MODELOS DE RESPOSTAS BINÁRIAS: ESPECIFICAÇÃO, ESTIMAÇÃO E INFERÊNCIA[1]

Ricardo Chaves Lima[2]

RESUMO

Esse trabalho analisa os aspectos relacionados à especificação, estimação e inferência de modelos econométricos em que a variável dependente é dicótoma. Será mostrado que o modelo linear de probabilidade pode ser estimado usando-se o método dos mínimos quadrados, mas essa técnica de estimação apresenta dois problemas: uma variância do erro heteroscedástica e a possibilidade de estimativas de probabilidades fora do intervalo entre 0 e 1. Os modelos logito e probito serão considerados como alternativas viáveis para estimação de modelos de resposta binárias. Quando dados não agrupados são usados, os modelos logito e probito são não lineares nos parâmetros e exigem um processo iterativo de estimação. O método da máxima verossimilhança é o mais comum na estimação desse tipo de modelo.

Palavras-chave: resposta binária, probito, logito, máxima verossimilhança.

BINARY RESPONSE MODELS: SPECIFICATION, ESTIMATE AND INFERENCE

SUMMARY

This paper analyzes aspects of specification, estimation and inference of econometric models with dichotomous dependent variable. It will be showed that the linear probability model can be estimated by using the least squares method, but this technique brings up two main shortcomings: an heteroscedastic error variance and the possibility of estimates of probabilities out of the 0 to 1 interval. The logit and probit models, then, will be considered as feasible alternatives of estimation for qualitative response models. When ungrouped data are used, logit and probit models are non-linear in parameters and must be estimated by interactive process. Maximum likelihood is a common estimation process used in such cases.

Key-words: binary response, probit, logit, maximum likelihood.

1 - INTRODUÇÃO

Algumas variáveis utilizadas em pesquisas socioeconômicas não são observadas de forma contínua. Estas, geralmente, representam respostas binárias dos indivíduos. É difícil, por exemplo, observar a propensão de um agricultor a adotar uma inovação tecnológica. O que se observa empiricamente é a decisão de adotar ou não a nova técnica. Nesse caso, a resposta do evento "adoção de tecnologia" é binária, do tipo sim ou não. Os modelos de respostas binárias são aqueles em que a variável dependente assume valores discretos. A probabilidade de ocorrência de cada resposta binária, de acordo com este modelo, é uma função de um conjunto de atributos dos indivíduos tais como renda, idade, sexo, estado civil, etc. Um dos principais objetivos dos modelos de respostas binárias, portanto, é calcular a probabilidade de um indivíduo com um determinado conjunto de atributos tomar uma decisão relativa a um evento dado (PINDYCK & RUBINFELD, 1981).

O presente trabalho objetiva discutir os aspectos relativos à especificação, estimação e inferência de modelos econométricos em que a variável dependente representa uma resposta binária. Serão discutidos os modelos: linear de probabilidade, logito e probito. O modelo linear de probabilidade será considerado apenas para efeito de discussão teórica,

[1]Recebido em 19/10/95. Liberado para publicação em 12/03/96.

[2]Ph.D em Economia Agrícola e Professor recém-doutor/CNPq do Departamento de Economia Agrícola da Universidade Federal do Ceará.

uma vez que este método pode gerar estimadores ineficientes e estimativa de probabilidades fora do intervalo zero e um. Os modelos logito e probito serão apresentados como alternativas viáveis para estimação de modelos de respostas binárias. O presente trabalho mostrará que o processo de estimação utilizado em modelos de respostas binárias é geralmente não linear, e que o método de máxima verossimilhança pode ser usado para obter estimadores eficientes. Será também discutido o caso do modelo logito com dados agrupados em que o método dos mínimos quadrados é capaz de produzir estimadores com propriedades estatísticas desejáveis.

2 - ESPECIFICAÇÃO DE MODELOS DE RESPOSTAS BINÁRIAS

Considere inicialmente que uma resposta binária é função de um índice latente (não observado) que varia de um mínimo a um máximo, passando por um nível limite o qual determina uma mudança de qualidade na resposta de um indivíduo. Considere também que as variações no referido índice são uma função dos atributos do indivíduo. Assim, pode-se concluir que as respostas binárias dos indivíduos são uma função dos atributos dos mesmos.

