regressão $\beta_j$,

$$\beta_j = \frac{\partial Y}{\partial X_j},$$

presume-se, implicitamente, ser Y uma função diferenciável em ordem a $X_j$. De modo análogo, a hipótese da normalidade das perturbações, em que assentam os procedimentos usuais de inferência, supõe que a variável dependente seja uma variável aleatória contínua.

Esses aspectos não são, contudo, essenciais na teoria do modelo de regressão linear. Se, por exemplo, $X_j$ for uma variável binária, caso em que a derivada parcial a que se aludiu acima não é definida, é possível modificar a forma de interpretação de $\beta_j$ para acomodar essa situação. E, se não for normal a distribuição de probabilidade das perturbações, haverá que modificar os procedimentos de inferência, mas os resultados fundamentais em matéria de estimação não são afectados.

Existe, no entanto, uma grande variedade de situações em que não é defensável admitir que a variável dependente tenha natureza contínua ou domínio real.

Em primeiro lugar, há casos em que a variável dependente é uma variável discreta: número de empresas falidas num ano, número de empregos criados ou destruídos, número de balcões de uma instituição de crédito, etc. O domínio da variável é, nestes exemplos, o conjunto dos números inteiros e a hipótese de normalidade é, claramente, inadequada.

Noutros casos, a variável dependente não é, sequer, quantitativa: ter ou não ter casa própria, deslocar-se para o trabalho de combóio, de autocarro, a pé, em viatura própria ou de outro modo, votar a favor, contra, ou abster-se numa deliberação, etc. Conquanto seja habitual fazer-se corresponder números inteiros às diversas categorias (por exemplo, 1 à posse de casa própria e 0 ao caso contrário), esses números não traduzem verdadeiramente uma quantificação, mas constituem, antes, um dispositivo de codificação essencialmente arbitrário.

Em terceiro lugar, citem-se casos em que, conquanto a natureza de variável contínua seja aceitável, não é válido admitir-se que Y tenha $\Re$ por domínio. Se a variável dependente for uma probabilidade, por exemplo, o domínio deverá ser restringido ao intervalo [0, 1].

Noutros casos, ainda, embora a variável de interesse possa ser uma variável contínua, pode não ser observável em todo o domínio. Num exercício de tiro ao alvo, a distância a que a bala passa do centro só é observada, tipicamente, para os projécteis que embateram num raio limitado em torno do centro; quanto aos outros, sabe-se apenas que a distância excedeu esse raio. As restrições à observabilidade da variável podem ser ainda mais drásticas e limitá-la a uma mera informação qualitativa, do tipo "acertou" ou "não acertou".

Dispõe-se, em Econometria, de uma gama muito vasta de modelos para analisar problemas em que a variável dependente é discreta, qualitativa, limitada, censurada ou truncada. Constituem objecto deste texto apenas alguns modelos elementares dessa classe, os chamados modelos de escolha binária. Trata-se de modelos cuja utilização mais frequente em Economia teve por objectivo o estudo das escolhas de um agente e em que a variável dependente é de natureza binária: a escolha faz-se entre duas alternativas e uma, ou outra, tem de ser escolhida. Além dos modelos com variável dependente discreta ou limitada, fora da análise ficarão os modelos de escolha multinomial, em que a escolha se faz entre mais de duas alternativas.

São muito numerosos os exemplos de emprego de modelos de escolha binária:

i) Em estudos da oferta de trabalho, é frequente modelizar-se a decisão de participação ou não na força de trabalho como função de uma série vasta de atributos individuais (sexo, grau de instrução, idade, etc.) ou familiares (estado civil, número e idade dos filhos, rendimento do agregado, etc.) e, ainda, de atributos dos empregos disponíveis (remunerações, horas de trabalho, etc.). A variável dependente é codificada com o valor 1 se o iº indivíduo participa na força de trabalho, ou com o valor 0 em caso contrário.

ii) Em estudos dos determinantes da aquisição de bens duradouros (automóveis, casas, etc.), é também frequente relacionar-se a posse, ou não, de um certo bem com factores explicativos como o rendimento, o preço, a taxa de juro, etc.

iii) Em estudos de comportamentos eleitorais, usam-se, para explicar a verificação do acontecimento "o iº indivíduo votou no partido Z", ou da alternativa "o iº indivíduo não votou no partido Z", variáveis explicativas como o rendimento do indivíduo, o local de residência, a religião ou outras.

iv) Em estudos da escolha de modos de transporte, a variável dependente poderia corresponder ao acontecimento "uso de transporte público" ou a "uso de outros modos de transporte", em função de variáveis como os preços dos transportes, o rendimento do utente, a distância a percorrer, etc.

v) Em estudos de comportamentos migratórios, a decisão de emigrar, ou não, é explicada em função dos salários na região de origem e na de destino, de características pessoais dos migrantes, etc.

vi) Em estudos da procura de educação e, em particular, de educação superior: concluído um curso de nível secundário, o indivíduo escolhe entre o ingresso imediato no mercado de trabalho ou a continuação dos estudos, em função das oportunidades de emprego, dos fluxos de rendimentos esperados numa e noutra alternativa, dos custos da frequência de uma universidade, etc.

vii) Em estudos sobre mercados monetários e financeiros, têm sido analisadas a probabilidade de recusa ou concessão de crédito pelas instituições financeiras e a probabilidade de cumprimento ou incumprimento de obrigações pelos devedores. A informação disponível, quanto à variável dependente, é limitada às alternativas "empréstimo concedido" vs. "empréstimo recusado", ou "prestações cumpridas" vs. "prestações não cumpridas".

## Modelos lineares de probabilidade

Considere-se a habitual equação de regressão linear

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i, \qquad [1]$$

ou, numa notação mais conveniente,

$$Y_i = \mathbf{X_i} \boldsymbol{\beta} + u_i \qquad [2]$$

(em que $\mathbf{X_i}$ é o vector (1×k) de componentes 1, $X_{2i}$, $X_{3i}$, ... , $X_{ki}$ e $\boldsymbol{\beta}$ o vector (k×1) de coeficientes de regressão), e admita-se que $Y_i$ designa uma variável binária codificada com o valor 1 ou o valor 0. Usualmente, o valor 1 é atribuído à presença de um certo atributo na iª observação, enquanto o valor 0 é atribuído à sua ausência. Noutra perspectiva, podem ver-se os dois valores possíveis como correspondendo à verificação, ou não, de um certo acontecimento pela iª observação.

De acordo com o modelo estatístico subjacente às hipóteses clássicas, num processo de amostragem repetida, para uma mesma sequência ordenada ($X_{2i}$, $X_{3i}$, ... , $X_{ki}$) deveria ser possível obter valores diferentes de Y. Na especificação em análise, seriam dois, apenas, os valores possíveis: o acontecimento em causa seria observado para alguns indivíduos — e ter-se-ia $Y_i = 1$ —, e não seria observado nos restantes — para os quais, $Y_i = 0$.

Seja $p_i$ a probabilidade de verificação do acontecimento, isto é, $p_i = \text{Prob}(Y_i=1|X_{2i}, X_{3i}, ... , X_{ki})$ e, por conseguinte, seja $1-p_i$ a probabilidade de não verificação do acontecimento. Então, o valor médio de $Y_i$ (condicional a $X_{2i}$, $X_{3i}$, ... , $X_{ki}$), caso exista, terá de ser igual a $p_i$:

$$E(Y_i) = 0(1-p_i) + 1\ p_i = p_i. \qquad [3]$$

Por conseguinte, se for $E(u_i) = 0$ na equação [1], deverá ter-se

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} = p_i. \qquad [4]$$

Sendo o valor médio condicional de Y a probabilidade (condicional a $X_2, X_3, ... , X_k$) de Y = 1, modelos com esta estrutura são conhecidos por modelos lineares de probabilidade ou pela sigla inglesa LPM (*linear probability models*).

De um ponto de vista teórico, os modelos LPM apresentam alguns aspectos insatisfatórios. Se o modelo for linear na variável $X_j$, e as derivadas abaixo existirem, então,

$$\frac{\partial E(Y)}{\partial X_j} = \frac{\partial\, Prob(Y=1)}{\partial X_j} = \beta_j,$$

ou seja, tudo o resto igual, a probabilidade de verificação do acontecimento em análise é afectada sempre de modo idêntico por uma variação de $X_j$, qualquer que seja o nível da variável. Por exemplo, um aumento de 100 para 200 no rendimento de um indivíduo, tudo o mais constante, faria variar a probabilidade de aquisição de casa própria pelo mesmo montante em que o faria um aumento de 10000 para 10100.

Uma vez que E(Y) é uma medida de probabilidade, terá de ser $0 \le E(Y) \le 1$. Mas é difícil compatibilizar a restrição $0 \le E(Y) \le 1$ com a possibilidade de variações absolutas constantes em E(Y) induzidas por variações unitárias de uma variável explicativa. Se E(Y) é uma função linear de $X_j$, é inevitável que, para alguns valores de $X_j$, E(Y) venha a situar-se fora do intervalo admissível. Seria mais natural que, exercendo $X_j$ um efeito positivo na probabilidade de ser Y = 1, esse efeito fosse praticamente nulo para valores muito baixos ou muito altos de $X_j$ e fosse mais sensível para algum intervalo de valores intermédios da variável. Isto é, seria de esperar que fosse

$$\lim_{\mathbf{x_i}\,\boldsymbol{\beta} \to +\infty} Prob(Y=1) = 1$$

e

$$\lim_{\mathbf{x_i}\,\boldsymbol{\beta} \to -\infty} Prob(Y=1) = 0 .$$

Por outro lado, a aplicação de algumas das técnicas de estimação e de análise estatística comummente empregues com o modelo clássico de regressão linear suscita dificuldades sérias em modelos LPM:

Em primeiro lugar, terá de ser, como se viu, $0 \le E(Y) \le 1$. Mas a estimação pelo método ordinário de mínimos quadrados (OLS) dos coeficientes em [4] não obedecerá necessariamente a essa restrição, pelo que é possível que origine valores estimados de Y negativos ou superiores à unidade.

