Introdução à análise estatística de variáveis dicotômicas e aplicações em dados socioeconômicos

Sergio Luiz de Bragança *

Esta sucinta exposição tem como objetivo apresentar algumas idéias ligadas ao problema da inferência estatística sobre variáveis dicotômicas, tendo em mente aplicações diante de uma grande massa de dados.

O autor não tem a pretensão de disputar a originalidade de quaisquer dos resultados aqui incluídos, todos podendo ser encontrados nas referências de cada item.

* Da SUEGE/IBGE. O autor agradece ao IBGE pelo incentivo e ambiente propício à concentração, tão necessários a qualquer trabalho científico; à FGV pela oportunidade de contacto com renomados cientistas através da organização de ciclos conferenciais, tendo esta publicação sido inspirada pelas palestras do Prof. Marc Nerlove sobre o seu trabalho *Univariate and multivariate log-linear and logistic model for the analysis of qualitative socio-economic data* em co-autoria com o Prof. S. J. Press, ambos de Chicago, EUA; ao Prof. José Luiz Carvalho, vice-diretor da EPGE, pela sugestão da adoção do modelo logístico no estudo do fenômeno de migrações; a J. B. Burle de Figueiredo, da Organização Internacional do Trabalho, pelo incentivo e oportunidade de aplicação efetiva das técnicas aqui apresentadas; a Luís Koodi Hotta e Sebastião Amorim pelas críticas, sugestões e um exaustivo trabalho de programação; a Geraldo Machado Costa pelo sempre paciente suporte computacional e a programação da rotina de regressão ponderada; e às inúmeras pessoas que porventura tiveram alguma contribuição, que certamente somam uma boa parcela do artigo, e que por injustiça da memória não tiveram seus nomes citados.

Em vista do cunho introdutório do artigo, este contém um estudo limitado ao caso de uma variável dicotômica, ficando o caso multivariado politômico para outra ocasião, ou mesmo para leitura individual das diversas referências que serão citadas no final. A ênfase está concentrada na apresentação dos resultados sem demonstração matemática, porém, arrumados objetivamente dentro de uma seqüência lógica que facilite sua compreensão, colocando o leitor em posição confortável para utilização das técnicas desenvolvidas ao longo das próximas páginas.

Variáveis dicotômicas aparecem em dados de qualquer espécie, seja naturalmente ou através da discretização e agregação de valores de variáveis contínuas ou politômicas. Sem a preocupação quanto à relevância da variável, mas com o intuito de apresentar uma ilustração, podemos citar vários exemplos. Evidentemente, variáveis dicotômicas sempre dizem respeito ao acontecimento ou não de um evento genérico. E que se convenciona denotar cada observação i como $y_i = 1$ se o evento ocorrer, e $y_i = 0$ caso contrário, como faremos a seguir:

1. Em estudos sobre a efetividade de uma determinada droga ou substância no tratamento de doenças ou extermínio de pestes ou animais nocivos; em estudos bioquímicos em geral, tem-se:

$y_i = 1$, quando a dose ministrada foi suficiente para a consecução dos objetivos definidos;

$y_i = 0$, caso contrário.

2. Em dados socioeconômicos, variáveis ligadas a fenômenos como migrações, fecundidade, força de trabalho, fornecem exemplos os mais variados:

a) $y_i = 1$, caso o indivíduo seja um migrante de determinada região;

$y_i = 0$, caso contrário; ou

b) $y_i = 1$, caso uma mulher pertença a população economicamente ativa;

$y_i = 0$, caso contrário; etc.

3. Assuntos relacionados com a engenharia, como resistência de materiais, no teste de um novo material fornecem observações como abaixo:

$y_i = 1$, caso o material tenha resistido a determinada pressão ou choque;

$y_i = 0$, caso contrário.

4. Estudos de comportamento humano em psicologia, através de respostas do tipo certo ou errado em testes dirigidos, etc.

O presente artigo discorrerá sobre critérios e técnicas para se acessarem problemas totalmente análogos aos que se estudam em análise de covariância. Cada observação será um vetor (y_i, x_i) sobre um mesmo indivíduo i, onde Y é uma variável dicotômica e X um vetor de variáveis de qualquer tipo, e será examinada a questão de se quantificar a explicação da probabilidade de Y tomar o valor 1 dadas as condições definidas pelo vetor X. Portanto, em resumo, a diferença entre a análise de covariância e o presente tópico reside no fato de no primeiro o interesse estar voltado para os valores da variável dependente Y e, no segundo, na probabilidade de Y tomar um determinado valor.

Por exemplo, se Y representar o resultado do teste em um indivíduo submetido a uma dose de alguma droga específica, ou seja, $Y = 1$ caso haja sucesso e $Y = 0$ caso contrário, poder-se-ia estar interessado no fato de a probabilidade de sucesso ser ou não independente da dosagem dentro de certos limites e, caso não seja, a que dosagem X essa probabilidade apresentar-se-ia superior a um valor prefixado.

No caso de um estudo sobre migrações poder-se-ia estar interessado em detectar quais as variáveis, de um elenco, que melhor explicam a decisão de uma pessoa migrar. Nesse exemplo, ter-se-iam observações do tipo $(Y, x_1, \ldots, x_n)$, onde Y toma os valores um ou zero se o indivíduo é um migrante ou não, respectivamente, e $x_1 \ldots x_n$ são códigos para as variáveis explicativas que se estivessem considerando, como idade, renda, nível educacional, etc.

O trabalho foi organizado em quatro itens. O primeiro item diz respeito à modelagem, especificamente, sendo apresentadas algumas opções, embora haja uma justificada tendenciosidade em favor do modelo logístico. A primeira parte deste item contém uma análise da inadaptabilidade do modelo linear no estudo de variáveis categóricas, que serve como excelente introdução à segunda parte, onde estão apresentados o modelo logístico e alternativas. O segundo item contém o estudo da estimação dos parâmetros dos modelos expostos no item anterior, subdividido em duas partes segundo as características da amostra em consideração. O terceiro item refere-se a problemas com que, certamente, se confronta em aplicações, seja devido a aspectos computacionais ou a características da amostra — apresentaremos sugestões de como contorná-los ou amenizá-los. No quarto item o autor apresenta uma análise simplificada de uma apli-

cação do material exposto nos itens anteriores a um estudo sobre migrações. Ao final do trabalho apresenta-se uma lista selecionada de referências bibliográficas que se julga útil aos interessados no estudo de variáveis categóricas.

1. Modelos de análise

Digamos que se queira analisar a probabilidade de ocorrência de um evento E e que para tal se disponha de um elenco de k variáveis que, supostamente, exerçam alguma influência sobre essa probabilidade. Precisando melhor, digamos que se tenham obtido, através de uma amostragem, as observações independentes (y_j, x_j), $1 \leqslant j \leqslant m$, do par (Y, X) onde Y toma os valores um e zero, respectivamente, caso o evento E ocorra ou não, nas condições especificadas pelo vetor X de k variáveis.

1.1 Modelo linear de probabilidade

Ao se adotar o modelo tradicional de regressão ficaria postulado o seguinte comportamento:

$$y_j = x_j' \beta + \varepsilon_j, \; 1 \leq j \leq n,$$

onde ε denota uma variável representando um desvio tal que $E(\varepsilon_j) = 0$, $Var(\varepsilon_j) = \sigma^2$ e $E(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$. Ou seja, os desvios têm média zero, variância constante e correlação nula. Portanto, tem-se que $E(y_j) = x'_j\beta$, onde $E(y_j)$ nada mais é do que a probabilidade de a variável Y tomar o valor 1 nas condições especificadas pelo vetor x_j uma vez que Y é uma variável dicotômica de códigos um ou zero, isto é,

$$P(y_j = 1) = P(Y = 1 \mid X = x_j) = E(y_j) = x_j' \beta$$

Todavia, na identidade acima tem-se, por um lado, um valor compreendido entre zero e um, e por outro, um produto escalar $X' \beta$ podendo tomar qualquer valor real. Assim, se numa aplicação específica o intervalo de valores para a variável X satisfizesse a identidade $0 \leqslant X \beta \leqslant 1$, ainda poder-se-ia considerar a equação obtida pela regressão como uma primeira aproximação linear da probabilidade $P(Y = 1 \mid X)$ dentro daquele intervalo. Caso contrário, não restaria outra alternativa senão postular que para os valores de X tais que $X' \beta > 1$ ou $X' \beta < 0$ então $P(Y = 1 \mid X) = 1$