Para ilustrar a situação acima, considere o exemplo da adoção de uma nova tecnologia por parte de um agricultor. Seja Y_i uma variável binária que representa a decisão do i-ésimo agricultor em adotar uma nova tecnologia. Para efeito de operacionalização considere Y_i igual a 1 quando o agricultor adota a referida tecnologia e 0 o caso contrário. Seja I_i um índice latente que represente a propensão do i-ésimo agricultor em adotar a nova tecnologia. O referido índice varia de um mínimo a um máximo, passando por um nível limite I^* o qual determina a mudança de atitude do agricultor com relação à adoção da nova tecnologia. Ou seja:

$$Y = \begin{cases} 1, \text{ quando } I_i > I^* \\ 0, \text{ quando } I_i \leq I^* \end{cases}$$

Pede-se então escrever a resposta binária dos indivíduos como uma função da variável latente I da seguinte forma:

$$Y_i = f(I_i) \qquad (1)$$

A propensão dos agricultores em adotar uma nova tecnologia (I_i), no entanto, é uma variável não observada empiricamente. Assumindo-se que I_i é uma função linear dos k atributos dos indivíduos (X_1, X_2, ...X_K), pode-se então escrever (1) como segue:

$$y_i = F(X_i' \beta) \qquad (2)$$

onde, para T observações ($i = 1, \ldots, T$), y_i é um vetor ($T \times 1$) de observações da variável dependente, x_i é um vetor ($K \times 1$) de variáveis independentes e β é um vetor ($K \times 1$) de parâmetros a serem estimados. O tipo de modelo utilizado na estimação estatística de (2) depende da escolha de F. As formas funcionais mais comuns em aplicações de modelos de respostas binárias são as seguintes (JUDGE *et al.*, 1985):

Modelo de linear probabilidade:

$$F(X_i'\beta) = X_i'\beta$$

Modelo Probito:

$$F(X_i'\beta) = \Phi(X_i'\beta) = \int_{-\infty}^{X_i'\beta} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

Modelo Logito:

$$F(X_i'\beta) = L(X_i'\beta) = \frac{1}{1 + e^{-X_i'\beta}}$$

onde $\Phi(.)$ representa a função de densidade normal cumulativa, $L(.)$ representa a função logística cumulativa, e representa a base do logaritmo natural e π uma constante com valor aproximado de 3.1416.

O modelo linear de probabilidade assume que as resposta binária dos indivíduos são uma função linear do conjunto de atributos dos mesmos. A utilização de métodos usuais de regressão linear para a estimação de modelos de respostas binárias, no entan-

to, apresenta algumas dificuldades. Duas destas dificuldades são: 1) a possibilidade de se obter estimativas de probabilidades fora do intervalo entre zero e um e 2) a obtenção de termos de erro não homoscedástico (JUDGE et al., 1985; AMEMIYA, 1981).

Para ilustrar estas dificuldades pode-se usar o exemplo da adoção de uma nova tecnologia por parte de um agricultor. Considere que a decisão de adotar a nova tecnologia (y_i) é uma função do número de anos de educação formal do agricultor (x_i), de forma que o modelo linear de probabilidade possa ser escrito da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (3)$$

onde β_0 e β_1 são os parâmetros do modelo e e_i o termo de erro. Considere também, para efeito de ilustração, que uma determinada amostra de agricultores foi pesquisada e que a variável x_i apresentou um valor mínimo (x_{min}) e valor um máximo (x_{max}). A utilização de observações da variável dependente fora da amostra considerada, tais que $x_i > x_{max}$ ou $x_i < x_{min}$, pode gerar estimativas de probabilidade maiores do que um ou menores do que zero, o que não tem significado estatístico. A outra complicação do modelo linear de probabilidade é que o termo de erro não pode ser considerado homoscedástico. Considerando-se a equação (3) acima como representando um modelo de escolha binária, y_i assume valores 0 e 1 com probabilidades de $(1 - P_i)$ e P_i, respectivamente. Assim, e_i assume valores $(1 - \beta_o - \beta_1 x_i)$ e $(- \beta_o - \beta_1 x_i)$ com probabilidade $(1 - P_i)$ e P_i, respectivamente. A variância do erro pode ser dada por:

$$Var\ [e_i] = E\ (e_i^2) = (1 - \beta_o - \beta_1\ x_i)^2\ P_i + (-\beta_o - \beta_1\ x_i)^2\ (1 - P_i)$$