Em segundo lugar, não é aceitável a hipótese da normalidade para a variável dependente (que, recorde-se, é uma variável discreta), nem para a perturbação $u_i$. Esta poderá assumir apenas dois valores,

$$1 - (\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki}),$$

com probabilidade $p_i$, ou

$$0 - (\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki}),$$

com probabilidade $1-p_i$, pelo que possuirá uma distribuição binomial. A consequência mais importante de tal facto será a de invalidar as técnicas de inferência estatística apresentadas anteriormente, em amostras de dimensão finita. O melhor que se poderá esperar é que, em amostras de grande dimensão, os resultados obtidos com essas técnicas sejam uma boa aproximação aos verdadeiros.

Em terceiro lugar, não é sustentável a hipótese de homoscedasticidade relativamente à perturbação u. Uma vez que $E(u_i) = 0$, será $Var(u_i) = E(u_i^2)$. Ora, $u_i^2$ poderá apenas assumir dois valores,

$$[1 - (\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki})]^2,$$

com probabilidade $p_i$, e

$$[ - (\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki})]^2,$$

com probabilidade $1-p_i$, pelo que

$$E(u_i^2) = [1 - (\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki})]^2 \, p_i +$$

$$+ (\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki})^2 \, (1-p_i).$$

Notando (da equação [4]) que

$$p_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki},$$

é fácil simplificar a expressão anterior para obter

$$Var(u_i) = E(u_i^2) = p_i \, (1-p_i), \qquad [5]$$

com a implicação de heteroscedasticidade.

Conquanto as dificuldades citadas (não normalidade, heteroscedasticidade, não obrigatoriedade de verificação de $0 \leq \hat{Y} \leq 1$) possam ser contornadas na estimação de modelos LPM, têm maior voga nas aplicações empíricas outras formulações que, do ponto de vista teórico, não apresentam inconvenientes como os que acima se apontaram. Avultam, entre elas, os modelos habitualmente designados pelas expressões *probit* e *logit*.

Modelos *probit* e *logit*

Uma abordagem mais satisfatória dos modelos em que a variável dependente é dicotómica é a que pressupõe que a variável Y, com Y = 1 ou Y = 0, é apenas a manifestação observável de uma variável não observável Y* (dita variável latente) tal que

$$Y_i^* = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i = \mathbf{X_i}\,\boldsymbol{\beta} + u_i \qquad [6]$$

e em que é especificada uma regra de determinação de Y em função de Y*. Essa regra é, tipicamente, da forma[1]

$$Y_i = \begin{cases} 1, \text{ se } Y_i^* \geq 0 \\ 0, \text{ se } Y_i^* < 0 \end{cases} . \qquad [7]$$

Note-se que, nesta formulação, quer a variável dependente latente, quer a perturbação aleatória, podem ser validamente definidas como variáveis aleatórias contínuas e o carácter discreto é reservado apenas para a contrapartida observável da variável de interesse.

Pode conceber-se a variável latente $Y_i^*$ como a diferença entre a utilidade, $U_{1i}$, que para o iº indivíduo teria a alternativa representada por $Y_i = 1$ e a utilidade, $U_{0i}$, associada com a alternativa $Y_i = 0$, isto é,

$$Y_i^* = U_{1i} - U_{0i} = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i.$$

O indivíduo optaria pela primeira alternativa se $U_{1i} > U_{0i}$ e pela segunda se $U_{1i} \leq U_{0i}$, ou seja, verificar-se-ia $Y_i = 1$ ou $Y_i = 0$ conforme fosse, respectivamente, $Y_i^* > 0$ ou $Y_i^* \leq 0$.

Outra formulação possível é a que vê $Y_i^*$ como um índice da propensão do iº indivíduo para a escolha de uma alternativa. Esse índice poderia corresponder à diferença entre $R_{1i}$, o rendimento ou benefício marginal esperado se o indivíduo fizer a escolha da alternativa associada com Y = 1, e $C_{1i}$, custo marginal dessa escolha. De maneira análoga à anterior, ter-se-ia agora

$$Y_i^* = R_{1i} - C_{1i} = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i,$$

$Y_i = 1$ se $R_{1i} \geq C_{1i}$ (e, portanto, $Y_i^* \geq 0$), ou $Y_i = 0$ se $Y_i^* < 0$.

[1] A escolha da constante 0 como limiar de separação entre Y = 0 e Y = 1 é, essencialmente, arbitrária. O assunto é discutido mais extensamente noutro ponto do texto.

Na classe de modelos caracterizada pelas relações [6] e [7], é

$$\begin{aligned}\mathrm{Prob}(Y_i = 1) &= \mathrm{Prob}(Y_i^* > 0) = \\ &= \mathrm{Prob}(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i > 0) = \\ &= \mathrm{Prob}(u_i > - \mathbf{X_i}\, \boldsymbol{\beta})\end{aligned}$$

e, por conseguinte,

$$\mathrm{Prob}(Y_i = 0) = \mathrm{Prob}(u_i \leq - \mathbf{X_i}\, \boldsymbol{\beta}).$$

Então, sendo $u_i$ uma variável aleatória com função de distribuição F(.), vem

$$\begin{aligned}\mathrm{Prob}(Y_i = 0) &= F(- \mathbf{X_i}\, \boldsymbol{\beta}), \\ \mathrm{Prob}(Y_i = 1) &= 1 - F(- \mathbf{X_i}\, \boldsymbol{\beta}).\end{aligned} \qquad [8]$$

As duas escolhas mais comuns para a forma funcional de F(.) são as que correspondem à distribuição normal reduzida e à distribuição logística. No modelo *probit*, é postulado que $u_i$ tem distribuição normal reduzida e F(.) designa, então, a função de distribuição normal,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t^2}\, dt\,; \qquad [9]$$

a função de densidade de probabilidade associada é dada, como é sabido, por

$$\phi(x) = \frac{d\Phi(x)}{dx} = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}. \qquad [10]$$

No modelo *logit*, a escolha de F(.) recai em

$$\Lambda(x) = \frac{1}{1 + e^{-x}}, \qquad [11]$$

a função de distribuição de uma variável logística de média nula e variância $\frac{\pi^2}{3}$. A função de densidade da logística é

$$\lambda(x) = \frac{d\Lambda(x)}{dx} = \frac{e^{-x}}{(1 + e^{-x})^2}; \qquad [12]$$

É fácil verificar que $\lambda(x) = \Lambda(x)\, [1 - \Lambda(x)]$.

As duas distribuições consideradas têm características de simetria de que resultam algumas propriedades interessantes. Uma delas, por exemplo, é a de que

$$F(x) = 1 - F(-x),$$

designe F(.) a função em [9] ou a função em [11]. Isso explica por que é possível dar às expressões em [8] um aspecto diferente:

$$\begin{aligned} &\text{Prob}(Y_i = 0) = 1 - F(\mathbf{X_i}\,\boldsymbol{\beta}), \\ &\text{Prob}(Y_i = 1) = F(\mathbf{X_i}\,\boldsymbol{\beta}). \end{aligned} \qquad [13]$$

Do confronto da última destas equações com a equação de regressão dos modelos LPM é fácil concluir que, nestes, se procede, implicitamente, como se fora $F(\mathbf{X_i}\,\boldsymbol{\beta}) = \mathbf{X_i}\,\boldsymbol{\beta}$. Não é difícil identificar concretizações de $\mathbf{X_i}$ e $\boldsymbol{\beta}$ para as quais se tenha $\mathbf{X_i}\,\boldsymbol{\beta} > 1$ ou $\mathbf{X_i}\,\boldsymbol{\beta} < 0$, em violação do requerido para uma probabilidade. Ao invés, a própria especificação dos modelos *probit* e *logit* garante (por ser $0 \leq F(\mathbf{X_i}\,\boldsymbol{\beta}) \leq 1$, quaisquer que sejam $\mathbf{X_i}$ e $\boldsymbol{\beta}$) que tal nunca sucederá.

Por outro lado, os coeficientes integrados no vector $\boldsymbol{\beta}$ nos modelos *probit* e *logit* não têm a interpretação usual dos modelos de regressão linear. De facto, tem-se, usando [13],

$$E(Y_i) = 0[1 - F(\mathbf{X_i}\,\boldsymbol{\beta})] + 1\,F(\mathbf{X_i}\,\boldsymbol{\beta}),$$

ou seja,

$$E(Y_i) = F(\mathbf{X_i}\,\boldsymbol{\beta}). \qquad [14]$$

Portanto,

$$\frac{\partial\, E(Y_i)}{\partial\, \mathbf{X}_i'} = \frac{d\,F(\mathbf{X}_i\,\boldsymbol{\beta})}{d(\mathbf{X}_i\,\boldsymbol{\beta})}\frac{\partial(\mathbf{X}_i\,\boldsymbol{\beta})}{\partial \mathbf{X}_i'} = f(\mathbf{X_i}\,\boldsymbol{\beta})\,\boldsymbol{\beta}, \qquad [15]$$

em que f(.) designa a função de densidade correspondente à função de distribuição F(.).

Na equação [15], o primeiro membro é um vector-coluna de derivadas parciais, de que a componente genérica é, supondo que $E(Y_i)$ é uma função linear de $X_j$,

$$\frac{\partial E(Y_i)}{\partial X_{ji}} = f(\mathbf{X_i}\,\boldsymbol{\beta})\,\beta_j. \qquad [16]$$

Resulta de [16] que, nos modelos *probit* e *logit*, $\beta_j$ não mede necessariamente o efeito marginal de $X_j$ sobre E(Y) (ou, o que é o mesmo, sobre Prob(Y = 1)). Relembre-se, contudo, que continua a ser, salvo não linearidade em relação a $X_j$,

$$\frac{\partial\, E(Y_i^*)}{\partial\, X_{ji}} = \beta_j$$

para qualquer i e qualquer j, relativamente à variável não observável Y*, à semelhança do que ocorria na generalidade dos modelos de regressão estudados anteriormente.