ou $P\ (Y = 1 \mid X) = 0$, respectivamente. Obter-se-ia, desta forma, uma aproximação da função $P\ (Y = 1 \mid X)$ por uma linha quebrada definida como abaixo (veja figura 1):

$$P(Y = 1 \mid X = x_j) = x_j' \beta \qquad \text{se} \quad 0 \leq x_j' \beta \leq 1$$

$$P(Y = 1 \mid X = x_j) = 1 \qquad \text{se} \quad x_j \beta > 1$$

$$P(Y = 1 \mid X = x_j) = 0 \qquad \text{se} \quad x_j' \beta < 0$$

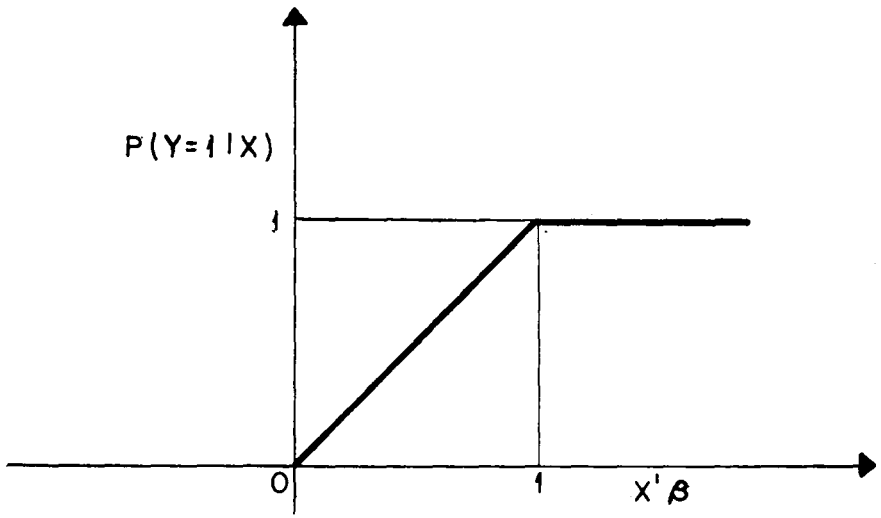

Figura 1 - Representação gráfica de P(Y=1|X)

Este procedimento não só carece de qualquer rigor científico, mas também não acrescenta nenhuma informação sobre o comportamento de função $P\ (Y = 1 \mid X)$ fora do intervalo $0 \leqslant X\beta \leqslant 1$. Portanto, qualquer estudo onde houvesse interesse nas caudas da distribuição seria inviável.

Até este ponto, temos explorado somente as conseqüências da média dos desvios ser nula, no modelo de regressão. A seguir, examinaremos a incompatibilidade entre a constância da variância no modelo de regressão e as propriedades particulares da aplicação em pauta.

Sendo Y uma variável dicotômica, tem-se que para um dado x_j esta variável tem uma distribuição de Bernoulli, portanto, sua variância pode ser expressa em função da média através de conhecida fórmula,

$$Var\ (Y \mid X = x_j) = E\ (Y \mid X = x_j)\ [1 - E\ (Y \mid X = x_j)].$$

Pelo modelo de regressão $E\ (Y \mid X = x_j) = x_j\beta$ e $Var\ (Y \mid X = x_j) = Var\ (\varepsilon \mid X = x_j)$, assim, $Var\ (\varepsilon \mid X = x_j) = x'_j\beta\ (1 - x'_j\beta)$. Ou seja, os

desvios da equação não podem ter variância constante, e mais, para os valores de X tais que $X' \beta < 0$ ou $X' \beta > 1$ a variância toma valores negativos. [1]

1.2 Modelo logístico

Mantendo a analogia com o modelo de regressão para variáveis contínuas, digamos que se queira obter um vetor de parâmetros β no sentido de se tentar explicar a probabilidade do evento E acontecer dentro das condições impostas por um vetor de variáveis X. Em outras palavras, se quer encontrar um vetor de parâmetros β de forma que quanto maior o valor $X' \beta$ tanto maior a probabilidade $P\ (Y = 1 \mid X)$. A própria formulação sugere a aplicação de uma relação funcional monotônica, compreendida no seu conjunto de valores dentro do intervalo [0, 1]. Pode-se, então reformular o problema como sendo de se encontrarem critérios estatísticos para a estimação de uma relação funcional do tipo onde F representa uma função com as propriedades acima (veja figura 2).

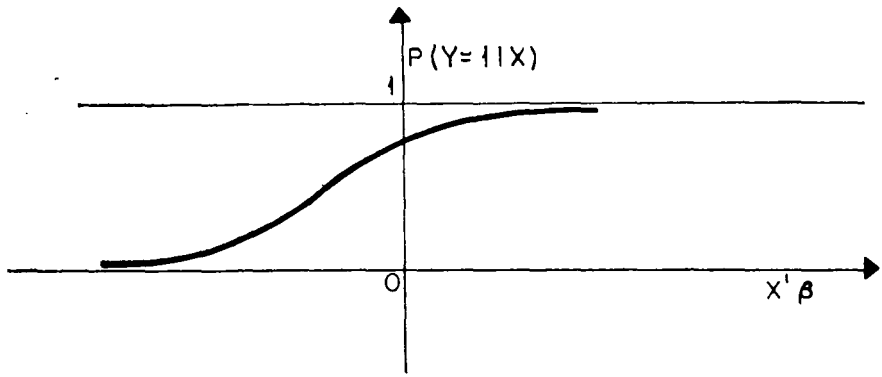

Figura 2 - A função logística

Para cada escolha de F que se faça tem-se um modelo com características próprias, e entre as opções possíveis encontram-se a função de probabilidade comulativa de qualquer distribuição. O modelo logístico utiliza a função cumulativa da distribuição logística, ou seja,

$$F(t) = \frac{1}{1 + e^{-t}}, \quad -\infty < t < \infty.$$

[1] Nerlove & Press (1973).

Assim, fazendo-se $p_j = P\ (Y = 1 \mid X = x_j)$ segue-se que

$$P_j = \frac{1}{1 + e^{-x_j' \beta}}$$

Resolvendo a equação acima em termos de $x_j' \beta$, obtém-se

$$x_j' \beta = log \left(\frac{P_j}{1 - p_j} \right).^2$$

Apresentaremos, a seguir, outras formas para F já estudadas, embora, no próximo item, enumeremos evidências em favor da função logística.

Função cumulativa da distribuição normal:

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx$$

Esta transformação já foi bastante utilizada, principalmente em problemas de estimação relacionados com respostas biológicas a diversos estímulos. [3]

Transformações angulares.

$$F_1(t) = \begin{cases} 1 & \text{se } t > \pi/4 \\ sen^2\ (t + \pi/4) & \text{se } |t| \leq \pi/4 \\ 0 & \text{se } t < -\pi/4 \end{cases}$$

$$F_2(t) = 1/2\ (1 + sent),\ -\pi/2 \leq t \leq \pi/2$$

$$F_3(t) = \frac{1}{2} + \frac{1}{\pi}\ arctg\ (t),\ -\infty < t < \infty$$

$$F_4(t) = tgh\ (t/2), \qquad -\infty < t < \infty.$$

Estas transformações são bem menos citadas, embora, dependendo da amostra, possam apresentar algumas vantagens em relação às anteriores. [4]

[2] J. Berkson batizou as quantidades $log \left(\frac{p_j}{1 - p_j} \right)$ de *Logit* (p_j), razão pela qual também se encontra o estudo do modelo logístico sob o título de análise de *logits* (*logit analysis*).

[3] Uma discussão bem ampla, incluindo aplicações, pode ser encontrada em Finney (1952).

[4] Cox (1970) e Claringbold-Biggers-Emmens (1953).