onde *Var* representa a variância e E a esperança. Pode-se então mostrar, de acordo com PINDYCK & RUBINFELD (1981), que:

$$Var\ [e_i] = P_i\ (1 - P_i)$$

Ou seja, quando P_i estiver próximo a zero ou um a variância do erro será mínima, e quando P_i estiver próximo a 0.5 a variância do erro será máxima. A variância do erro, portanto, depende de P_i, o que significa que o erro aleatório é heteroscedástico. De acordo com GREENE (1993), o problema da heteroscedasticidade no modelo linear de probabilidade é menor e pode ser contornado aplicando-se o método dos Mínimos Quadrados Generalizados Factíveis (Feasible Generalized Least Squares). O grande problema é que não se pode garantir previsões de probabilidades restritas ao intervalo de zero a um. A figura 1 é uma representação gráfica de um modelo linear de probabilidade. De acordo com o referido exemplo, para valores de x que tornem o índice latente I menor que -3 ou maior do que 3, o modelo linear de probabilidade produzirá estimativas de P_i fora do intervalo entre 0 e 1.

Sendo o maior problema do modelo linear a possibilidade de estimativas de probabilidade fora do intervalo entre 0 e 1, é natural que se procure modelos que produzam estimativas de probabilidade dentro desse intervalo. A solução óbvia é transformar o modelo original de forma que as estimativas de probabilidade estejam restritas ao intervalo entre 0 e 1 para qualquer x_i. Essa transformação deve ser tal que um aumento (diminuição) na probabilidade de um determinado evento binário ocorrer deve estar associado a um aumento (diminuição) nos valores das variáveis explicativas. Ou seja, deve-se assumir uma transformação monotônica (PINDYCK & RUBINFELD, 1981). Os modelos probito e logito permitem essa transformação.

O modelo probito é baseado na função de distribuição normal cumulativa, a qual possibilita uma transformação no modelo garantindo que, para qualquer x, as estimativas de probabilidade estejam no intervalo entre zero e um. Assim, a probabilidade do evento qualitativo ocorrer tende a 0 quando I_i decresce para $-\infty$, e tende a 1 quando I_i cresce para $+\infty$. O modelo logito é baseado na função de probabilidade logística cumulativa, a qual garante que as estimativas de probabilidade cairão dentro do intervalo entre zero e um. Da mesma forma que o probito, a probabilidade de um evento qualitativo ocorrer no modelo logito tende a zero quando I_i decresce para $-\infty$, e tende a um quando I_i cresce para $+\infty$. A figura 2 é uma representação gráfica das funções de probabilidade normal cumulativa e logística cumulativa, mostrando que, para ambas as distribuições, a inclinação da curva é maior

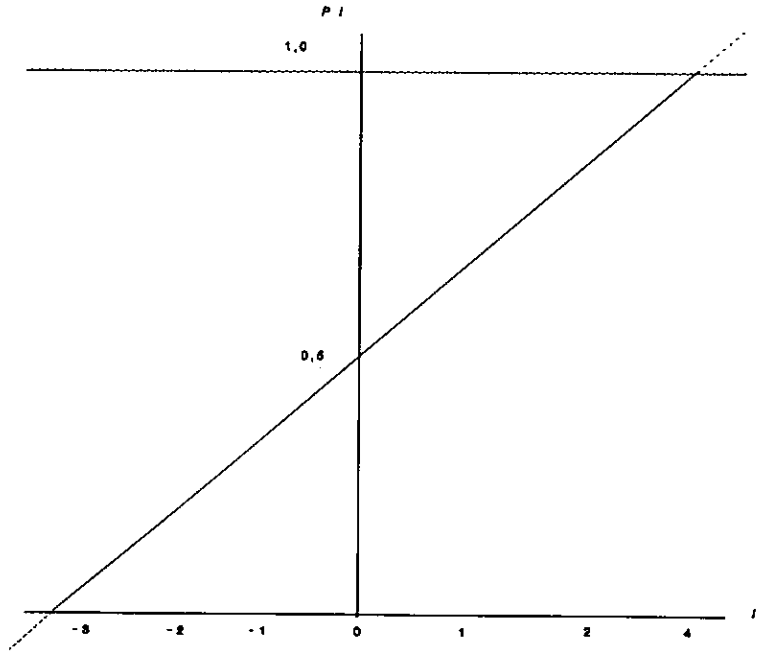

FIGURA 1 - Representação Gráfica das Funções de Probabilidade Linear.