Três factos merecem realce em [16]:

i) O efeito marginal de $X_j$ sobre Prob(Y = 1) é variável de indivíduo para indivíduo (devido à presença de $X_{2i}$, ..., $X_{ji}$, ..., $X_{ki}$ em $f(\mathbf{X_i}\ \boldsymbol{\beta})$).

ii) Esse efeito é também variável com $X_j$; isto é, para o mesmo indivíduo, esse efeito será diferente de um valor de $X_{ji}$ para outro.

iii) Por último, note-se que, para o iº indivíduo, os efeitos marginais descritos em [16] dependem, além do coeficiente associado, de um factor de proporcionalidade idêntico, $f(\mathbf{X_i}\ \boldsymbol{\beta})$, qualquer que seja o regressor ($X_2$, ou $X_3$, etc., ou $X_k$) considerado. Uma vez que esse factor é, geralmente, positivo (recorde-se que f(.) designa uma função de densidade), pode concluir-se que uma variação infinitesimal de $X_{ji}$, tudo o resto igual, induz uma variação positiva ou negativa na probabilidade de ser Y = 1, conforme for positivo ou negativo o coeficiente $\beta_j$.

Em conclusão, enquanto na generalidade dos modelos de regressão linear a estimativa de um coeficiente fornece indicação sobre o sentido, positivo ou negativo, da influência de uma variável explicativa, $X_j$, e sobre a grandeza dessa influência (idêntica para todos os indivíduos e, salvo não linearidade relativamente às variáveis independentes, independente do valor de $X_j$), nos modelos *probit* e *logit* a estimativa apenas dá, de imediato, informação sobre o sentido da influência. Informação sobre a grandeza do efeito requererá cálculos adicionais, que conduzirão, geralmente, a resultados diferentes de indivíduo para indivíduo e dependentes do valor de $X_j$.

Embora, de um ponto de vista teórico, seja mais fácil justificar a adopção da especificação normal para a função F(.), a especificação que conduz à análise *logit* tem sido a utilizada com maior frequência, por razões que se prendem com a simplicidade da expressão analítica da sua função de distribuição (por contraste com o integral requerido em [9]) e com a facilidade de cálculo na fase de estimação. As distribuições normal e logística têm grande semelhança para argumentos na vizinhança de 0, e é nas abas da distribuição que a diferença é mais perceptível, com a logística a apresentar abas "mais espessas". Para amostras em que a proporção de observações com Y = 1 e a proporção com Y = 0 sejam equilibradas, não é de esperar grandes diferenças de resultados entre as duas especificações.

As vantagens calculatórias da opção pelo *logit* são óbvias se se notar que, sendo

$$E(Y) = \text{Prob}(Y = 1) = \frac{1}{1+e^{-x}}$$

e, por conseguinte,

$$\text{Prob}(Y = 0) = 1 - \frac{1}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}},$$

vem

$$\frac{\text{Prob}(Y = 1)}{\text{Prob}(Y = 0)} = e^{x}$$

e

$$\ln\left(\frac{\text{Prob}(Y = 1)}{\text{Prob}(Y = 0)}\right) = x.$$

Tendo presente que o argumento x nas expressões acima é, neste contexto, $\mathbf{X_i}\ \boldsymbol{\beta}$, conclui-se que

$$\ln\left(\frac{\text{Prob}(Y_i = 1)}{\text{Prob}(Y_i = 0)}\right) = \mathbf{X_i}\ \boldsymbol{\beta} = \beta_1 + \beta_2\ X_{2i} + \beta_3\ X_{3i} + ... + \beta_k\ X_{ki}, \qquad [17]$$

com a variável no primeiro membro da equação a exibir linearidade relativamente aos coeficientes de interesse. Essa variável é o logaritmo natural da razão de probabilidades (*odds ratio*, na expressão em inglês). A razão de probabilidades tem uma interpretação sugestiva: se as probabilidades dos acontecimentos Y = 1 e Y = 0 fossem, respectivamente, 0,8 e 0,2, dir-se-ia que "as chances são de 4 para 1" em favor da verificação do acontecimento a que corresponde Y = 1.

Se o logaritmo da razão de probabilidades fosse observado (excepto por uma perturbação aleatória), a equação em [17] poderia servir de base a uma regressão linear com as características habituais. Veja-se que esse logaritmo é uma variável contínua que assume valores de $-\infty$ a $+\infty$, à medida que a probabilidade do acontecimento varia de 0 a 1. A simplicidade da equação [17] contrasta com

$$\text{Prob}(Y_i = 1) = \frac{1}{1+e^{-(\beta_1+\beta_2 X_{2i}+...+\beta_k X_{ki})}}, \qquad [18]$$

em que é patente a não linearidade da variável do primeiro membro relativamente aos coeficientes $\beta_1$, $\beta_2$, ..., $\beta_k$. Se o objectivo for a estimação das probabilidades em [18], uma estratégia a considerar seria a de basear numa equação como [17] a estimação dos coeficientes e, depois, usar essas estimativas na expressão [18].

Estimação de modelos lineares de probabilidade

Admita-se disponível uma amostra de observações $(X_{2i}, X_{3i}, ..., X_{ki})$, i = 1, 2, ..., n, com a informação, para cada uma das observações, sobre a verificação do acontecimento em apreço, codificada pela atribuição dos valores 1 ou 0 a $Y_i$. O modelo pode ser escrito na forma habitual,

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{u},$$

em que **X** é a matriz (n×k) cuja iª linha é o vector $\mathbf{X_i}$ que se definiu acima,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \cdots \\ \mathbf{X_n} \end{bmatrix},$$

e **Y** é um vector (n×1) de componentes $Y_1, Y_2, ..., Y_n$, iguais, no caso, ou a 0, ou a 1.

Desde que **X** tenha característica igual ao número de colunas, k, é possível calcular estimativas de $\boldsymbol{\beta}$ por OLS segundo $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\,\mathbf{X'Y}$ e, a partir delas, estimativas de $\text{Prob}(Y_i = 1)$ de acordo com

$$\text{Pr}\,\widehat{\text{ob}(Y_i} = 1) = \mathbf{X_i}\,\hat{\boldsymbol{\beta}}.$$

Como já se afirmou, não está garantido que as estimativas da probabilidade assim obtidas pertençam, para todo i, ao intervalo [0; 1].

As perturbações aleatórias em **u** são heteroscedásticas e, por razões já conhecidas, o estimador OLS de $\boldsymbol{\beta}$ não é eficiente. Sabe-se (de [5]) que **Var(u)** será uma matriz diagonal, em que o iº elemento da diagonal principal é

$$\text{Var}(u_i) = p_i\,(1-p_i) = \mathbf{X_i}\,\boldsymbol{\beta}\,(1 - \mathbf{X_i}\,\boldsymbol{\beta}). \qquad [19]$$

Uma vez que $\text{Var}(u_i)$ depende do vector $\boldsymbol{\beta}$, desconhecido, o estimador generalizado de mínimos quadrados (GLS), que seria o estimador linear e cêntrico de variância mínima, não é exequível. O estimador exequível (EGLS) é

$$\hat{\boldsymbol{\beta}}_{\text{EGLS}} = \{\mathbf{X'}\,[\,\widehat{\mathbf{Var}}(\mathbf{u})\,]^{-1}\,\mathbf{X}\}^{-1}\,\mathbf{X'}\,[\,\widehat{\mathbf{Var}}(\mathbf{u})\,]^{-1}\,\mathbf{Y}, \qquad [20]$$

em que, por $\widehat{\mathbf{Var}}(\mathbf{u})$, se designou um estimador da matriz diagonal **Var(u)**.

Esse estimador pode ser construído substituindo $\boldsymbol{\beta}$ em [19] por $\hat{\boldsymbol{\beta}}_{\text{OLS}}$. O procedimento equivaleria a, inicialmente, estimar $\boldsymbol{\beta}$ por OLS, utilizar as estimativas dos coeficientes para formar estimativas das variâncias em [19] e, por último, a reestimar $\boldsymbol{\beta}$ por EGLS. Embora não cêntrico em amostras finitas, o estimador assim definido é consistente. Contudo, para as observações em que seja $\mathbf{X_i}\,\hat{\boldsymbol{\beta}} \leq 0$ ou $\mathbf{X_i}\,\hat{\boldsymbol{\beta}} \geq 1$, a estimativa da variância segundo [19] conduzirá a valores negativos ou nulos. Na literatura, dois procedimentos *ad hoc* que têm sido usados para lidar com o problema consistem na exclusão das observações em que tal suceda, ou na substituição de $\mathbf{X_i}\,\hat{\boldsymbol{\beta}}$ por um número próximo de 0 (v.g., 0,01), quando $\mathbf{X_i}\,\hat{\boldsymbol{\beta}} \leq 0$, ou por um número próximo de 1 (v.g., 0,99), se $\mathbf{X_i}\,\hat{\boldsymbol{\beta}} \geq 1$.