2. Estimação dos parâmetros

2.1 Muitas observações por cela

2.1.1 Regressão simples

Seja, novamente, Y uma variável dicotômica e X um vetor de variáveis quaisquer. Digamos que se tenha uma amostra que permita o agrupamento de muitas observações por cela. Isto é, digamos que a amostra seja constituída por observações como abaixo:

$$(y_{ij_i},\ x_i) \qquad \begin{matrix} i = 1, \ldots, n \\ j_i = 1, \ldots, m_i \end{matrix}$$

Dado que cada cela x_i contém m_i observações, pode-se tomar para estimador das probabilidades de ocorrência do evento $(y_i = 1)$, dentro das diversas celas, a função

$$\hat{p}_i = \frac{1}{m_i} \sum_{j_i=1}^{m_i} y_{ij_i}$$

Assim, para um dado F ter-se-ia

$$\hat{P}_i \simeq F\,(x_j' \beta), \text{ e como consequência } F^{-1}\,(\hat{p}_i) = x_i'\ \beta + \varepsilon_i$$

No caso do modelo logístico, ter -se-ia

$$Logit\ (\hat{p}_i) = log\ \left(\frac{\hat{p}_i}{1 - \hat{p}}\right) = x_i'\ \beta + \varepsilon_i.$$

Portanto, tanto no caso geral como no modelo logístico, a estimação poderia ser feita através de uma regressão simples. Evidentemente, a qualidade dessa estimação estaria vinculada ao comportamento dos resíduos ε_i, e note-se que no caso extremo em que $m_i = 1$, então $\hat{p}_i = 1$ ou $\hat{p} = 0$, impossibilitando a formação dos $logit\ (\hat{p}_i) = log\ (\hat{p}_i/\mathrm{v} - \hat{p}_i)$, e quando m_i for muito pequeno tanto menor será a confiabilidade nos estimadores $\hat{p}$ quanto maior será a probabilidade de todas as observações na cela x_i serem todas iguais, ou seja, $\hat{p}_i = 1$ ou $\hat{p}_i = 0$. Assim, a utilização desta estratégia só seria aconselhável nos casos de amostras em que $m_i > 5$. Não que com isso seja eliminada a chance de se obterem estimações $\hat{p} = 1$ ou $\hat{p} = 0$, porém, esta seria diminuída e em se obtendo uma freqüência baixa

para esses valores extremos, não ficaria invalidada esta aplicação da regressão. No entanto, havendo uma percentagem significativamente alta de celas nas condições postas, o modelo de regressão deveria ser abandonado em favor de estimadores de máxima verossimilhança que apresentaremos mais adiante. Note, também, que as outras transformações compartilham estes mesmos problemas.

2.1.2 Regressão ponderada

Não é difícil imaginar situações em que o número de observações por cela, m_i, possa variar bastante. Na realidade, em se trabalhando com dados de natureza socioeconômica, o número de observações por cela depende exclusivamente da amostra, pois nesse caso não se aplica a idéia de controle de experimentos como se pode fazer em biologia, por exemplo. Assim, não havendo nenhuma regra que condicione o número de observações por cela seria mera coincidência, e uma informação preciosa sobre a população em questão, o fato de as quantidades m_i terem uma variação pequena. Donde se conclui que a estimação pelo critério de mínimos quadrados simples, diante da informação disponível, não é o método mais eficiente uma vez que nesse caso não se considera o número de observações por cela.

Joseph Berkson [5] sugeriu solução definida pela minimização do funcional $logit\ (\chi^2) = \sum_i m_i\ p_i\ (1 - p_i)\ (Logit\ (p_i) - x_j'\ \beta)^2$, denominada Logit ($\chi^2$) mínimo.

Esse estimador, sendo do tipo χ^2 mínimo, possui as mesmas propriedades assintóticas que os estimadores de máxima verossimilhança. Não obstante, a parte computacional de soluções *Logit* (χ^2) mínimo reduz-se à aplicação de uma rotina de mínimos quadrados ponderados, cada cela x_i tendo peso dado por $m_i\ p_i\ (1 - p_i)$. Portanto, sob esse ponto de vista, o estimador acima apresenta a vantagem de prescindir de rotinas iterativas na sua resolução, estas quase sempre onerosas por problemas de convergência.

Não se conclua com isso a existência de uma superioridade do *Logit* (χ^2) mínimo em relação ao critério de máxima verossimilhança, pois estamos pressupondo nessa comparação uma agregação dos dados para tornar possível a estimação das probabilidades dentro de cada cela. Na estimação

[5] Berkson, (1953).

por máxima verossimilhança não há necessidade de se agregarem os dados, processo esse que implica perda de informação a que corresponde uma perda na eficiência da estimação.

A sugestão de Berkson [6] para a análise do Logit é extensiva às outras transformações citadas, pois o princípio geral subjacente à escolha das ponderações do funcional *Logit* (χ^2) baseia-se no estudo da variância assintótica das probabilidades de ocorrência dentro de cada cela.

Considerando o modelo $p_j = F\ (x_j' \beta)$ com m_j observações na cela j, então para m_j suficientemente grande e p_j fora das caudas da distribuição F, $z_j = F^{-1}\ (\hat{p}_j)$ tem distribuição aproximadamente normal, e a média e a variância de z_j convergem para $F^{-1}\ (p_j)$ e

$$[(F^{-1})'\ (p_j)]^2 \cdot p_j\,(1 - p_j)/m_j \text{ [7]}$$

E mais, obtém-se um estimador consistente dessa variância ao se substituir p_j por $\hat{p}_j$, qual seja, fazendo-se

$$Var\ (z_j) = [(F^{-1})'\ (\hat{p}_j)]^2\ \hat{p}_j\,(1 - \hat{p})/m_j$$

Como já era esperado, quando $F\ (t) = \dfrac{1}{1 + e^{-t}}$, é imediato ver que $Var\ (z_j) = \dfrac{1}{m^t\ p_j\ \ (1 - p_j)}$, isto é, os pesos dos quadrados no funcional *Logit* (χ^2) são exatamente $1/Var\ (z_j)$.

2.2 Poucas observações por cela

2.2.1 Máxima verossimilhança

Na impossibilidade de se estimar ou sendo baixo o grau de confiança ao se estimarem as probabilidades dentro de cada cela, tem-se a opção de adotar estimadores de máxima verossimilhança.

Novamente, consideremos uma amostra constituída pelas observações (y_{ij_i}, x_i), $1 \leq j_i \leq m_i$, $1 \leq i \leq n$, e o modelo geral representado por

$$P\,(Y = 1 \mid X) = F\,(X'\,\beta)$$

[6] Berkson. op. cit.

[7] Cox (1970).

A função de verossimilhança para este modelo, que denotaremos por $L_\beta\ (y_{11}, y_{12} \ldots, y_n \mid x_1 \ldots x_n)$ *com base na amostra considerada é por* definição.

$$L_\beta\ (y_{11}, y_{12} \ldots y_{nm} \mid x_1 \ldots x_n) = \prod_{i=1}^{n} \prod_{j_i=1}^{m_i} [F\ (x_i' \beta)]^y ij_i\ [1 - F\ (x_i' \beta)]^{1-y} ij_i$$

Note que agora o número de observações por cela torna-se transparente, mesmo no caso extremo em que $m_j = 1$ para todas as celas, pois fazendo-se $y_{ij_i} = y_i$ obtém-se

$$L_\beta\ (y_1 \ldots y_n \mid x_1 \ldots x_n) = \prod_{i=1}^{n} [F\ (x_i' \beta)]^y\ i\ [1 - F\ (x_i' \beta)]^{1-y}\ i.$$

Pela expressão da função de verossimilhança, não se perde em generalidade adotando-se esta notação para o caso geral, ou seja, m_i qualquer, pois basta que se enumere a amostra sem grupar as observações de uma mesma cela.

Examinemos alguns aspectos da estimação por máxima verossimilhança do modelo logístico. A função de verossimilhança, para este modelo, toma a seguinte forma:

$$L_\beta\ (y_1 \ldots y_n \mid x_1 \ldots x_n) = \prod_{i=1}^{n} \left[\frac{1}{1 + e^{-x_i' \beta}}\right]^{y_i} \left[\frac{1}{1 + e^{x_i' \beta}}\right]^{1-y_i} =$$

$$= \prod_{i=1}^{n} \left[\frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}}\right]^{y_i} \left[\frac{1}{1 + e^{x_i' \beta}}\right]^{1-y_i} = \frac{e^{\sum_{i=0}^{n} \beta_i t_j}}{\prod_{i=1}^{n} (1 + e^{x_i' \beta})} = \frac{e^{\beta t}}{\prod_{i=1}^{n} (1 + e^{x_i' \beta})}$$

onde $t_j = \sum_{i=1}^{n} x_i^j\ y_i$, isto é, soma dos valores da j-ésima componente dos vetores x_i nas observações em que $y = 1$; e k representa o número de variáveis no vetor X.