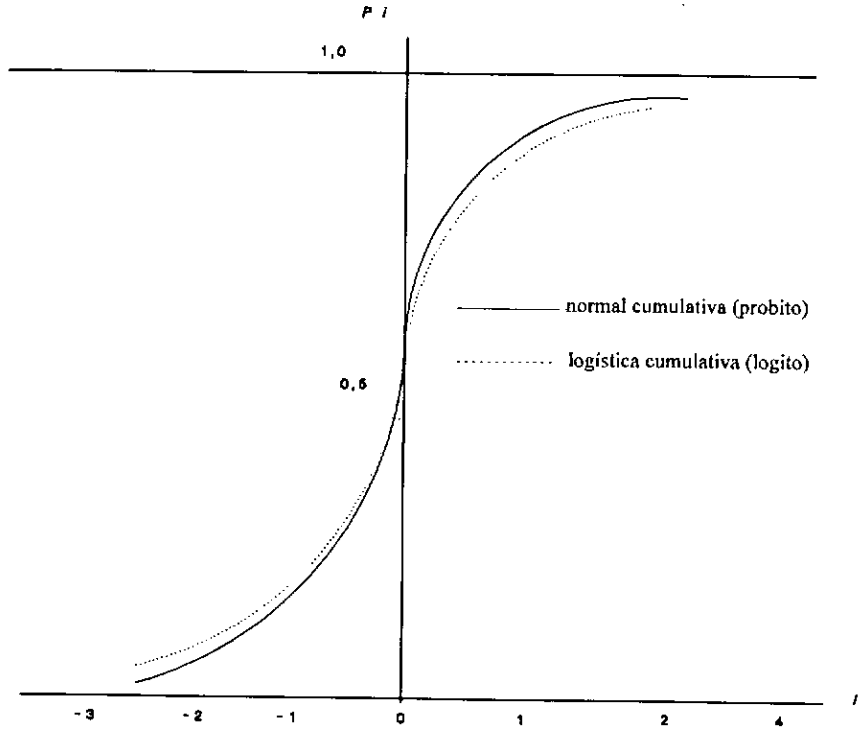

FIGURA 2 - Representação Gráfica das Funções de Probabilidade Normal Cumulativa e Logística Cumulativa.

quando P_i está próximo a 0.5, e é menor para valores de P_i próximos a 0 ou 1. Em termos de modelo de regressão isso significa que mudanças na variável dependente próximas ao ponto médio da distribuição terão um impacto maior em P_i do que as mudanças próximas aos extremos da distribuição (PINDYCK & RUBINFELD, 1981).

3 - ESTIMAÇÃO DE MODELOS DE RESPOSTAS BINÁRIAS

O método de estimação utilizado em modelos de respostas binárias depende de os dados coletados na amostra estarem ou não agrupados. No caso de dados agrupados o método dos mínimos quadrados pode ser utilizado. Para ilustrar o referido caso considere o exemplo anterior da adoção de uma nova tecnologia por parte de um agricultor. Suponha, por exemplo, que a amostra pesquisada é constituída de T grupos, e que em cada grupo existem n indivíduos, dentre os quais r adotaram a nova tecnologia. Uma estimativa de probabilidade de adoção da nova tecnologia para cada grupo será:

$$\hat{P} = \frac{r_t}{n_t}$$

O modelo logito com dados agrupados, portanto, pode ser especificado da seguinte forma:

$$log \frac{\hat{P_t}}{1-\hat{P_t}} = log \frac{r_t/n_t}{1-r_t/n_t} = \qquad (4)$$
$$log = \frac{r_t}{n_t - r_t} = \beta_0 + \beta_1 X_t + \epsilon_t$$

A equação (4) é linear em parâmetros e pode ser estimada pelo método dos mínimos quadrados. Os parâmetros estimados são consistentes (PINDYCK & RUBINFELD, 1981).