Em amostras de configuração peculiar, é possível recorrer a uma variante, proposta por Goldberger, do procedimento de estimação que se acaba de descrever. A configuração em causa requer que, para cada sequência ordenada ($X_{2i}$, $X_{3i}$, ..., $X_{ki}$) haja multiplicidade de observações. Então, é possível calcular, para cada grupo de observações que apresentam exactamente os mesmos valores para todas as variáveis explicativas, a proporção das observações em que Y = 1, e utilizá-la como estimativa da probabilidade (condicional a $X_2 = X_{2i}$, $X_3 = X_{3i}$, ..., $X_k = X_{ki}$) de verificação do acontecimento em causa. Essa proporção, que se designará por $\hat{p}_i$, poderá assumir valores diferentes de 0 e 1. Pelo contrário, numa amostra em que não houvesse nenhum par de observações "repetidas" (isto é, com os mesmos valores para as variáveis explicativas), todas as proporções assim calculadas teriam de ser ou 0, ou 1.

Uma ilustração do tipo de amostra referido ocorre em estudos sobre a escolha de modos de transporte. Suponha-se que, para cada indivíduo, i, e para cada um de $n_i$ dias úteis de certo mês, é registado o modo de transporte empregue, público ou privado. As variáveis que determinam a escolha (rendimento do indivíduo, distância da residência ao trabalho, etc.) assumem valores imutáveis para as $n_i$ observações relativas a uma pessoa inquirida, mas o modo de transporte utilizado pode não ser sempre o mesmo e verificar-se que houve recurso aos transportes públicos em, por exemplo, 30% das deslocações efectuadas.

O expediente que esta configuração da amostra viabiliza é o da substituição da série de 0's e 1's de valores de Y na amostra original por uma série de proporções $\hat{p}$ no intervalo real [0, 1]. Se as $n_i$ observações relativas ao iº indivíduo no dia t (t = 1, 2, ..., $n_i$) obedecem a

$$Y_{it} = \beta_1 + \beta_2\, X_{2it} + \beta_3\, X_{3it} + ... + \beta_k\, X_{kit} + u_{it},$$

em que $Y_{it} = 1$ ou $Y_{it} = 0$, $X_{jit} = X_{jit'}$, para quaisquer t, t' = 1, 2, ..., $n_i$ e j = 2, 3, ..., k, e $u_{it}$ é uma perturbação aleatória de média nula e variância $p_i(1-p_i)$, somando essas observações e dividindo por $n_i$, obtém-se

$$\frac{\sum_{t=1}^{n_i} Y_{it}}{n_i} = \beta_1 + \beta_2\, X_{2i} + \beta_3\, X_{3i} + ... + \beta_k\, X_{ki} + \frac{\sum_{t=1}^{n_i} u_{it}}{n_i}.$$

Para $n_i$ finito, a variável no primeiro membro desta equação é a proporção $\hat{p}_i$ e a perturbação no segundo membro, que se designará por $\overline{u_i}$, tem valor esperado nulo e variância dada por $p_i(1-p_i)/n_i$.

Suponha-se que se formaram, a partir das observações originais e de acordo com a regra explicada acima, n observações agrupadas (uma por cada indivíduo inquirido, na ilustração) e que se calculou $\hat{p}_i$ para cada um desses grupos. Seja $\hat{\mathbf{p}}$ o vector-coluna em que se reuniram essas n proporções e $\overline{\mathbf{u}}$ o vector-coluna das perturbações correspondentes. Então, é

$$\hat{\mathbf{p}} = \mathbf{X}\,\boldsymbol{\beta} + \overline{\mathbf{u}}. \qquad [21]$$

O estimador OLS de $\boldsymbol{\beta}$ seria dado por

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X'}\,\mathbf{X})^{-1}\,\mathbf{X'}\;\hat{\mathbf{p}}; \qquad [22]$$

como se sabe, conquanto cêntrico e consistente, esse estimador não é eficiente. Uma alternativa seria o estimador GLS de $\boldsymbol{\beta}$ dado por

$$\hat{\boldsymbol{\beta}}_{GLS} = \{\mathbf{X'}\,[\mathbf{Var}(\overline{\mathbf{u}})]^{-1}\,\mathbf{X}\}^{-1}\,\mathbf{X'}\,[\mathbf{Var}(\overline{\mathbf{u}})]^{-1}\;\hat{\mathbf{p}}, \qquad [23]$$

com $\mathbf{Var}(\overline{\mathbf{u}})$ a designar uma matriz diagonal em que o elemento genérico da diagonal principal é $p_i(1-p_i)/n_i$,

$$\mathbf{Var}(\overline{\mathbf{u}}) = \text{diag}[p_1(1-p_1)/n_1,\ p_2(1-p_2)/n_2,\ ...,\ p_n(1-p_n)/n_n]\ .$$

Uma vez que as probabilidades $p_i$ não são observadas, o estimador GLS exequível a empregar seria

$$\hat{\boldsymbol{\beta}}_{EGLS} = \{\mathbf{X'}\,[\,\widehat{\mathbf{Var}}(\overline{\mathbf{u}})\,]^{-1}\,\mathbf{X}\}^{-1}\,\mathbf{X'}\,[\,\widehat{\mathbf{Var}}(\overline{\mathbf{u}})\,]^{-1}\;\hat{\mathbf{p}}, \qquad [24]$$

em que $\widehat{\mathbf{Var}}(\overline{\mathbf{u}})$ designa um estimador de $\mathbf{Var}(\overline{\mathbf{u}})$. Oferecem-se duas vias para a construção desse estimador: corresponde uma à utilização directa das proporções amostrais,

$$\widehat{\mathbf{Var}}(\overline{\mathbf{u}}) = \text{diag}[\,\hat{p}_1(1-\hat{p}_1)/n_1,\ \hat{p}_2(1-\hat{p}_2)/n_2,\ ...,\ \hat{p}_n(1-\hat{p}_n)/n_n]; \qquad [25]$$

parece preferível uma segunda via em que essas proporções seriam substituídas pelas estimativas resultantes da estimação da equação [21] por OLS,

$$\hat{\hat{\mathbf{p}}} = \mathbf{X}\;\hat{\boldsymbol{\beta}}_{OLS},$$

e

$$\widehat{\mathbf{Var}}(\overline{\mathbf{u}}) = \text{diag}[\,\hat{\hat{p}}_1(1-\hat{\hat{p}}_1)/n_1,\ \hat{\hat{p}}_2(1-\hat{\hat{p}}_2)/n_2,\ ...,\ \hat{\hat{p}}_n(1-\hat{\hat{p}}_n)/n_n]. \qquad [26]$$

Os estimadores definidos por [24] e [26] serão, geralmente, consistentes e assimptoticamente eficientes.

Embora o procedimento que se acaba de descrever para amostras com observações repetidas permita, ao substituir uma variável dependente dicotómica por outra, contínua no intervalo [0, 1], minorar o problema da possível obtenção de estimativas de probabilidades fora desse intervalo, não o resolve por completo. Quando isso acontece, não é sequer possível construir a estimativa de $\mathbf{Var}(\bar{\mathbf{u}})$ segundo [26], devido à presença de estimativas (de variâncias) negativas. Se é certo que com a fórmula em [25] tal não sucede, uma vez que se trata de proporções amostrais, não é raro que entre essas proporções se incluam algumas que são iguais a 0 ou a 1; nesse caso, haverá elementos nulos na diagonal principal da matriz da equação [25], essa matriz não será invertível e, por consequência, as estimativas GLS definidas em [24] não são determinadas.

Estimação de modelos *probit* e *logit*

Seja o modelo

$$Y_i^* = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i = \mathbf{X_i}\, \boldsymbol{\beta} + u_i,$$

em que $\boldsymbol{\beta}$ é um vector de parâmetros a estimar. Apesar da linearidade de $Y_i^*$ relativamente a $\boldsymbol{\beta}$, a equação não pode servir de base à estimação, porquanto a variável dependente não é observável. Observável é uma variável dicotómica Y tal que

$$Y_i = \begin{cases} 1, \text{ com probabilidade } F(\mathbf{X}_i\boldsymbol{\beta}) \\ 0, \text{ com probabilidade } 1 - F(\mathbf{X}_i\boldsymbol{\beta}) \end{cases},$$

conforme se viu atrás (equações [13]). Seja F(.) uma ou outra das funções de distribuição que se consideraram atrás, $F(\mathbf{X_i}\, \boldsymbol{\beta})$ não é função linear de $\boldsymbol{\beta}$ e a estimação pelos métodos habituais no contexto do modelo de regressão linear não é possível.

A estimação de modelos *probit* ou *logit* é feita, geralmente, pelo método da máxima verosimilhança. Admita-se disponível uma amostra aleatória em que há $n_0$ observações com $Y_i = 0$ e $n_1$ ($= n - n_0$) observações em que $Y_i = 1$. A função de verosimilhança é dada por

$$L = \Pi_0 \, Prob(Y_i = 0) \times \Pi_1 \, Prob(Y_i = 1),$$

em que $\Pi_0$ pretende significar o produto iterado das $n_0$ observações para as quais $Y_i = 0$, enquanto $\Pi_1$ corresponde ao produto iterado das $n_1$ observações restantes. É possível dar à função de verosimilhança uma expressão com notação mais simples, fazendo uso do facto de que Y apenas assume os valores 0 e 1:

$$L = \prod_{i=1}^{n} \left\{ \left[ \text{Prob}(Y_i = 0) \right]^{1-Y_i} \left[ \text{Prob}(Y_i = 1) \right]^{Y_i} \right\}$$

ou, substituindo pelas expressões apropriadas,

$$L = \prod_{i=1}^{n} \left\{ \left[ 1 - F(\mathbf{X_i}\, \boldsymbol{\beta}) \right]^{1-Y_i} \left[ F(\mathbf{X_i}\, \boldsymbol{\beta}) \right]^{Y_i} \right\}. \qquad [27]^{2}$$

Os estimadores de máxima verosimilhança do vector $\boldsymbol{\beta}$ são as funções de valores amostrais que maximizam L em ordem a $\boldsymbol{\beta}$. Como se sabe, é geralmente mais fácil determinar o máximo da função logarítmica de verosimilhança,

$$\ln L = \sum_{i=1}^{n} \left\{ (1 - Y_i) \ln\left[ 1 - F(\mathbf{X_i}\, \boldsymbol{\beta}) \right] + Y_i \ln\left[ F(\mathbf{X_i}\, \boldsymbol{\beta}) \right] \right\}. \qquad [28]$$

Os estimadores de máxima verosimilhança dos parâmetros do modelo (*probit* ou *logit*) são os que maximizam a função L e, por implicação, também a função em [28].