Pelo teorema da fatorização (Neyman-Pearson), o vetor $t = (t_0, t_1 \ldots t_k)$ é uma estatística suficiente para β, dados $x_1 \ldots x_n$, uma vez que $L\beta\ (y_1 \ldots y_n \mid x_1 \ldots x_n)$ só depende da variável Y por intermédio da função $\beta.t$.

Uma outra propriedade importante da função de verossimilhança $L\beta$ para o modelo logístico é dada pelo fato de *Log* $L\beta$ ser côncava. Assim, a

escolha da rotina de maximização, a ser utilizada na estimação, pode ser mais bem especificada propiciando, em conseqüência, um ganho em eficiência. Senão vejamos,

$$Log\ L_\beta = \sum_{j=0}^{n} \beta_j\ t_j - \sum_{i=1}^{n} log\ (1 + e^{x_i'\beta})$$

$$\frac{\delta\ Log_\beta}{\delta \beta_j} = t_j - \sum_{i=1}^{n} \frac{x_j^i \cdot e^{x_i'\beta}}{1 + e^{x_i'\beta}}$$

$$\frac{\delta^2\ Log\ L_\beta}{\delta\beta_{j1}\ \delta\beta_{j2}} = - \sum_{i=1}^{n} \frac{e^{x_i'\beta}\ x_i^{j1}\ x_i^{j2}}{(1 + e^{x_i'\beta})^2} = - \sum_{i=1}^{n} K\ (i)\ x_i^{j\,1} \cdot x_i^{j\,2}$$

Como $K\ (i) > 0$, fica demonstrada a forma côncava da função $L\beta$. [8]

Por definição, o estimador de máxima verossimilhança, $\hat{\beta}$, satisfaz as identidades.

$$\left[\frac{\delta\ Log\ L_\beta}{\delta\ \beta_j}\right]_{\beta = \hat{\beta}} = 0,\ j = 0,\ \ldots,\ k$$

que fornecem as seguintes equações normais para este estimador

$$\sum_{i=1}^{n} \frac{1 + e^{x_i'\beta}}{e^{x_i'\beta}}\ x_i^j = t,\ j = 0,\ \ldots,\ k.$$

Sendo L côncava, a raiz dessas equações é um maximando para L, e, portanto, sua resolução apresenta uma nova alternativa computacional para obtenção do estimador β.

2.3 Pequena apologia da transformação logística

Todas as transformações aqui citadas têm um comportamento bem semelhante fora das caudas das respectivas distribuições, isto é, somente quando os dados apresentarem as probabilidades das celas muito concentradas em intervalos próximos de um ou zero podem-se esperar discrepâncias nas estimações. Assim, na ausência de tais concentrações, a transformação logística facilita a parte computacional por ser a de expressão algébrica mais simples. Por outro lado, teoricamente dever-se-ia escolher a transformação em função das condições particulares da aplicação que se queira fazer. Porém, pelas vagas referências encontradas nos diversos textos sobre o assunto, parece ainda não haver evidência empírica suficiente no sentido de apoiar uma escolha deste tipo.

[8] Nerlove-Press (1973).

Um outro atrativo da transformação logística decorre da existência de estatísticas suficientes para os parâmetros do modelo logístico, colocando, portanto, um possível usuário em posição mais confortável com relação às questões de eficiência.

Por último, apresentaremos uma justificativa teórica para o modelo logístico demonstrando como este aparece naturalmente, no contexto em que estamos examinando, quando se adotam hipóteses adicionais sobre o comportamento do vetor X.

Seja Y uma variável dicotômica tomando os valores um ou zero dependendo se o evento E ocorre ou não, e X um vetor de variáveis com uma função densidade contínua dada por $h\ (X)$. Pelo teorema de Bayes, podemos escrever:

$$P\,(Y = 1 \mid X) = \frac{p \cdot h\,(X \mid Y = 1)}{p \cdot h\,(X \mid Y = 1) + (1 - p)\, h\,(X \mid Y = 0)} =$$

$$= \frac{1}{1 + \dfrac{1 - p}{p}\ \dfrac{h\,(X \mid Y = 0)}{h\,(X \mid Y = 1)}},$$

onde $p = P(Y = 1)$.

Admitindo-se que a distribuição do vetor X dado que $Y = 1$ seja normal com vetor de médias M_1 e matriz de covariância Σ, fato que denotaremos por $L\ (X \mid Y = 1) \sim N\ (M_1, \Sigma)$ e que $L\ (X \mid Y = 0 \sim N\ (M_2, \Sigma)$, então é imediato que $P\ (Y = 1 \mid X)$ pode ser colocada na forma da função logística (os termos quadráticos anulam-se uma vez que as matrizes de covariância são idênticas). [9]

3. Aspectos empíricos

3.1 Dados agregados

Várias modificações da transformação logística podem ser propostas no sentido de tornar possível a manipulação de dados nos casos em que todas as observações dentro de uma mesma cela são todas iguais. Vamos examinar sucintamente algumas possibilidades deixando os detalhes para serem consultados em Cox (1970).

[9] Nerlove-Press (1973) e Berkson (1951).

Considerando a cela x_j, vamos chamar de r_j o número de vezes, na amostra, que $y_{ij} = 1$, isto é, $r_j = \Sigma\, y_{ij}$ e n_j o número total de observações nesta cela. Então, fazendo-se $\hat{p}_j = r_j/n_j$,

$$z_j = Logit\ (\hat{p}_j) = log\left(\frac{r_j}{n_j - r_j}\right)$$

A tentativa mais simples de se modificar z_j seria através de uma translação dos operandos dentro da expressão dessa variável. Neste caso

$$z_j\ (a) = log\left(\frac{r_j + a}{n_j - r_j + a}\right)$$

Admitindo-se algumas hipóteses adicionais sobre a distribuição de r_j, pode-se deduzir que

$$E\,[z_j\ (a)] - Logit\ p_j = \frac{(1 - 2\,p_j)\ (a - 1/2)}{p_j\ (1 - p_j)\ n_j} + 0\ (1/n_j)$$

onde $0\ (1/n_j)$ representa um termo de ordem inferior a $1/n_j$ que, portanto, será desprezado.

Este valor aproximado para a tendenciosidade de $z_j\ (a)$ como estimador de $Logit\ (p_j)$, sugere a escolha de $a = 1/2$, pois esta forneceria um estimador praticamente não-tendencioso, ou o menos tendencioso, ao nível de amostra de tamanho n_j. Portanto, ficamos com $z_j = log\ \dfrac{(r_j + 1/2)}{(n_j - r_j + 1/2)}$ e definindo-se $v_j = \dfrac{(n_j + 1)\ (n_j + 2)}{n_j\ (r_j + 1)\ (n_j - r_j + 1)}$ tem-se que v_j é um estimador aproximadamente não-tendencioso de $Var\ (z_j)$, que são os elementos necessários à análise ponderada já considerada no item 2.

Para um procedimento alternativo, tomando-se $a = -\ 1/2$, isto é, $z_j = log\ \dfrac{(r_j - 1/2)}{(n_j - r_j - 1/2)}$ e fazendo-se $v_j = \dfrac{(n_j - 1)}{r_j\ (n_j - r_j)}$, então, para $n_j > 1$, tem-se que

$$E\left[\frac{r_j\ (n_j - r_j)}{n_j - 1}\right] = n_j \cdot p_j \cdot (1 - p_j) \quad \text{e}$$

$$E\left[\left(\frac{r_j\ (n_j - r)_j}{n_j - 1}\right)\ log\ \frac{(r_j - 1/2)}{(n_j - r_j - 1/2)}\right] \simeq n_j\ p_j\ (1 - p_j)\ logit\ (p_j).$$

aproximadamente. Assim, nesse caso, as celas com $r_j = 0$ ou $r_j = n_j$ são eliminadas pela ponderação escolhida, e mais,

$$\frac{r_j (n_j - r_j)}{n_j - 1} \cdot log \left(\frac{(r_j - 1/2)}{n_j - r_j - 1/2} \right)$$

constitui-se num estimador quase não-tendensioso dos termos do funcional *Logit* (χ^2).