No caso da utilização de observações individuais, o método mais comum de estimação dos modelos probito e logito é o de máxima verossimilhança. O método da máxima verossimilhança objetiva estimar parâmetros que maximizem a probabilidade de uma determinada amostra pertencer a uma população dada. Considerando-se $(y_1, y_2, \ldots y_n)$, as observações de uma amostra aleatória normalmente distribuída, e $[p(y_1), p(y_2), \ldots, p(n)]$ as respectivas probabilidades associadas à distribuição normal, a função de máxima verossimilhança (L) é dada por:

$$L = P(y_1).P(y_2).\cdots P(y_n)$$

Os estimadores de máxima verossimilhança de um modelo linear do tipo $y_t = x_t'\beta + \epsilon_t$ podem ser obtidos da expressão:

$$L = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(y_t - x_t'\beta)^2\right]$$

pela derivada parcial de L com relação ao respectivo parâmetro.

Para ilustrar este método considere o exemplo da adoção de uma nova tecnologia por agricultores. Seja P_i a probabilidade de um determinado agricultor adotar uma nova tecnologia, e Y_i a variável observada que assume valores 1 quando o agricultor adotar a referida tecnologia, e 0 caso contrário. O objetivo do processo de estimação é encontrar parâmetros que maximizem a probabilidade de o padrão de escolha da amostra ter ocorrido na população. Assim, assumindo que Y_i é obtido de uma distribuição binomial, a função de verossimilhança (L) pode ser escrita da seguinte forma:

$$L = P(y_1).P(y_2).\cdots P(y_n) = \prod_{t=1}^{t_1} P_t \prod_{t=t_1+1}^{T} (1-P_t) \qquad (5)$$

onde t_1 é o número de vezes em que o agricultor adota a referida tecnologia. Considerando $F(x_i'\beta)$ a forma funcional do modelo de resposta binária, pode-se escrever a função de verossimilhança da seguinte forma (GREENE, 1993):

$$L = \prod_{i=1} [F(x_i'\beta)]^{y_i} [1 - F(x_i'\beta)]^{1-y_i} \qquad (6)$$

Tirando-se o logaritmo natural de (6) tem-se a função de log-verossimilhança:

$$lnL = \ell = \sum_t [y_t \ln F(x_t'\beta) + (1 - y_t) \ln (1 - F(x_t'\beta))] \qquad (7)$$

A obtenção de estimadores para o vetor de parâmetros β é feita diferenciando-se ℓ com relação a cada ele-

mento de β e igualando-se a zero. O processo de estimação é não linear e requer uma solução iterativa. Os parâmetros obtidos na convergência final do processo iterativo têm matriz de covariância assintótica dada pelo inverso da matriz de informação[3]. Essas estimativas de variância e covariância permitem a realização de teste de hipótese dos parâmetros e uma análise da dimensão do termo de erro (MADDALA, 1983).

A estimação de equações do tipo (7) torna-se simples com a utilização de programas de computador que trazem algoritmos específicos para a estimação de modelos como logito e probito. Um exemplo desses pacotes estatísticos são os programas LINDEP, SAS e SHAZAM.

4 - INFERÊNCIA COM MODELOS DE RESPOSTAS BINÁRIAS

No modelo linear de probabilidade, os coeficientes das variáveis explicativas têm o mesmo significado que no método dos mínimos quadrados. Ou seja, o coeficiente de uma variável explicativa mede uma mudança na variável explicada como resultado de uma variação unitária na referida variável explicativa *ceteris paribus*. Nos modelos probito e logito os coeficientes estimados medem o impacto de cada variável explicativa no índice latente, e não na variável explicada (WHITE, 1993). O impacto da variável explicativa na variável explicada nos modelos logito e probito, o qual é denominado efeito marginal, representa a inclinação das curvas logística cumulativa e normal cumulativa, respectivamente, para cada observação.

Sejam x_{ik} o k-ésimo elemento do vetor de variáveis explicativas $\mathbf{x}_i$ e β_k o k-ésimo elemento de β, o efeito marginal para uma determinada variável x_i é calculado como a derivada parcial da função de resposta binária com relação a x_i, como é mostrado abaixo (AMEMIYA, 1981).

Modelo linear de probabilidade:

$$\frac{\partial}{\partial x_{ik}} (X_i' \beta) = \beta_k$$

Modelo Probito:

$$\frac{\partial}{\partial x_{ik}} \Phi (X_i' \beta) = \Phi (X_i' \beta) \cdot \beta_k$$

Modelo Logito:

$$\frac{\partial}{\partial x_{ik}} L (X_i' \beta) = \frac{e^{X_i' \beta}}{(1 + e^{X_i' \beta})^2} \cdot \beta_k$$

O efeito marginal para o modelo linear de probabilidade é o próprio coeficiente β_k como no caso dos mínimos quadrados. No caso dos modelos probito e logito, o efeito marginal não é dado diretamente por β_k, mas pelas derivadas parciais mostradas acima. O efeito marginal para uma determinada variável representa uma mudança na probabilidade de um dado evento ocorrer quando o valor da referida variável experimenta uma mudança unitária.