A condição usual para resolução do problema requer o anulamento das k derivadas parciais de primeira ordem, ou seja, requer

$$\frac{\partial \ln L}{\partial \beta_j} = 0, \quad \forall j = 1, 2, ..., k. \qquad [29]$$

A condição de 2ª ordem, por sua vez, será preenchida se, na vizinhança do óptimo, for negativa definida a matriz das derivadas parciais de 2ª ordem de ln L em ordem a $\boldsymbol{\beta}$.

O sistema de k equações em [29] não é linear em $\beta_1$, $\beta_2$ , ..., $\beta_k$, e não é possível, por isso, definir por uma expressão analítica a solução do problema, seja F(.) a função de distribuição normal dos modelos *probit* ou a função logística dos modelos *logit*. Note-se, contudo, que a primeira classe de modelos é, do ponto de vista calculatório, bastante mais difícil de tratar, por envolver a avaliação de n integrais. Esse é um dos factores que, antes do advento de meios de cálculo automático poderosos, tornava popular o recurso à especificação *logit* em detrimento do *probit*. Hoje em dia, sabe-se que é côncava a função de verosimilhança para ambas as classes de modelos e que, por consequência, o máximo da função é único (se existir um máximo); e há já *software* adequado para determinação desse máximo sem grande

---

[2] Embora se tenha usado, nas expressões acima, o símbolo L, simplesmente, para facilitar a notação, deve ter-se presente que L se refere a uma função dos parâmetros do modelo, dados os valores amostrais das variáveis. Entre os parâmetros do modelo incluem-se, de modo óbvio, os coeficientes das variáveis explicativas que são as componentes de $\boldsymbol{\beta}$, e, eventualmente, também os parâmetros da matriz de variâncias e covariâncias das perturbações. Entenda-se, portanto, que

$$L \equiv L(\boldsymbol{\beta} \mid \mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}, Y_1, Y_2, ..., Y_n).$$

custo. Não se entrará aqui, contudo, nos pormenores dos métodos de optimização numérica que permitem a identificação do máximo.

É instrutivo, no entanto, e útil para um desenvolvimento a apresentar adiante, analisar neste contexto a formulação mais simples possível do problema, aquela em que o vector $\boldsymbol{\beta}$ tem uma única componente, $\beta_1$. Ter-se-á, nesse caso,

$$\ln L = \sum_{i=1}^{n} \left\{ (1-Y_i)\ \ln\left[1\ -\ F(\beta_1)\right] + Y_i\ \ln\left[F(\beta_1)\right] \right\}\ ,$$

e, pela condição de 1ª ordem,

$$\frac{d \ln L}{d\,\beta_1} = 0 \ \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{n} \left\{ (1-Y_i)\ \frac{-\ f(\beta_1)}{1\ -\ F(\beta_1)} + Y_i\ \frac{f(\beta_1)}{F(\beta_1)} \right\} = 0$$

(recorde-se que $\dfrac{d\,F(\beta_1)}{d\,\beta_1} = f(\beta_1)$, por definição de função de densidade)

$$\Rightarrow \frac{f(\beta_1)}{F(\beta_1)} \sum_{i=1}^{n} Y_i\ - \frac{f(\beta_1)}{1\ -\ F(\beta_1)} \sum_{i=1}^{n} (1-Y_i) = 0$$

$$\Rightarrow n_1\ \frac{f(\beta_1)}{F(\beta_1)} = (n-n_1)\ \frac{f(\beta_1)}{1\ -\ F(\beta_1)}$$

(usando a convenção introduzida anteriormente de simbolizar por $n_1$ o número de observações com $Y_i = 1$ e por $n_0$ ($n_0 = n - n_1$) o número de observações com $Y_i = 0$)

$$\Rightarrow F(\beta_1) = \frac{n_1}{n}\ . \qquad [30]$$

O valor máximo da função logarítmica de verosimilhança, atingido quando $\beta_1$ satisfizer a condição em [30], será

$$\ln L_0 = \sum_{i=1}^{n} \left\{ (1-Y_i)\ \ln\left[1\ -\ \frac{n_1}{n}\right] + Y_i\ \ln\left[\frac{n_1}{n}\right] \right\};$$

após simplificação,

$$\ln L_0 = n_0\ \ln\frac{n_0}{n} + n_1\ \ln\frac{n_1}{n}\ . \qquad [31]$$

Note-se que os resultados em [30] e [31] foram obtidos sem que alguma vez se particularizasse se F(.) se referia a uma distribuição normal reduzida ou à função logística, pelo que são válidos para ambos os casos.

A interpretação de [30] é, no entanto, diferente nos dois casos. Num modelo *probit*, o estimador de máxima verosimilhança de $\beta_1$ é $\hat{\beta}_1$ tal que

$$\Phi(\hat{\beta}_1) = \int_{-\infty}^{\hat{\beta}_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = \frac{n_1}{n}; \qquad [32]$$

num modelo *logit*, o estimador de máxima verosimilhança de $\beta_1$ é $\hat{\beta}_1$ tal que

$$\Lambda(\hat{\beta}_1) = \frac{1}{1+e^{-\hat{\beta}_1}} = \frac{n_1}{n},$$

donde resulta

$$\hat{\beta}_1 = \ln \frac{n_1}{n_0}. \qquad [33]$$

Por exemplo, se for de 50% a percentagem de observações com Y = 1 na amostra, a estimativa de máxima verosimilhança de $\beta_1$ será igual a 0, quer se trate de um modelo *probit* ou de um *logit*.

Prova-se que, sob certas condições, os estimadores de máxima verosimilhança são consistentes e têm distribuição assimptótica normal, que pode ser aproximada, em amostras finitas, por uma distribuição normal de média $\boldsymbol{\beta}$ e matriz de variâncias e covariâncias

$$\widehat{\mathbf{Var}}(\hat{\boldsymbol{\beta}}_{ML}) = -\left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right)^{-1}, \qquad [34]$$

em que a matriz das derivadas parciais de 2ª ordem indicada na expressão deve ser entendida como avaliada para $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{ML}$.[3]

Embora, com o *software* modernamente disponível, a estimação de modelos *probit* e *logit* pelo método da máxima verosimilhança não suscite dificuldades de maior, é possível, apenas para o *logit* e para amostras com observações repetidas como se descreveu atrás, recorrer também à estimação por métodos de mínimos quadrados. Recordando (ver equação [17]) que

[3] Há outras aproximações possíveis à matriz $\mathbf{Var}(\hat{\boldsymbol{\beta}}_{ML})$; a escolha por uma ou outra aproximação depende, geralmente, do algoritmo usado na busca do máximo da função, questão que não será discutida neste texto.

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \mathbf{X_i}\,\boldsymbol{\beta} = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki},$$

ocorre imediatamente que poderia obter-se estimativas OLS dos coeficientes, por ajustamento de

$$\ln\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \text{erro},$$

desde que nenhuma das proporções amostrais $\hat{p}_i$ fosse igual a 0 ou a 1. Em amostras sem observações repetidas, o uso desse expediente está, evidentemente, excluído.

Avaliação de resultados e análise estatística

As classes de modelos LPM, *probit* e *logit* têm em comum o facto de, neles, a variável dependente ser uma variável qualitativa com dois estados possíveis, a que, habitualmente, se faz corresponder o símbolo 1 ou o símbolo 0. Dispondo-se de uma amostra aleatória de observações ($X_{2i}$, $X_{3i}$, ..., $X_{ki}$), i = 1, 2, ..., n, e ainda, para cada observação, de informação sobre a qual dos dois grupos possíveis pertence, seria concebível proceder-se à estimação, com os mesmos dados, de um LPM (por OLS), de um *probit*, ou de um *logit* (por máxima verosimilhança). Que comparação se pode estabelecer entre os resultados segundo as três especificações?

No que toca a estimativas dos coeficientes, as comparações feitas tendem a sugerir a validade aproximada das relações seguintes, para as estimativas obtidas num LPM, $\hat{\beta}_{j,LPM}$, e num modelo *logit*, $\hat{\beta}_{j,L}$:

$$\hat{\beta}_{1,LPM} \approx 0{,}25\ \hat{\beta}_{1,L} + 0{,}5,$$

para o termo independente, e

$$\hat{\beta}_{j,LPM} \approx 0{,}25\ \hat{\beta}_{j,L} \quad , \quad j = 2, 3, ..., k,$$

para os coeficientes das variáveis independentes.

Por sua vez, as estimativas obtidas por *probit*, $\hat{\beta}_{j,P}$, e por *logit*, $\hat{\beta}_{j,L}$, tendem a verificar, aproximadamente,

$$\hat{\beta}_{j,L} \approx 1{,}6\ \hat{\beta}_{j,P} \quad , \quad j = 2, 3, ..., k,$$

em que o factor 1,6 está associado com a desigualdade da variância das perturbações nos modelos *logit* ($\pi^2/3$) e *probit* (1).

Um segundo elemento de comparação dos resultados é o que respeita às estimativas das probabilidades. Quanto a esse aspecto, os modelos LPM, de um lado, e os modelos *probit* e *logit*, do outro, podem produzir resultados radicalmente diferentes. Como já se apontou, pode acontecer, com os primeiros, que as estimativas se situem fora do intervalo [0, 1]; com os outros, tal nunca sucede. De resto, tem-se constatado frequentemente que as estimativas de probabilidades obtidas pela formulação *probit* e pela formulação *logit* são muito semelhantes, o que não surpreende se se atender à similitude de comportamento das duas funções de distribuição. Essa semelhança é, geralmente, mais acentuada quando, na amostra, são próximas as proporções de ocorrência dos dois valores possíveis, Y = 1 e Y = 0.