3.2 Dados desagregados

Independentemente das características da amostra, já se constatou, no item 2, não haver dificuldade numérica na montagem da função de verossimilhança. Assim, o problema que se apresenta na estimação por máxima verossimilhança é meramente computacional uma vez que se trata de uma maximização de um funcional relativamente complicado que, portanto, exige uma disponibilidade de rotinas iterativas eficientes. Não obstante, tem-se meio caminho andado caso se disponha de boas soluções iniciais para estas rotinas iterativas. Enumeraremos, a seguir, algumas indicações no sentido de se obterem soluções iniciais razoáveis.

Uma primeira alternativa baseia-se na estimação por mínimos quadrados. Para isso, é bastante que se agrupem os dados por celas criadas, mesmo que, arbitrariamente, se utilize a estimação de vetor β, obtido através de uma rotina de regressão simples ou ponderada, como solução inicial.

Outras alternativas podem ser estudadas pelo exame direto da função de verossimilhança, tentando sempre simplificar sua expressão com base nas características específicas dos dados em questão. Vejamos, por exemplo, duas situações simplificadas, deixando os detalhes para uma eventual consulta em Cox. [10]

Suponhamos que:

a) todas as probabilidades de sucesso são pequenas, ou seja, as quantidades $\frac{1}{1 + e^{-x_i' \beta}}$ são pequenas, e

b) as variações nos termos $X' \beta$ são pequenas.

[10] Cox (1970).

Antes de efetuar as simplificações que podem ser obtidas das condições postas, vamos definir uma notação matricial que nos facilitará a manipulação algébrica.

Seja $Z = (z_{ij})$ a matriz de elementos $z_{ji} = x_i - x_j$, onde x_j representa a média amostral da variável x^j, ou seja, $x_j = \sum_{i=1}^{n} x_j^i$. Da mesma forma definiremos a matriz $X = (x_{ij})$, onde $x_{ji} = x_i^j$ e z_i' será a i-ésima linha da matriz transposta de Z.

Então, separando a interseção ou fator constante $\beta 0$ dos coeficientes das variáveis explicativas na expressão de $Log\ L_\beta$ e utilizando as hipóteses (a) e (b) podemos escrever:

$$log\ L_\beta = \beta 0 \sum_{i=1}^{n} y_i + \beta'\ ZY - \sum_{i=1}^{n} log\ (1 + e^{\beta 0 + z_i'\ \beta})$$

$$\simeq \beta 0 \sum_{i=1}^{n} y_i + \beta'\ ZY - e^{\beta 0} \sum_{i=1}^{n} e^{z_i'\ \beta}$$

$$\simeq \beta 0 \sum_{i=1}^{n} y_i + \beta'\ ZY - e^{\beta 0}\ [n + (1/2)\ \beta'\ ZZ'\ \beta]$$

Tomando-se as derivadas em relação a $\beta 0$ e β, obtêm-se as seguintes equações normais de fácil resolução à semelhança das equações correspondentes para a regressão simples:

$$e^{\hat{\beta}0} = \sum_{i=1}^{n} y_i$$

e

$$\hat{\beta} = e^{-\hat{\beta}0}\ (ZZ')^{-1}\ ZY.$$

Um segundo caso em que a função de verossimilhança pode ser bastante simplificada é quando as probabilidades ajustadas no modelo estão fora das caudas, ou seja, não estão próximas de um ou zero. Assim sendo, pode-se aproximar $F\ (t)$ dentro de um intervalo prefixado por, digamos, um pedaço de sua expansão em série de Taylor, como por exemplo:

$$F\ (t) = \begin{cases} 1 & \text{se } t > 3 \\ 1/2 + 1/6t & \text{se } |t| \leq 3 \\ 0 & \text{se } t < -3 \end{cases}$$

Então, se $| x_i \beta | \leqslant 3$ para todas as observações, é fácil notar-se que a nova função de verossimilhança para a amostra em questão será

$$L_\beta = \prod_{i=1}^{n} \left(\frac{1}{2} + \frac{1}{6} x_i' \beta\right)^{y_i} \left(\frac{1}{2} - \frac{1}{6} x_i' \beta\right)^{1-y_i} \quad \text{e}$$

$$Log\ L_\beta = \Sigma\, y_i \cdot log \left(\frac{1}{2} + \frac{1}{6} x_i' \beta\right) + \Sigma\,(1 - y_i) \cdot log \left(\frac{1}{2} - \frac{1}{6} x_i' \beta\right)$$

Tomando-se as derivadas parciais e igualando-se a zero, obtém-se que

$$X' \left(Y - \frac{1}{2}\right) - \frac{1}{6} XX' \hat{\beta} = 0 \quad \text{e}$$

$$\hat{\beta} = 6\,(XX')^{-1} X \left(Y - \frac{1}{2}\right)$$

Em conseqüência da fórmula para a solução β, a sua obtenção torna-se uma simples aplicação de uma rotina de regressão, depois de se ter devidamente recodificado os dados primários.

Um outro método de obtenção de uma solução inicial, denominado pelo Prof. M. Nerlove *reverse Taylor series,* oferece, segundo evidência empírica encontrada pelo próprio, resultados surpreendentemente próximos da solução final para o estimador de máxima verossimilhança, especialmente no caso complementar ao dos alternativos acima, ou seja, quando os dados apresentam-se bastante espalhados em todo seu intervalo de variação. [11]

Este método consiste em interpretar a regressão ordinária sobre os dados desagregados como uma aproximação linear por séries de Taylor da transformação logística; desta forma, invertendo o processo obtêm-se os coeficientes desejados como demonstraremos adiante.

Consideremos inicialmente a expansão em séries de Taylor da transformação logística

$$F(x) = F(x_1, \ldots, x_k) = F(\bar{x}) + \sum_{j=1}^{k} \frac{\delta}{\delta x_j} F(\bar{x}) \cdot (x_j - \bar{x}_j) + R(x)$$

$$= \frac{1}{1 + e^{-(\beta_0 + \bar{x}' \beta)}} - \frac{(\bar{x}' \beta) \cdot e^{-(\beta_0 + \bar{x}' \beta)}}{[1 + e^{-(\beta_0 + \bar{x}' \beta)}]^2} + \frac{(x' \beta) \cdot e^{-(\beta_0 + \bar{x}' \beta)}}{[1 + e^{-(\beta_0 + \bar{x}' \beta)}]^2} + R(x)$$

[11] Nerlove-Press (1973).

Assim, estimando-se $F(x)$ por uma função linear, digamos $F(x) = a + x' b$, podemos interpretar a e b como

$$a = \frac{1}{1 + e^{-(\beta_0 + \bar{x}' \beta)}} - \bar{x}' b$$

e

$$b_i = \frac{\beta_i e^{-(\beta_0 + \bar{x}' \beta)}}{[1 + e^{-(\beta_0 + \bar{x}' \beta)}]^2}, \qquad i = 1, \ldots k$$

E, resolvendo este sistema de equações, obtém-se

$$\beta_i = \frac{b_i}{(a + \bar{x}' b)(1 - a - \bar{x}' b)}$$

$$\beta_0 = - \bar{x}' b - log \left(\frac{1}{a + \bar{x}' b} - 1 \right).$$

4. Uma aplicação ao estudo das migrações

Esta aplicação faz parte de um documento preparado dentro do programa Fundação Instituto Brasileiro de Geografia e Estatística/Organização Internacional do Trabalho, em andamento, cujo objetivo é construir um modelo de simulação socioeconômico para o Brasil. Os responsáveis pelo projeto, visando um modelo articulado em um grande número de variáveis endógenas, interessaram-se em obter uma função de migrações rural-urbana que fosse sensível a modificações nos níveis de algumas dessas variáveis. No caso, as variáveis escolhidas foram, por ordem de importância segundo indicadores estatísticos, diferencial de renda, nível educacional, número de filhos e idade. Faremos a seguir uma pequena apresentação da amostra estudada e das estimações através das diferentes técnicas utilizadas.