Para ilustrar o significado do efeito marginal pode-se usar o exemplo da adoção de tecnologia por um agricultor. Considere o modelo apresentado na equação (3), onde a adoção de uma nova tecnologia por um agricultor é representada por uma variável binária (1 = adoção e 0 = caso contrário), que é explicada pelo nível de educação, representado pelos anos de escolaridade do agricultor. O efeito marginal, nesse caso, representa o impacto de cada ano de escolaridade do agricultor na probabilidade da ocorrência de adoção da nova tecnologia. Esse resultado, portanto, pode ser usado para a previsão da probabilidade de um determinado evento ocorrer dado um conjunto de atributos do indivíduo.

Outro resultado que pode ser usado para efeito de inferência é a elasticidade da probabilidade de um determinado evento ocorrer com relação a um determinado atributo do indivíduo. Enquanto o efeito marginal representa o impacto na probabilidade da resposta binária decorrente de uma mudança unitária em x_i, a elasticidade representa uma mudança percentual na probabilidade da resposta binária como resultado de uma variação de 1% em x_i. A elasticidade, portanto, permite comparar o efeito relativo de cada variável explicativa na probabilidade de ocorrência do

[3]A matriz de informação $| I(\hat{\beta}) |$ é dada pelo valor negativo da esperança das derivadas segundas da função de máxima verossimilhança com relação aos parâmetros a serem estimados, da seguinte forma:

$$I(\hat{\beta}) = E\left(-\frac{\partial^2 \log L}{\partial \beta \partial \beta'} \right)$$

evento binário. A elasticidade da probabilidade, assim como o efeito marginal, varia ao longo das curvas normal cumulativa e logística cumulativa assumindo um valor para cada x_i. Pode-se, no entanto, calcular a elasticidade para um valor médio de x_i. A elasticidade da probabilidade P_i com relação a x_i é calculada no ponto como segue:

$$E_{ik} = \left(\frac{\partial P_i}{\partial x_{ik}} \right) \cdot \frac{x_{ik}}{F(X_i' \beta)}$$

onde, para o valor médio de x_i

$$E_k = \left(\frac{\partial P_i}{\partial x_{ik}} \right) \cdot \frac{x_{ik}}{F(\bar{X}_i' \beta)}$$

onde $F(x_i'\beta)$ corresponde à forma funcional a ser considerada.

5 - CONSIDERAÇÕES FINAIS

O propósito do presente trabalho foi discutir alguns aspectos relacionados à especificação, estimação e inferência de modelos econométricos em que a variável dependente representa uma resposta binária. Considerou-se apenas o caso em que a resposta é binária, ou seja, em que o indivíduo está diante de uma situação com apenas duas opções de escolha. Existem situações, no entanto, em que a resposta qualitativa é multinomial. Ou seja, o indivíduo deve escolher entre múltiplas opções. Seria o caso, por exemplo, de um agricultor que teria de optar por uma entre várias inovações tecnológicas. Nessa situação deve-se utilizar modelos de respostas qualitativas modificados tais como logito multinomial ou probito multinomial. Uma discussão aprofundada sobre esse tipo de modelo pode ser encontrada em MADDALA (1983), GREENE (1993) e AMEMIYA (1981).

LITERATURA CITADA

AMEMIYA, Takeshi. Qualitative response models: a survey. **Journal of Economic Literature**, California, v.19, p.1493-1536, Dec. 1981.

GREENE, William H. **Econometric analysis**. New York: MacMillan, 1993.

JUDGE, George G. et al. **The theory and practice of econometrics**. New York: Wiley, 1985. 101p.

MADDALA, G.S. **Limited dependent and qualitative variables in econometrics**. Madison: Cambridge University, 1983.

PINDYCK, Robert S. & RUBINFELD, D. **Econometric models and economic forecasts**. New York: MacGraw-Hill, 1981.

WHITE, K. **SAHAZAM User's reference manual**: version 7.0. New York: MacGraw-Hill, 1993.