Para a iª observação, a probabilidade de ser $Y_i = 1$ num LPM é estimada segundo

$$\hat{Y}_i = \Pr ob(\hat{Y}_i = 1) = \mathbf{X_i}\ \hat{\boldsymbol{\beta}}, \qquad [35]$$

em que se designou por $\hat{\boldsymbol{\beta}}$ o estimador (OLS ou EGLS) utilizado e se admitiu ser $0 \leq. \mathbf{X_i}\ \hat{\boldsymbol{\beta}} \leq 1$.

Num modelo *probit*, essa probabilidade seria estimada de acordo com

$$\Pr ob(\hat{Y}_i = 1) = \Phi(\mathbf{X_i}\ \hat{\boldsymbol{\beta}}) = \int_{-\infty}^{\mathbf{X_i}\hat{\boldsymbol{\beta}}} \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}t^2} dt \qquad [36]$$

e, num modelo *logit*, por

$$\Pr ob(\hat{Y}_i = 1) = \Lambda(\mathbf{X_i}\ \hat{\boldsymbol{\beta}}) = \frac{1}{1 + e^{-\mathbf{X_i}\hat{\boldsymbol{\beta}}}}. \qquad [37]$$

Quer em [36], quer em [37], designou-se, agora, por $\hat{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança do vector de coeficientes.

Como se viu em secção anterior, ao contrário do que se passa nos modelos LPM, não é imediata a interpretação das estimativas dos coeficientes dos modelos *probit* e *logit*. Nestes, a única informação óbvia que se deriva das estimativas é a do sinal, positivo ou negativo, da influência marginal do regressor em causa sobre a probabilidade de verificação do acontecimento. É usual, por isso, que, a par dessas estimativas, seja fornecida informação sobre a grandeza dessa influência; e, uma vez que ela é variável de observação para observação e, também, com os valores das variáveis explicativas, a prática corrente é a de calcular estimativas desses efeitos referidas às médias amostrais dos regressores. Quando se contarem variáveis *dummy* entre estes regressores, a abordagem mais correcta será a de confrontar as estimativas de probabilidades calculadas para o valor 1 e para o valor 0 dessa variável explicativa binária, para um mesmo vector de valores das outras variáveis explicativas.

Poderá ter algum interesse dispor, para os modelos *probit* e *logit*, de um indicador sintético da qualidade do "ajustamento", similar ao coeficiente de determinação $R^2$ da regressão clássica. Um indicador referido por vezes é o chamado pseudo-$R^2$, definido por

$$ps.R^2 = 1 - \frac{\ln L(\hat{\beta}_{ML})}{\ln L_0}, \qquad [38]$$

em que $\ln L(\hat{\beta}_{ML})$ é o valor da função logarítmica de verosimilhança avaliada para $\hat{\beta}_{ML}$ (e, portanto, igual ao máximo dessa função), enquanto $\ln L_0$ designa o máximo dessa função, sob a restrição de serem nulos todos os coeficientes das variáveis independentes ($\beta_2 = \beta_3 = ... = \beta_k = 0$). Este máximo foi já calculado anteriormente (ver equação [31]).

À semelhança de $R^2$, o pseudo-$R^2$ assume apenas valores no intervalo [0, 1]; diferentemente de $R^2$, o pseudo-$R^2$ não tem interpretação intuitiva imediata. Ter-se-ia $ps.R^2 = 1$ quando $\ln L(\hat{\beta}_{ML}) = 0$ e, portanto, $L(\hat{\beta}_{ML}) = 1$. Recorde-se que a função de verosimilhança é um produto de n factores, cada um dos quais dado por

$$\left[\text{Prob}(Y_i = 0)\right]^{1-Y_i} \left[\text{Prob}(Y_i = 1)\right]^{Y_i};$$

logo, são todos não negativos e não superiores a 1 (por corresponderem a probabilidades). Então, o produto só seria igual a 1 se esses n factores fossem todos iguais a 1, o que exigiria que, sendo $Y_i = 1$, fosse igual a 1 a probabilidade estimada de ser $Y_i = 1$; e, sendo $Y_i = 0$, fosse também igual a 1 a estimativa da probabilidade de ser $Y_i = 0$ (ou, por outras palavras, que fosse 0 a estimativa da probabilidade de ser $Y_i = 1$). O modelo estimado seria então um "previsor perfeito", no sentido de predizer correctamente, para todas as observações sem excepção, a qual dos dois grupos pertencia. Tenha-se em atenção, contudo, que, seja F(.) a função de distribuição normal reduzida ou a função logística, apenas se aproximaria de 1 (ou 0) se o argumento tendesse para infinito (ou $-\infty$). Um valor de $ps.R^2$ virtualmente igual a 1 seria mais sugestivo de uma má especificação do que de um "perfeito ajustamento".

No outro extremo, ter-se-ia $ps.R^2 = 0$ quando $\ln L(\hat{\beta}_{ML}) = \ln L_0$ e, por conseguinte, quando fossem nulas todas as componentes do vector $\hat{\beta}_{ML}$, excepto a primeira (relativa ao termo independente). Nesse caso, nenhuma das variáveis "explicativas" afectaria a probabilidade de ocorrência do acontecimento em análise. Como se viu na equação [30], a estimativa dessa probabilidade seria, simplesmente, igual à proporção de observações na amostra com Y = 1.

Outro elemento de apreciação dos resultados de estimação que, por vezes, se usa é uma matriz em que se confrontam as predições do modelo com a repartição efectiva das observações da amostra pelos dois grupos em presença. Calculada, para cada observação, uma estimativa da probabilidade de Y = 1 (usando fórmulas já apresentadas e tendo em conta se se trata de um modelo *probit* ou de um *logit*), é costume predizer-se $Y_i = 1$ ou $Y_i = 0$ conforme essa estimativa exceda ou não 0,5. Depois, a informação é organizada numa tabela da forma:

| | $Y_i$ predito = 1 | $Y_i$ predito = 0 | Totais |
|---|---|---|---|
| $Y_i$ observado = 1 | A | B | $n_1$ |
| $Y_i$ observado = 0 | C | D | $n_0$ |
| Totais | E | F | n |

Os números representados por A e D correspondem a previsões acertadas do modelo, enquanto B e C quantificam previsões erradas. A proporção de observações erradamente classificadas, (B+C)/n, é a chamada taxa aparente de erro.

A esse respeito, refira-se que, como métodos de classificação ou de previsão, as análises *probit* e *logit* têm um competidor sério em previsores *ad hoc*, como aquele que prevê, para toda e qualquer observação, o acontecimento com maior frequência relativa na amostra. Por exemplo, se 90% das observações numa amostra apresentam Y = 1, o previsor "$\hat{Y}_i = 1, \forall i$" fará predições correctas em 90% das vezes, o que não deixa grande escopo para melhoria a métodos mais complexos. Não é raro que, em termos de capacidade preditiva, os métodos *probit* e *logit* apenas consigam uma melhoria modesta sobre a desse previsor *ad hoc*, ainda quando consigam explicar razoavelmente os determinantes da probabilidade do acontecimento em estudo. Essa é uma das razões por que, apesar do seu carácter atraente, a valia da informação na tabela acima é, frequentemente, questionada.

A análise estatística dos modelos *probit* e *logit* é, geralmente, conduzida em termos análogos aos que se conhecem para o modelo de regressão linear, apenas com as modificações requeridas pelo carácter de aproximação assimptótica da matriz de variâncias estimadas em [34]. Assim, testes de hipóteses sobre coeficientes individuais podem ser conduzidos da forma usual. Embora se continue a designar, por abuso de linguagem, o quociente da estimativa do coeficiente pela estimativa do desvio-padrão do estimador por rácio t, os valores críticos aproximados devem obter-se nas tabelas da distribuição normal reduzida.

Testes de hipóteses sobre restrições aos parâmetros envolvem uma comparação entre os máximos da função logarítmica de verosimilhança atingidos quando as restrições são observadas e quando são ignoradas, à semelhança do que na regressão linear clássica se faz com as somas de quadrados dos resíduos. A estatística de teste (chamada razão de verosimilhança) é calculada como

$$-2\,(\ln L_r - \ln L_u), \qquad [39]$$

em que $\ln L_r$ designa o valor da função logarítmica de verosimilhança quando maximizada sob as restrições fixadas na hipótese nula sob teste e $\ln L_u$ o valor da função logarítmica de verosimilhança sem restrições (para que se usou atrás o símbolo $\ln L(\hat{\beta}_{ML})$). O valor crítico aproximado é obtido nas tabelas da distribuição do qui-quadrado, com número de graus de liberdade igual ao de restrições consideradas na hipótese nula.

Um caso particular de uso da razão de verosimilhança que tem particular interesse em modelos estimados pelo método da máxima verosimilhança ocorre no teste da hipótese nula $\beta_2 = \beta_3 = ... = \beta_k = 0$, que tem alcance análogo ao do teste de significância global na regressão clássica. Sob essa hipótese, tem-se, assimptoticamente,

$$-2\,[\ln L_0 - \ln L(\hat{\beta}_{ML})] \;\sim\; \chi^2_{(k-1)}, \qquad [40]$$

em que o significado dos símbolos é o exposto a propósito do pseudo-$R^2$. É frequente, no relato dos resultados de estimação destes modelos, dar-se conta do valor amostral da estatística em [40] como indicador sumário da qualidade do "ajustamento", em papel semelhante ao que, na regressão clássica, se conferiria à estatística F para teste da mesma hipótese nula.