Todas as decisões tomadas na especificação do universo sobre o qual atuar tiveram sempre a finalidade de tornar o estudo o mais significativo possível. Assim, filtramos inicialmente as observações no sentido de nos restringirmos aos indivíduos entre 20 e 50 anos de idade, na condição de chefe de família com tempo de migração inferior a três anos. A unidade adotada, ou seja, a família representada pelo chefe, tem fundamento na hipótese de o centro de decisões familiares, assim como seu próprio

perfil socioeconômico, estarem determinados na pessoa do chefe. O filtro de idade incluído mantém no universo a grande maioria das famílias migrantes e tem a propriedade de eliminar os casos atípicos dentro do contexto que estamos analisando. E mais, uma vez resolvido ter o diferencial de renda como variável explicativa do fenômeno de migrações, só nos interessava considerar chefes de família economicamente ativos. A restrição sobre o tempo de migração foi imposta com o objetivo de limitar a descaracterização do migrante, com o correr do tempo, devido ao contacto com facilidades de, por exemplo, elevar seu nível educacional dentro dos grandes centros urbanos. Especificado o universo, passemos à descrição da amostra propriamente dita.

Como os indivíduos que migram para regiões metropolitanas apresentam um perfil bem amplo e como estas são as que mais recebem migrantes rurais, escolhemos a amostra com base nas principais regiões metropolitanas do País e nas regiões rurais que mais contribuem com migrantes para essas regiões. Apresentaremos na tabela 1 os números obtidos sem deixar de lembrar que estes representam contagens de chefes de famílias dentro das especificações citadas.

Tabela 1

Tamanho das amostras utilizadas segundo origem e destino dos migrantes *

Origem → / Destino ↓	Paraíba	Pernambuco	Bahia	Minas Gerais	São Paulo	Paraná	Santa Catarina	Rio Grande do Sul	Goiás	Total
Recife	228	2.366	—	—	—	—	—	—	—	2.594
Rio de Janeiro	1.684	500	282	1.531	—	—	—	—	—	3.997
Salvador	—	—	1.074	—	—	—	—	—	—	1.074
Belo Horizonte	—	—	60	3.095	—	—	—	—	—	3.155
São Paulo	618	2.384	3.988	4.939	11.185	3.575	104	47	—	26.840
Florianópolis	—	—	—	—	—	—	354	—	—	354
Curitiba	—	—	—	—	—	1.378	—	—	—	1.378
Porto Alegre	—	—	—	—	—	—	—	4.092	—	4.092
Goiânia	—	—	—	—	—	—	—	—	1.230	1.230
Total	2.530	5.250	5.404	9.565	11.185	4.953	458	4.139	1.230	44.714

População Rural = 1.361.265

* Censo demográfico de 1970.

Foram criadas categorias para as diversas variáveis incluídas no tema, não só com o objetivo de obter tabelas de freqüência e cruzamento mas também para tornar viável a utilização das técnicas de estimação com dados agregados. As tabelas de freqüência e cruzamentos referidos foram elaboradas no sentido de se obter uma caracterização do perfil do migrante em cotejo com o do não-migrante e os resultados obtidos já seriam suficientes para definir os sinais encontrados para os parâmetros das variáveis na função de probabilidades. Deixaremos de transcrever esses dados para não fugir aos propósitos de uma ilustração das técnicas de análise que foram apresentadas no artigo, não obstante comentarmos a construção da variável diferencial de renda uma vez que todas as outras foram obtidas através de recodificações simples dos dados do censo demográfico (1970).

O ideal para obtenção da variável diferencial de renda seria fazê-la controlada pelo maior número de variáveis pertinentes ao tema ou que exerçam influência na renda do indivíduo, desde que os dados oferecessem riqueza suficiente para permitir tal procedimento. Assim, podemos citar classes de ocupações, nível educacional e idade como os exemplos mais conspícuos de variáveis de controle. A primeira delas foi descartada em função da acentuada transferência de setores e ocupações acarretada pelo fenômeno das migrações, isto é, um migrante com determinada ocupação dentro de um setor dificilmente já se encontrava com essa classificação na origem rural. A variável idade na presença do nível educacional demonstrou perder bastante poder explicativo sobre a renda individual, talvez mais em função dos dados que apresentam concentração nas primeiras classes de renda e níveis educacionais do que como regra geral. Desta forma, restringimo-nos ao nível educacional como fator de controle na montagem do diferencial de renda que foi realizada como se segue. Tomamos como base a renda média dentro de cada um dos nove níveis educacionais, descritos adiante, e definimos o diferencial de renda como a diferença entre renda familiar anual e a renda familiar anual média (rural) obtida para o nível educacional a que pertence o chefe da família migrante.

Os níveis educacionais foram os seguintes:

1 — zero anos de estudo
2 — 1 ou 2 anos de estudo
3 — 3 ou 4 anos de estudo (primário completo)
4 — 5 ou 6 anos de estudo
5 — 7 ou 8 anos de estudo (ginasial completo)

6 — 9 ou 10 anos de estudo
7 — 11 ou 12 anos de estudo (colegial completo)
8 — 13, 14 ou 15 anos de estudo (1.º, 2.º ou 3.º ano universitário)
9 — 16, 17 ou 18 anos de estudo (4.º, 5.º ou 6.º ano universitário)

Foram obtidos os seguintes valores para renda familiar média dentro de cada um dos nove níveis educacionais nas regiões rurais e para os migrantes (rural-urbano):

Tabela 2

Renda familiar anual média por nível de educação para não-migrantes e para migrantes (rural-urbano) *

Não-Migrante		Migrante	
Nível de educação	Renda média anual	Nível de educação	Renda média anual
1	1.551,74	1	3.436,40
2	1.969,56	2	3.807,04
3	2.638,58	3	4.559,64
4	3.241,39	4	5.135,17
5	6.493,91	5	8.154,34
6	7.853,26	6	9.699,25
7	9.773,59	7	14.115,69
8	20.748,14	8	13.422,71
9	22.291,42	9	26.635,29

* Censo demográfico de 1970.

Passemos à análise das estimações e estatísticas correspondentes, lembrando que os resultados apresentados na tabela 3 dizem respeito à estimação dos coeficientes β de uma função $P\ (y = 1 \mid X = x) = F\ (x'\ \beta)$, onde $P\ (y = 1 \mid X = x)$ é a probabilidade de um indivíduo da região rural migrar, dado que seu perfil, segundo o vetor das variáveis X, é especificado pelo vetor de códigos x, e $F\ (X'\ \beta)$ é a sua representação através dos modelos propostos para a análise de variáveis binárias.

A organização da tabela 3 foi feita de modo a separar por grupos de duas colunas as estimações referentes a cada transformação aplicada aos percentuais por celas através das técnicas de regressão simples e ponderada sobre os dados tarnsformados. Em vista do grande volume de dados

Tabela 3

Transformação (variável dependente)	$Y = log \frac{R + 1/2}{N - R + 1/2}$		$Y = log \frac{R + 1/2}{N - R + 1/2}$		$Y = log \frac{R - 1/2}{N - R - 1/2}$		$Y = log \frac{R - 1/2}{N - R - 1/2}$		$Y = arcsen \sqrt{R/N} - \Pi/4$		$Y = arcsen \sqrt{R/N} - \Pi/4$	
Estimação	Regressão ponderada		Regressão simples		Regressão ponderada		Regressão simples		Regressão ponderada		Regressão simples	
Variáveis explicativas	Coef. reg.	Est. t	Coef. reg.	Est. t	Coef. reg.	Est. t	Coef. reg.	Est. t	Coef. reg.	Est. t	Coef. reg.	Est. t
Diferencial renda	0,911	48,61	0,725	21,46	0,915	41,25	0,753	21,91	0,089	41,19	0,067	10,18
Nível educacional	1,143	32,30	1,125	23,32	1,157	27,30	1,293	27,24	0,122	26,56	0,079	8,41
Número de filhos	—0,284	23,17	0,204	9,07	—0,285	19,62	—0,209	9,64	—0,021	17,93	—0,040	9,09
Idade	—0,210	73,79	—0,207	6,93	—0,210	11,64	—0,167	6,06	—0,011	7,01	—0,024	4,10
Interseção	— 6,230		— 6,310		— 6,532		— 6,606		— 0,934		— 0,698	
R^2	0,924		0,614		0,904		0,739		0,985		0,287	
F	2.352		304		1.231		367		13.100		77	

disponíveis, pudemos prescindir das técnicas de máxima verossimilhança; as estimações das probabilidades, $\hat{p}$, dentro de cada cela tendo sido realizadas com uma proporção aceitável de casos degenerados, ou seja, casos em que $\hat{p} = 1$ ou $\hat{p} = 0$. Depois de agregados os dados, obtivemos um total de celas da ordem de 800 com aproximadamente 200 degeneradas.