Refira-se, a terminar, que, ao contrário do que se passa no modelo clássico de regressão linear, ou em mais alto grau do que nesse, se tem constatado, em modelos como o *probit* ou *logit*, forte sensibilidade dos resultados e propriedades dos estimadores a situações como a não normalidade da distribuição dos erros, heteroscedasticidade ou erros de especificação. Dado o papel crucial da hipótese da normalidade em algumas das formulações expostas, não é, talvez, surpreendente essa sensibilidade. Tal como no modelo linear clássico se enxertaram numerosíssimos desenvolvimentos e extensões, também nos modelos com variável qualitativa ou dependente há uma vasta literatura que prolonga as potencialidades dos modelos basilares ora estudados.

## Identificação dos parâmetros de modelos *probit* e *logit*

A possibilidade de identificar os parâmetros de uma população, a partir de uma amostra dela extraída, não é ilimitada. Esse é o facto subjacente a um problema complexo de Econometria, conhecido pela questão da identificação.

Afloramentos dessa questão surgem logo no modelo clássico de regressão linear. É conhecido, por exemplo, que com uma amostra de dimensão inferior ao número de parâmetros do modelo não é possível estimá-los ou, pelo menos, estimá-los a todos. Embora possibilidade de estimação e identificação não sejam sinónimos, os dois tipos de problema aparecem, frequentemente, associados.

Um outro exemplo, ainda no domínio da regressão clássica, ocorre a propósito do termo independente. Considerem-se os modelos $Y = \beta_1 + \beta_2 X + u$, em que u é uma perturbação aleatória de média nula, e $Y = \beta_1^* + \beta_2 X + v$, em que é $v \equiv u + 2$ e $\beta_1^* = \beta_1 - 2$. Por muito grande que seja a dimensão da amostra de observações dos pares (X, Y), não é possível distinguir se foram geradas por um ou outro dos dois modelos, o primeiro com termo constante $\beta_1$ e perturbações de média 0, o segundo com termo constante $\beta_1^*$ e perturbações de média 2. Se a perturbação aleatória de um modelo tiver valor esperado $\mu$, constante, mas não igual a 0, a soma $(\beta_1 + \mu)$ é identificável e pode ser estimada, mas nem $\beta_1$, nem $\mu$, isoladamente, o serão. Essa é, de resto, uma das razões por que, na generalidade das análises econométricas, não é conferido grande interesse às estimativas de termos constantes.

Para uma terceira ilustração, seja o modelo não linear $Y = \beta^2 X + u$. É óbvio que o parâmetro $\beta$ não pode ser identificado, já que qualquer amostra gerada com um valor de $\beta$ seria idêntica à gerada com o valor simétrico. No entanto, $\beta^2$ é identificável e, por essa razão, costuma dizer-se que $\beta$ é identificável, excepto pelo sinal.

Mais prementes nos modelos não lineares e nos modelos multiequacionais, os problemas de identificação colocam-se, também, naqueles em que a variável dependente é qualitativa ou limitada. Duas características comuns destes são:

1) uma relação estrutural que envolve uma variável dependente, $Y^*$, que não é observável ou, pelo menos, não é observável para todos os indivíduos da população;

2) uma regra que determina uma variável observada, Y, em função da variável latente, $Y^*$, ou, se $Y^*$ é observável para apenas alguns indivíduos da população, que discrimina entre esses e os restantes.

A restrição à observabilidade de $Y^*$ actua como um filtro que distorce a imagem da relação que se pretende estimar e, em alguns casos, a distorce a ponto de impedir a identificação de certos parâmetros. Por exemplo, num exercício de tiro em que a informação disponível se cinja a "acertou, ou falhou", não é possível identificar o parâmetro associado com a dispersão dos tiros em torno do alvo; já o seria, contudo, se se dispusesse de informação quanto à distância do ponto de impacto relativamente ao centro do alvo, para os projécteis que embateram numa vizinhança desse centro.

Para discutir a questão da identificação em modelos *probit* e *logit*, considere-se uma população A, caracterizada pelas relações

$$Y_i^{A*} = \beta_1 + \beta_2 X_i + u_i, \quad E(u_i) = 0, \quad Y_i = \begin{cases} 1, \text{ se } Y_i^{A*} > 0 \\ 0, \text{ se } Y_i^{A*} \leq 0 \end{cases}.$$

Em primeiro lugar, confronte-se essa população com outra, B, caracterizada por

$$Y_i^{B*} = \alpha_1 + \beta_2 X_i + v_i, \quad \alpha_1 = \beta_1 - \mu, \quad v_i \equiv u_i + \mu, \quad E(v_i) = \mu, \quad Y_i = \begin{cases} 1, \text{ se } Y_i^{B*} > 0 \\ 0, \text{ se } Y_i^{B*} \leq 0 \end{cases}.$$

Para todos os pares $(X_i, u_i)$, é $Y_i^{A*} = Y_i^{B*}$; por conseguinte, também a variável observada $Y_i$ será idêntica. Não é possível determinar, a partir de qualquer amostra de observações $(X_i, Y_i)$, se foi gerada pela população A ou pela população B e, portanto, se os parâmetros relevantes são $\beta_1, \beta_2, 0$, ou $\alpha_1, \beta_2, \mu$.

Em segundo lugar, confronte-se A com a população C, em que

$$Y_i^{C*} = \alpha_1 + \beta_2 X_i + u_i, \quad \alpha_1 = \beta_1 + \lambda, \quad E(u_i) = 0, \quad Y_i = \begin{cases} 1, \text{ se } Y_i^{C*} > \lambda \\ 0, \text{ se } Y_i^{C*} \leq \lambda \end{cases}.$$

Desta vez, são diferentes, para cada par ($X_i$, $u_i$), as variáveis latentes $Y_i^{C*}$ e $Y_i^{A*}$ ($Y_i^{C*} = Y_i^{A*} + \lambda$), mas a variável observada, $Y_i$, é a mesma, tornando impossível distinguir se, na população, o limiar de transição de Y = 0 para Y = 1 é 0 e o termo constante é $\beta_1$, ou se são $\lambda$ e $\alpha_1$, respectivamente.

As comparações efectuadas permitem concluir que não é possível identificar três constantes: o termo independente da relação estrutural, o valor médio das perturbações e o limiar de transição. É possível, no entanto, identificar uma constante que descreve, em condições que se precisarão abaixo, o efeito conjugado desses três factores. Para assegurar comparabilidade de resultados, a convenção de normalização usualmente adoptada é a de postular serem iguais a 0 as duas últimas constantes referidas e deixar, portanto, que seja a estimativa de $\beta_1$ a reflectir a influência conjunta.

Para uma terceira ilustração, comparem-se A e D, com

$$Y_i^{D*} = \alpha_1 + \alpha_2 X_i + v_i, \quad \alpha_1 = \sigma\beta_1, \quad \alpha_2 = \sigma\beta_2, \quad v_i = \sigma u_i, \quad Y_i = \begin{cases} 1, \text{ se } Y_i^{D*} > 0 \\ 0, \text{ se } Y_i^{D*} \leq 0 \end{cases},$$

com $\sigma > 0$. Para cada par ($X_i$, $u_i$), as variáveis latentes $Y_i^{D*}$ e $Y_i^{A*}$ são, novamente, diferentes (agora, $Y_i^{D*} = \sigma Y_i^{A*}$), mas é óbvio que, sempre que for $Y_i^{A*} > 0$, também será $Y_i^{D*} > 0$. Por conseguinte, a variável observada, $Y_i$, será a mesma, quer a população tenha por parâmetros $\beta_1$, $\beta_2$ e Var(u), ou $\sigma\beta_1$, $\sigma\beta_2$ e $\sigma^2$Var(u).

Do que acaba de expor-se neste último exemplo, devem reter-se duas consequências. A primeira é a de que a variância das perturbações em modelos *probit* ou *logit* não é identificável: há uma infinidade de populações, cada uma com o seu valor para $\sigma$, de que poderá ter sido extraída uma certa amostra de pares ($X_i$, $Y_i$). A segunda conclusão a tirar é a de que não é possível distinguir amostras de populações com coeficientes $\beta_1$, $\beta_2$ das geradas por populações com coeficientes $\sigma\beta_1$, $\sigma\beta_2$. Costuma referir-se esta situação dizendo que os coeficientes da relação são identificados, excepto por um factor de proporcionalidade constante.

Abordada a questão da identificação com recurso a alguns exemplos simples, passar-se-á, na sequência, a um tratamento mais formal do problema. Analisar-se-á, apenas, o caso dos modelos *probit*, mas é fácil a extensão dos resultados à especificação *logit*.

Mostrar-se-á, em primeiro lugar, que, em condições a explicitar adiante, o mesmo conjunto de observações pode ter sido gerado indistintamente por uma de várias estruturas, o que impossibilita a identificação de alguns dos parâmetros estruturais. Provar-se-á, em segundo lugar, que não é possível obter estimativas para todos os parâmetros requeridos por uma especificação supostamente mais geral.

Com esse objectivo, considerem-se as especificações

$$Z_i^* = \delta_1 + \delta_2 X_{2i} + \delta_3 X_{3i} + ... + \delta_k X_{ki} + v_i = \mathbf{X_i}\,\boldsymbol{\delta} + v_i \qquad [41]$$

$$Y_i = \begin{cases} 1, \text{ se } Z_i^* > 0 \\ 0, \text{ se } Z_i^* \leq 0 \end{cases} \qquad [42]$$

$$v_i \sim N(0,1) \qquad [43]$$

e

$$W_i^* = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + ... + \gamma_k X_{ki} + w_i = \mathbf{X_i}\,\boldsymbol{\gamma} + w_i \qquad [44]$$

$$Y_i = \begin{cases} 1, \text{ se } W_i^* > \lambda \\ 0, \text{ se } W_i^* \leq \lambda \end{cases} \qquad [45]$$

$$w_i \sim N(\mu,\sigma^2) \qquad [46]$$

São parâmetros do primeiro modelo as k componentes do vector $\boldsymbol{\delta}$, $\delta_1$, $\delta_2$, ..., $\delta_k$; no segundo, além dos k coeficientes $\gamma_1$, $\gamma_2$, ..., $\gamma_k$, incluem-se $\lambda$, $\mu$ e $\sigma^2$ como parâmetros adicionais.