Cabe ressaltar que a construção de classes para as variáveis independentes, possibilitando a criação de celas e a conseqüente estimação das probabilidades, não seguiu um critério muito rigoroso; no caso, essas classes visavam atender ao nível de agregação do modelo de simulação mencionado. Portanto, sob esse ponto de vista, os resultados deveriam ser entendidos de uma maneira mais qualitativa, isto é, examinando-se o tipo de influência de cada variável através do sinal dos coeficientes sem desviar muita atenção para seus valores.

Ressaltamos também o fato de não ter sido efetuado nenhum teste de aderência, porém, a variância explicada por todos os modelos, com exceção do obtido através da transformação $Y = arcsen\ (\sqrt{R}/N - \pi/4)$ que comentaremos mais adiante, apresenta-se suficientemente alta para transmitir confiabilidade pelo menos na representatividade desses modelos. De resto, podemos obter somente uma ordenação das variáveis segundo a importância medida pelo seu maior afastamento da hipótese do coeficiente correspondente ser zero, pois com graus de liberdade da ordem de 800 unidades, os coeficientes dificilmente deixariam de ser significativos quando testados pelo valor da estatística *t*.

Nesta aplicação utilizamos apenas as transformações logísticas modificadas e a transformação angular citada, e pode-se notar que para as primeiras os resultados apresentam uma certa uniformidade no que diz respeito aos valores dos coeficientes, o mesmo não acontecendo com a última, que forneceu valores totalmente instáveis ao se variar a técnica de estimação da regressão simples para regressão ponderada.

Os valores dos coeficientes para as transfarmações logísticas encontram-se entre 0,15 e 1,3 e para a transformação angular *arcsen*, entre 0,01 e 0,12. Esta disparidade na ordem de grandeza dos parâmetros já deveria ser esperada, pois as duas distribuições correspondentes não foram calibradas no sentido de parametrizá-las de forma a coincidirem em determinado valor de probabilidade.

Vamos nos deter um pouco na análise do comportamento anômalo da transformação angular.

Esta transformação tem a característica de que para valores de probabilidade não muito concentrados em vizinhanças dos pontos zero e um ela possui um efeito estabilizador de variâncias, quando o número de observações por cela é significativo e aproximadamente constante, podendo mesmo dispensar rotinas de regressão ponderada. Isto devido ao fato de o estimador *arcsen* $\hat{p}$ ter variância assintótica igual a $4N$, onde N é o número de observações. Porém, na aplicação que estamos examinando ocorre exatamente o oposto, ou seja, não só os valores de probabilidades dentro das celas estão concentrados em torno dos valores zero e um, principalmente o valor zero, mas também o número de observações por cela é bastante variável. Desta forma, como a transformação *arcsen* $(\sqrt{\hat{p}} - \pi/4)$ tem seu conjunto de valores truncado no intervalo $[-\pi/4, \pi/4]$, uma massa muito grande de observações restou concentrada em pequenas vizinhanças dos pontos $-\pi/4$ e $\pi/4$. Conseqüentemente, obtivemos um comportamento bastante instável no que diz respeito à variância explicada (R^2) que no caso da regressão simples foi de apenas 28,4% enquanto que na regressão ponderada foi surpreendentemente alta, 98,5%.

Os resultados das estimações utilizando as outras transformações são bastante compatíveis, o que em caso contrário colocaria todo o procedimento em dúvida de vez que se trata de duas modificações da transformação logística com o objetivo de lidar com observações degeneradas. Todavia, daremos sempre preferência à transformação $Y = log \frac{R + 1/2}{N - R + 1/2}$, pois acreditamos ser esta a mais eficiente, uma vez que permite a utilização de todas as observações na estimação dos parâmetros.

Chamaremos a atenção do leitor para dois aspectos interessantes surgidos nas estimações com as transformações logísticas. O primeiro, já mencionado quando do exame da transformação angular, diz respeito ao aumento sistemático da variância explicada (R^2) quando se passa da regressão simples para a regressão ponderada, e o segundo à inversão na ordem de influência das variáveis diferencial de renda e nível educacional segundo os valores da estatística t para os respectivos coeficientes na regressão simples e ponderada. Esse último fenômeno parece-nos uma coincidência numérica, enquanto que o primeiro poderia ser explicado em função da expansão da nuvem de pontos decorrentes da multiplicação pelo inverso dos desvios-padrão.

Bibliografia útil ao estudo de variáveis categóricas

Abramowitz, M., & Stegun, I. A. *Handbook of mathematical functions.* New York, Dover, 1965. App. A.

Aitchison, J., & Silvey, S. D. The generalization of probit analysis to the case of multiple responses. *Biometrika,* v. 44, p. 131-40, 1957. § 7.4.

Anscombe, F. J. On estimating binomial response relations. *Biometrika,* v. 43, p. 461-64, 1956. § 3.2, App. A.

Anscombe, F. J. Examination of residuals. *Proc. 4th Berkley Symp.* 1961. v. 1, p. 1-36. § 6.6.

Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics,* v. 11, p. 375-86. 1955. § 1.2.

Ashford, J. R. An approach to the analysis of data for semi-quantal responses in biological response. *Biometrics,* v. 15, p. 573-81, 1959. § 7.4.

Bartlett, M. S. Contingency table iteractions. *J. R. Statist. Soc.,* Suppl., v. 2, p. 248-52, 1935. App. A, b.

Berkson, J. Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.,* v. 39, p. 357-65, 1944. App. B.

————. Why I prefer logits to probits. *Biometrics,* v. 7, p. 327-39, 1951. App. B.

————. A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *J. Amer. Statist. Assoc.,* v. 48, p. 565-99, 1953. §§ 3.1, 6.3, App. B.

————. Maximum likelihood and minimum χ^2 estimates of the logistic function. *J. Amer. Statist. Assoc.,* v. 50, p. 130-62, 1955. § 3.5, App. B.

————. Estimation by least squeres and by maximum likelihood. *Proc. 3rd Berkley Symp.* 1955. v. 1, p. 1-11. App. A, B.

————. Tables for the maximum likelihood estimate of the logistic function. *Biometrics,* v. 13, p. 28-34, 1957. App. B.

————. Nomograms for fitting the logistic function by maximum likelihood. *Biometrika,* v. 47, p. 121-41, 1960. App. B.

———. Application of minimum logit χ^2 to a problem of Grizzle with a notation on the problem of no interaction. *Biometrics,* v. 24, p. 75-95, 1968. App. B.

Birch, M. W. Maximum likelihood in three-way contingency tables. *J. R. Statist. Soc.,* v. B 25, p. 220-33, 1963.

———. The detection of partial association, I: the 2x2 case. *J. R. Statist. Soc.,* v. B 26, p. 313-24, 1964.

———. The detection of partial association, II: the general case. *J. R. Statist. Soc.,* v. B 27, p. 111-24, 1965.

Blom, G. Transformations of the binominal, negative binominal, Poisson and χ^2 distributions. *Biometrika,* v. 41, p. 302-16, 1954. App. A.

Chernoff, H. On the distribution of the likelihood ratio. *Ann. Math. Statist.,* v. 25, p. 573-78, 1954. App. A.

Claringbold, P. J., Biggers, J. D., & Emmens, G. W. The angular transformation in quantal analysis. *Biometrics,* v. 9, p. 467-84, 1953. § 2.7.

Cox, D. R. The regression analysis of binary sequences (with discussion). *J. R. Statist. Soc.,* v. B 20, p. 215-42, 1958. App. A, B.

———. Two further applications of a model for binary regression. *Biometrika,* v. 45, p. 562-65, 1958. App. A, B.

———. Large sample sequencial tests of composite hypoteses. *Sankhyā,* v. A 25, p. 5-12, 1963. App. A.

———. A simple example of a comparison involving quantal data. *Biometrika,* v. 53, p. 215-20, 1966. § 5.3, App. B.

———. Some procedures connected with the logistc qualitative response curve. *Research papers in statistics: essays honour of J. Neyman's 70th birthday.* In: F. N. David, ed., London, Wiley, 1966. p. 55-71. App. A, B.

———. The analysis of binary data. *Methuen's monographs on applied prosability and statistics.* 1960.

Cox, D. R., & Lauh, E. A note on the graphical analysis of multi-dimensional contingency tables. *Technometrics,* v. 9, 481-488, 1967. § 3.4.

Cox, D. R., & Snell, E. J. A general definition of residuals (with discussion). *J. R. Statist. Soc.,* v. B 30, p. 248-75, 1968. § 6.6, App. A.

Cramer, E. M. Some comparisons of methods of fitting the dosage response curve for small samples. *J. Amer. Statist. Assoc.*, v. 59, p. 779-93, 1964.

Darroch, J. N. Interactions in multi-factor contingency tables. *J. R. Statist. Soc.*, v. B 24, p. 251-63, 1962.

Draper, N. R., & Smith, H. *Applied regression analysis.* New York, Wiley, 1966. § 6.4.

Edwards, A. W. F. The measure of association in a 2x2 table. *J. R. Statist. Soc.*, v. A 126, p. 109-14, 1963. App. A.

Elashoff, J. D.; Elashoff, R. M. & Goldman, G. E. On the choice of variables in classification problems with dichotomous variables. *Biometrika*, v. 54, p. 668-70, 1967.

Feldstein, M. S. A binary variable multiple regression method of analysing factors affecting perinatal mortality and other outcomes of pregnancy. *J. R. Statist. Soc.*, v. A 129, p. 61-73, 1966. § 1.2.

Feller, W. *An introduction to probability and theory and its applications.* 3 ed. New York, Wiley, 1968. v. 1, § 4.3.

Finney, D. J. *Probit analysis.* Cambridge University Press, 2. ed., 1952. § 2.7, App. B.

————. *Statistical method in biological assay.* 2. ed., London, Griffin, 1964. App. B.

Fisher, R. A. The analysis of variance with various binomial transformations (with discussion). *Biometrics*, v. 10, p. 130-39, 1954.

Freeman, M. F., & Tukey, J. W. Transformations related to the angular and the square root. *Ann. Math. Statist.*, v. 21, p. 607-11, 1950. App. A.

Gabriel, K. R. Analysis of variance of proportions with unequal frequencies. *J. Amer. Statist. Assoc.*, v. 58, p. 1133-57, 1963.

Goodman, L. A. On Plackett's test for contingency table interations. *J. R. Statist. Soc.*, v. B 25, p. 179-88, 1963. § 7.5.

————. On methods for comparing contingency tables. *J. R. Statist. Soc.*, v. A 126, p. 94-108, 1963.

————. Simple methods of analyzing threefactor interation in contingency tables. *J. Amer Statist. Assoc.*, v. 59, 319-52, 1964.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. *J. Amer. Statist. Assoc.*, v. 49, p. 732-64, 1964. App. B.

————. Measures of association for cross classification. II. Further discussion and references. *J. Amer. Statist. Assoc.*, v. 54, p. 123-63, 1959. App. B.

————. Measures of association for cross classifications. III. Approximate sampling theory. *J. Amer. Statist. Assoc.*, v. 58, 310-64, 1963.

Grizzle, J. E. A new method of testing hypotheses and estimating parameters for the logistic model. *Biometrics,* v. 17, p. 372-85, 1961.

————. Asymptotic power of tests of linear hypotheses using the probit and logit transformations. *J. Amer. Statist. Assoc.*, v. 57, p. 877-94, 1962.

Gurland, J.; Lee, L., & Dolan, P. A. Polychotomous quantal response in biological assay. *Biometrics,* v. 16, p. 382-98, 1960. § 7.4.

Haldane, J. B. S. The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Hum. Genetics,* v. 20, p. 309-11, 1955. § 3.2.

Hewlett, P. S., & Plackett, R. L. A unified theory for quantal responses to mixtures of drugs: competitive action. *Biometrics,* v. 20, p. 566-75, 1964.

Hitchcock, S. E. A note on the estimation of the parameters of the logistic function, using the minimum logit χ^2 method. *Biometrika,* v. 49, p. 250-52, 1962.

————. Tests of hypotheses about the parameters of the logistic distribution. *Biometrika,* v. 53, p. 535-44, 1966. § 5.5, App. B.

Hodges, J. L. Fitting the logistic by maximum likelihood. *Biometrics,* v. 14, p. 453-61, 1958. App. A.

Kastenbaum, M. A., & Lamphiear, D. E. Calculation of chi-square to test the no three-factor interation hypothesis. *Biometrics,* v. 15, p. 107-15, 1959.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics.* 2. ed. London, Griffin, 1963. v. 1, § 5.4.

————. *The advanced theory of statistics.* 2. ed. London, Griffin, 1967. v. 2, § 5.4.

Lancaster, H. O. Complex contingency tables treated by partition of χ^2. *J. R. Statist. Soc.*, v. B 13, p. 242-49, 1951.

Lewis, B. N. On the analysis of interation in multidimensional contingency tables. *J. R. Statist. Soc.*, v. A 125, p. 88-117, 1962. § 7.6, App. B.

Lindley, D. V. The Bayesian analysis of contingency tables. *Ann. Math. Statist.*, v. 35, p. 1622-43, 1964. App. A.

Little, R. E. A note on estimation for quantal response data. *Biometrika*, v. 55, p. 578-79, 1968. § 6.5.

Mielke, P. W., & Siddiqui, M. M. A combinatorial test for independence of dichotomous responses. *J. Amer. Statist. Assoc.*, v. 60, p. 437-41, 1965.

Miettinen, O. S. The matched pairs design in the case of all-or-none responses. *Biometrics*, v. 24, p. 339-52, 1968.

Mosteller, F. Association and estimation in contingency tables. *J. Amer. Statist. Assoc.*, v. 63, p. 1-28, 1968.

Naylor, A. F. Comparison of regression constants fitted by maximum likelihood to four commom transformations of binominal data. *Ann. Hum. Genet.*, v. 27, p. 241-46, 1964. § 2.7.

Nerlove, M., & Press, J. Univariate and multivariate log-linear and logistic models. *Rand Corporation Report*, R-1306, 1973, EDA/NIH.

Reiersol, O. Linear and non-linear multiple comparisons in logit analysis. *Biometrika*, v. 48, p. 359-65, 1961.

Ries, P. N., & Smith, H. The use of chi-square preference testing in multidimensional problems. *Chemical Engineering Progress*, v. 59, p. 39-43, 1963. § 3.4.

Scheffe, H. *The analysis of variances*. New York, Wiley, 1959. § 2.1.

Silverstone, H. Estimating the logistic curve. *J. Amer. Statist. Assoc.*, v. 52, p. 567-77, 1957.

Simpson, E. H. The interpretation of interation in contingency tables. *J. R. Statist. Soc.*, v. B 13, p. 238-41, 1951. App. A.

Snell, E. J. A scaling procedure for ordered categorical data. *Biometrics*, v. 20, p. 592-607, 1964.

Stuart, A. The estimation and comparison of strengths of association in contingency tables. *Biometrika,* v. 39, p. 105-10, 1953.

Stuart, A. The comparison of frequences in matched samples. *Brit. J. Statist. Psycol.,* v. 10, p. 29-32, 1957.

Wald, A., & Wolfowitz, J. On a test of whether two samples are from the same population. *Ann. Math. Statist.,* v. 11, p. 147-62, 1940. Reimpresso em Wald. A. *Selected papers in statistics and probability.* New York, Mc-Graw Hill, 1955. § 5.7.

Walker, S. H., & Duncan, D. B. Estimation of the probability of an event as a function of several independent variables. *Biometrika,* v. 54, p. 167-79, 1967.

Yates, F. The use of transformation and maximum likelihood in the analysis of quantal experiments involving two treatments. *Biometrika,* v. 42, p. 382-403, 1955.