As observações das populações em causa são, genericamente, do tipo ($X_{2i}$, $X_{3i}$, ..., $X_{ki}$, $Y_i$), em que $Y_i$ é um mero indicador, codificado com os valores 0 e 1, da pertença da iª observação a um de dois grupos possíveis. Buscar-se-á, na sequência, resposta para duas questões. Podem as observações ser geradas, indistintamente, por qualquer das estruturas em confronto, apesar da disparidade de especificações quanto a aspectos como os parâmetros da distribuição das perturbações aleatórias, ou como a definição da variável observada a partir da variável latente? Em caso afirmativo, em que condições?

Para verificar que a resposta à primeira questão é afirmativa, e que os dois modelos são equivalentes, sob o ponto de vista de poderem gerar amostras perfeitamente coincidentes, analisem-se as consequências das três acções seguintes:

1) Some-se e subtraia-se $\mu$ ao segundo membro da equação [44]; obter-se-á

$$W_i^* = (\gamma_1 + \mu) + \gamma_2 X_{2i} + \gamma_3 X_{3i} + ... + \gamma_k X_{ki} + (w_i - \mu)$$

e, evidentemente, não se altera com isso o valor de $W_i^*$, nem a regra em [45] conduz a observações de Y diferentes das anteriores.

2) Subtraia-se $\lambda$ a ambos os membros da equação anterior; vem

$$W_i^* - \lambda = (\gamma_1 + \mu - \lambda) + \gamma_2 X_{2i} + \gamma_3 X_{3i} + ... + \gamma_k X_{ki} + (w_i - \mu).$$

Se se modificar [45] para

$$Y_i = \begin{cases} 1, \text{ se } W_i^* - \lambda > \lambda - \lambda \\ 0, \text{ se } W_i^* - \lambda \leq \lambda - \lambda \end{cases},$$

ou seja, se

$$Y_i = \begin{cases} 1, \text{ se } W_i^* - \lambda > 0 \\ 0, \text{ se } W_i^* - \lambda \leq 0 \end{cases},$$

as observações de $Y_i$ serão as mesmas que anteriormente.

3) Por último, dividam-se ambos os membros da equação acima por $\sigma > 0$; virá

$$\frac{W_i^* - \lambda}{\sigma} = \frac{\gamma_1 + \mu - \lambda}{\sigma} + \frac{\gamma_2}{\sigma} X_{2i} + \frac{\gamma_3}{\sigma} X_{3i} + ... + \frac{\gamma_k}{\sigma} X_{ki} + \frac{w_i - \mu}{\sigma}. \qquad [47]$$

É imediato que a divisão de $(W_i^* - \lambda)$ por uma constante positiva não trará modificação alguma aos valores observados para Y, que continuarão a ser dados por

$$Y_i = \begin{cases} 1, \text{ se } W_i^* - \lambda > 0 \\ 0, \text{ se } W_i^* - \lambda \leq 0 \end{cases}.$$

Deve já ser claro que o modelo resultante das transformações indicadas é, essencialmente, o postulado em [41]-[43]: a perturbação $\frac{w_i - \mu}{\sigma}$ verifica [43] (isto é, tem distribuição normal de parâmetros 0 e 1), a relação entre $Y_i$ e a variável latente é a prevista em [42] e, do confronto da equação [47] com a equação [41], depreende-se que as observações extraídas das duas populações serão coincidentes se for

$$\delta_1 = \frac{\gamma_1 + \mu - \lambda}{\sigma},$$

$$\delta_j = \frac{\gamma_j}{\sigma}, \text{ para } j = 2, 3, ..., k. \qquad [48]$$

De facto, se as condições em [48] se verificarem, ter-se-á, para qualquer $\mathbf{X_i}$,

$$E(Z_i^* \mid \mathbf{X_i}) = E(\frac{W_i^* - \lambda}{\sigma} \mid \mathbf{X_i}),$$

$$Var(Z_i^* \mid \mathbf{X_i}) = Var(\frac{W_i^* - \lambda}{\sigma} \mid \mathbf{X_i}),$$

e, como $Z_i^*$ e $\frac{W_i^* - \lambda}{\sigma}$ têm ambas distribuição normal, a identidade da média e da variância é suficiente para garantir a identidade das distribuições.

Do que precede, conclui-se que as observações geradas pelo modelo descrito pelas relações [41]-[43] não são distinguíveis das provenientes de qualquer outra população normal cujos coeficientes das variáveis independentes satisfaçam as condições [48]. Os coeficientes de um modelo *probit* são identificáveis excepto por um factor de proporcionalidade comum e, no caso do termo independente, a menos de uma constante.

Analisar-se-á, em seguida, a questão da estimação. Como se viu, a função de verosimilhança correspondente à especificação em [41]-[43] é

$$L(\boldsymbol{\delta}) = \prod_{i=1}^{n} \left\{ \left[1 - \Phi(\mathbf{X_i}\, \boldsymbol{\delta})\right]^{1-Y_i} \left[\Phi(\mathbf{X_i}\, \boldsymbol{\delta})\right]^{Y_i} \right\}, \qquad [49]$$

em que $\Phi(.)$ designa a função de distribuição normal reduzida. Seguindo um caminho análogo, deduzir-se-á aqui a função de verosimilhança, $L(\boldsymbol{\gamma}, \lambda, \mu, \sigma^2)$, para o modelo em [44]-[46].

Tem-se

$$\begin{aligned}
\mathrm{Prob}(Y_i = 0) &= \mathrm{Prob}(W_i^* \le \lambda) \\
&= \mathrm{Prob}(\mathbf{X_i}\, \boldsymbol{\gamma} + w_i \le \lambda) \\
&= \mathrm{Prob}(w_i \le \lambda - \mathbf{X_i}\, \boldsymbol{\gamma}) \\
&= \mathrm{Prob}\left(\frac{w_i - \mu}{\sigma} \le \frac{\lambda - \mu - \mathbf{X_i}\, \boldsymbol{\gamma}}{\sigma}\right) \\
&= \Phi\left(\frac{\lambda - \mu - \mathbf{X_i}\, \boldsymbol{\gamma}}{\sigma}\right).
\end{aligned}$$

Por consequência,

$$\mathrm{Prob}(Y_i = 1) = 1 - \Phi\left(\frac{\lambda - \mu - \mathbf{X_i}\, \boldsymbol{\gamma}}{\sigma}\right).$$

Usando o facto de que, para a distribuição normal reduzida,

$$\Phi(x) = 1 - \Phi(-x)\,,$$

a função de verosimilhança vem dada por

$$L(\boldsymbol{\gamma}, \lambda, \mu, \sigma^2) = \prod_{i=1}^{n} \left\{ \left[1 - \Phi\left(\frac{\mu - \lambda + \mathbf{X_i}\, \boldsymbol{\gamma}}{\sigma}\right)\right]^{1-Y_i} \left[\Phi\left(\frac{\mu - \lambda + \mathbf{X_i}\, \boldsymbol{\gamma}}{\sigma}\right)\right]^{Y_i} \right\}. \qquad [50]$$

Seja $\boldsymbol{\beta}$ o vector $(k\times 1)$ definido por

$$\boldsymbol{\beta} = \begin{bmatrix} (\gamma_1 + \mu - \lambda) / \sigma \\ \gamma_2 / \sigma \\ \gamma_3 / \sigma \\ \dots \\ \gamma_k / \sigma \end{bmatrix}.$$

Então, a função de verosimilhança em [50] pode ser escrita como

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \left[ 1 - \Phi(\mathbf{X_i}\ \boldsymbol{\beta}) \right]^{1-Y_i} \left[ \Phi(\mathbf{X_i}\ \boldsymbol{\beta}) \right]^{Y_i} \right\}. \qquad [51]$$

Sejam $\hat{\boldsymbol{\delta}}_{ML}$ e $\hat{\boldsymbol{\beta}}_{ML}$ os vectores que maximizam, respectivamente, as funções de verosimilhança em [50] e em [51]. Alguns momentos de reflexão devem ser os bastantes para persuadir o leitor de que terá de ser

$$\hat{\boldsymbol{\delta}}_{ML} = \hat{\boldsymbol{\beta}}_{ML}$$

e para extrair desse facto as consequências seguintes:

i) se a especificação em [44]-[46] for a correcta, não é possível estimar $\sigma$ separadamente dos outros parâmetros do modelo, nem estimar $\gamma_1$, $\mu$ e $\lambda$, individualmente;

ii) estimado o modelo [41]-[43], as estimativas obtidas para os coeficientes das variáveis explicativas são, de facto, estimativas de quocientes da forma (coeficiente "verdadeiro"/$\sigma$), em que $\sigma$ designa o desvio-padrão da perturbação aleatória.

## BIBLIOGRAFIA


DAVIDSON, Russell, e MACKINNON, James G. (1993), *Estimation and Inference in Econometrics*, Oxford University Press, New York.

GREENE, William H. (1997), *Econometric Analysis*, 3rd ed., Prentice-Hall, Upper Saddle River, New Jersey.

GUJARATI, Damodar N. (1995), *Basic Econometrics*, 3rd ed., McGraw-Hill, New York.

JUDGE, George G.; GRIFFITHS, W. E.; HILL, R. Carter; LUTKEPOHL, Helmut, e LEE, Tsoung-Chao (1985), *The Theory and Practice of Econometrics*, 2nd ed., John Wiley and Sons, New York.

MADDALA, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge.