UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E CIÊNCIA DA COMPUTAÇÃO

IMECC

USO DE UM MODELO DE REGRESSÃO LOGÍSTICA E TÉCNICAS DE DIGNÓSTICO NA IDENTIFICAÇÃO DE FATORES DE RISCO EM PARTOS PÉLVICOS

Enrico Antonio Colosimo

Orientador: Flávio Celso Bartmann

Campinas 1985

Aos meus pais,

Antonio e Lia

Agradecimentos...

... ao Professor Flávio Celso Bartmann, pela orientação e amizade demonstrada durante a elaboração do trabalho;

... à Marta, pela revisão do texto e compreensão durante todo este tempo;

... ao Dr. Anibal Faundes Lathan, por ceder o arquivo de dados e pela prestativa orientação com relação aos mesmos;

... ao Masanao Ohira, cuja colaboração, na gravação e transporte da fita com os dados, foi indispensável;

... ao pessoal do Deptº de Estatística da UFMG, por reduzir a minha carga horária neste último semestre;

... à Cibele, Fátima, Creusa, Douglas, dentre outras pessoas, pela ajuda na parte computacional;

... à Emília, pela colaboração na parte médica;

... à Maria Júlia, pelo eficiente trabalho de datilografia.

ÍNDICE

INTRODUÇÃO

Resultados adversos no parto estão associados a um grande número de fatores de risco. É amplamente conhecido, por exemplo, que a mortalidade perinatal varia consideravelmente com a idade da mãe e com o número de partos anteriores por ela realizados. Entre outros fatores, incluem-se a hipertensão arterial, a isoimunização (fator RH), o posicionamento do bebê no útero da mãe, etc. O efeito de cada um desses fatores de risco é afetado pela possível presença de outros. É, pois, de extrema importância o estudo de tais interações.

Partos onde o bebê está na posição de cócoras, ao invés do usual são referidos na literatura como partos pélvicos. Eles correspondem a mais de 2% do total. Embora a percentagem seja pequena, no Brasil, anualmente são realizados mais de 100.000 partos em tais condições. Partos pélvicos estão associados com altos índices de resultados adversos e em particular com uma alta frequência de mortalidade perinatal. Nosso objetivo neste estudo é identificar os fatores de risco mais importantes, presentes nos partos pélvicos.

O estudo de associação de fatores de risco em situações como essa, eram limitados até pouco tempo por restrições de duas naturezas: i) disponibilidade restrita de recursos computacionais e ii) inadequação na metodologia estatística para a análise de tais dados.

O grande número de possíveis fatores de risco e a fre

quência relativamente pequena com que aparecem alguns destes, tornam necessário o acúmulo de um grande número de observações. Os arquivos de dados resultantes são enormes e de difícil tratamento, mesmo em computadores de grande porte.

A outra limitação mencionada acima decorre do fato de que até o fim da década de 60, a análise estatística disponível se restringia basicamente em cruzar os dados, apresentando-os em tabelas de contingência de duas dimensões. Utilizando o teste do qui-quadrado proposto por Pearson no início deste século, testa-se a independência entre as duas variáveis, ou contata-se se uma variável seria ou não um fator de risco para a resposta em questão.

Ao aplicar tal teste, estamos supondo que a probabilidade associada a cada cela da tabela de contingência é constante (hipótese de homogeneidade). O que se sabe na realidade, é que em várias tabelas existem fatores de confundimento, variáveis que não foram levadas em consideração, mas que influenciam nos resultados, invalidando essa suposição. À medida que estas variáveis são incorporadas ao estudo, a dimensão das tabelas aumentam, tornando a análise através desta técnica extremamente complexa.

O aparecimento do livro de Cox (1970) tratando da análise de dados binários, fez com que os modelos de regressão logística, como aquele que será desenvolvido mais adiante, se tornassem populares. Esses modelos permitiram, pela primeira vez, a representação simples e econômica da dependência de uma probabilidade com variáveis explicativas.

Outros métodos de análise estatística foram propostos com a mesma finalidade. Os modelos log-lineares, por exemplo (Everitt,1977; Fienberg, 1981) vem sendo usados com cada vez mais frequência. Eles apresentam algumas vantagens sobre o uso do teste do qui-quadrado usual: (i) dão uma aproximação sistemática para a análise de tabelas multidimensionais complexas; (ii) dão estimativas da magnitude aos efeitos de interesse, consequentemente permitem julgar a importância relativa dos diferentes efeitos.

Neste estudo vamos nos ater única e exclusivamente ao modelo de regressão logística.

Recentemente tem-se desenvolvido uma série de técnicas de diagnósticos (Pregibon, 1981), que nos permitem avaliar de maneira mais profunda a adequação do modelo proposto e a detecção de suas possíveis limitações.

Neste trabalho usaremos essas técnicas na detecção de fatores de risco presentes em partos pélvicos usando o arquivo de dados referentes a mais de 41000 partos coletados no Setor de Obstetrícia do Hospital Barros Luco de Santiago (Chile).

O texto foi dividido em três capítulos. No Capítulo 1, o modelo de regressão logística será apresentado, bem como as razões para o seu uso. As técnicas de diagnóstico usadas neste modelo são tratadas no Capítulo 2. No Capítulo 3 o arquivo de dados citado acima será visto em detalhes. O problema será definido e os resultados apresentados.

CAPÍTULO 1

REGRESSÃO LOGÍSTICA

Em diversas situações práticas nos vemos frente a problemas onde a nossa variável resposta é dicotômica. Nestes casos e em outros a análise estatística consiste em desenvolver um modelo onde a variável resposta será relacionada com variáveis explicativas. Por exemplo, em um estudo sobre morte perinatal, a variável resposta é dicotômica, e informações provenientes da mãe, tais como peso, idade, escolaridade, número de abortos, altura uterina e outras são variáveis que estarão, possivelmente, relacionadas com a resposta.

Neste tipo de problema é importante desenvolver métodos de análise, para determinar o grau de dependência entre as variáveis explicativas e a resposta, e possivelmente prever com base nelas a probabilidade de ocorrerem as respostas de interesse.

Em geral, para simplificar os procedimentos nós consideramos que tais observações binárias são disponíveis em N indivíduos, usualmente supostas como independentes.

Seria razoável que se tentasse, inicialmente, ajustar os modelos clássicos de regressão linear. Na seção 1.1 discutiremos a inadequação destes modelos quando temos respostas binárias. Na seção 1.2 vamos comparar algumas alternativas para mostrar que o modelo logístico é muito razoável para este tipo de situação e discutiremos alguns aspectos de sua estrutura. Na seção 1.3 será visto o processo numérico de Newton-Raphson necessário para determinar as estimativas dos parâmetros do modelo . Os testes de hipóteses de adequação do modelo são apresentados na

seção 1.4 e, finalizando este capítulo, na seção 1.5, serão abordados problemas que surgem com as variáveis explicativas, como interação, seleção de variáveis, transformação e a não ordenação.

1. REGRESSÃO LINEAR COM RESPOSTA DICOTÔMICA

Consideremos o conjunto de dados representados por uma variável resposta quantitativa, a ser denotada por y, e p variáveis explicativas ou independentes que serão denotadas por $x_1, x_2, \ldots, x_p$. O processo usual de análise neste caso é ajustar um modelo linear que consiste em tomar

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

onde y_i é a resposta para o i-ésimo caso, β_i, $i = 0,1,\ldots,p$, são os parâmetros a serem estimados e ε_i uma realização de uma variável com média zero, usualmente referida como erro aleatório. Considerando o estudo com N casos e passando a usar a notação matricial, temos

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

onde $\underset{\sim}{Y}$ é o vetor das observações (respostas) de dimensão N, $\underset{\sim}{X}$ é a matriz de delineamento* N x p+1 das variáveis explicativas que assumiremos ter posto completo, $\underset{\sim}{\beta}$ é o vetor de parâmetros de dimensão p+1 e $\underset{\sim}{\varepsilon}$ o vetor de erros de dimensão N.

---

* $\underset{\sim}{X}$ também é conhecida por matriz de modelo

O estimador de $\underset{\sim}{\beta}$ é computado usando-se o método de mínimos quadrados que consiste em minimizar, com respeito a $\underset{\sim}{\beta}$ a forma quadrática

$$(\underset{\sim}{Y} - \underset{\sim}{\underset{\sim}{X}}\underset{\sim}{\beta})'(\underset{\sim}{Y} - \underset{\sim}{\underset{\sim}{X}}\underset{\sim}{\beta})$$

onde cada termo é exatamente o erro aleatório, nos levando ao seguinte estimador para $\underset{\sim}{\beta}$ (Draper e Smith, 1981),

$$\underset{\sim}{\hat{\beta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y} \tag{1.1}$$

Usando (1.1), o vetor de valores estimados é

$$\underset{\sim}{\hat{Y}} = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$$

A soma de quadrados dos resíduos é dada por:

$$SQR = \underset{\sim}{Y}'[I - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}']\underset{\sim}{Y}$$

Sob a suposição de que

$$\underset{\sim}{\varepsilon} \sim N_N(\underset{\sim}{\phi}, \sigma^2 I_N) \tag{1.2}$$

ou seja, que as observações $Y_1, Y_2, \ldots, Y_N$ são independentes e distribuídas segundo uma normal N-variada com vetor de médias igual a $\underset{\sim}{\underset{\sim}{X}}\underset{\sim}{\beta}$ e matriz de variância-covariância $\sigma^2 I_N$, nós garantimos certas propriedades ótimas para os estimadores. A suposição (1.2) também nos permite usar as técnicas estatísticas de análise de variância e covariância.

Retornemos ao caso binário, isto é, quando y_i assume somente dois valores. Em geral, na literatura usa-se 1 (um) para representar a resposta de maior interesse e 0 (zero) para a outra resposta. Poderíamos ainda usar o modelo de regressão linear,

$$\theta_i = P(y_i = 1) = E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad (1.3)$$

Tal procedimento seria computacionalmente simples, pois todas as técnicas estatísticas mencionadas acima estão disponíveis em vários pacotes estatísticos. Neste caso, porém, surgem algumas dificuldades básicas.

Desde que y_i assume somente os valores 0 e 1, temos $y_i^2 = y_i$ e

$$VAR(y_i) = E(y_i^2) - [E(y_i)]^2$$

$$= \theta_i - \theta_i^2 = \theta_i(1 - \theta_i) \qquad (1.4)$$

Como para cada y_i nós temos um θ_i, a condição de variância constante na suposição (1.2) não é satisfeita, não sendo razoável, portanto, o uso daquelas técnicas estatísticas mencionadas anteriormente. Esta limitação pode ser contornada atribuindo-se pesos para as observações e, usando mínimos quadrados ponderados, podemos encontrar os estimadores para as probabilidades de interesse. A parte computacional torna-se, entretanto, bem mais complicada.

Uma segunda limitação decorre do fato de que os y_i's são variáveis discretas, logo não são normalmente distribuídas e nenhum método de estimação que seja linear nos y_i's será em geral,

completamente eficiente.

A mais séria limitação ao uso de (1.3) advém, entretanto, do fato que necessariamente

$$0 \leq \theta_i \leq 1 \qquad (1.5)$$

Como esta restrição não foi levada em consideração ao tratarmos o modelo, poderíamos encontrar θ_i fora do intervalo (1.5), o que seria ridículo, pois θ_i é uma probabilidade. Nesta situação, será razoável formular o problema da escolha dos estimadores dos β_i's como um problema de programação não linear, onde se levaria em consideração a restrição (1.5). Os problemas computacionais se complicariam ainda mais e os parâmetros do modelo teriam uma interpretação limitada.

Diante de tais limitações o mais razoável é procurar um modelo que seja mais adequado a este tipo de situação.

2. O MODELO LINEAR LOGÍSTICO

Consideremos inicialmente a situação mais simples, ou seja, aquela onde existe apenas uma variável explicativa e nossa variável resposta dicotômica. Essa relação de dependência é tal que para valores baixos da variável explicativa a probabilidade de falha θ_i é muito baixa, subindo rapidamente numa fase de transição e aproximando-se lentamente de 1. Uma discussão exaustiva é feita

no livro de Finney (1952).

A Figura 1 abaixo mostra uma curva sigmóide, que satisfaz a restrição (1.5) e condiz com a descrição acima.

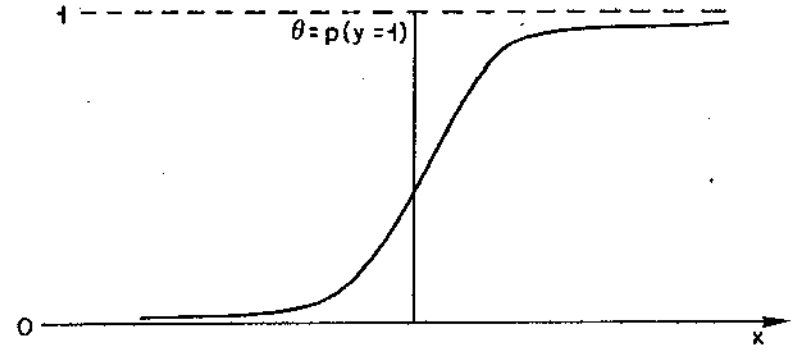

Fig. 1 - Exemplo de uma curva sigmóide

As funções de distribuição são muitas vezes da forma acima, o que as tornam candidatas naturais para serem usadas como a forma funcional do nosso modelo. Duas funções de distribuição têm merecido atenção, a da normal padrão e a da logística padrão dadas respectivamente por:

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

e

$$F(x) = \frac{e^x}{1+e^x}$$

Uma comparação entre estas duas escalas mais a linear e a angular pode ser encontrada no livro de Cox (1970).

A logística e a normal praticamente se igualam em todos os valores de probabilidade com exceção somente no caso dos dois extremos, ou seja, quando a probabilidade de sucesso é muito pequena ou muito perto de um. A curva normal aproxima-se do seu limite mais rapidamente que a logística. A linear e a angular entre 0,1 e 0,9, apesar de assumirem valores de probabilidade ligeiramente maiores, praticamente se igualam às outras duas. Fora deste intervalo de probabilidade as duas últimas curvas atingem seus limites rapidamente e os mesmos são finitos, o que usualmente restringe seu uso.

Na maioria das situações práticas, a escolha da logística ou da normal nos conduzem às mesmas conclusões. A preferência ao uso do modelo logístico pode ser explicada pela existência de métodos computacionais mais simples para a estimação dos parâmetros. A existência de estatística suficiente para o modelo logístico, como veremos abaixo, é uma vantagem teórica adicional.

Consideremos agora o modelo logístico com maior detalhe. Suponhamos que existem N indivíduos onde para cada um existe uma resposta associada:

$y_i = 1$, se o i-ésimo indivíduo é um sucesso e

$y_i = 0$, se o i-ésimo indivíduo é uma falha

Suponhamos que para cada um dos N indivíduos, p variáveis explicativas $x_{i1}, x_{i2}, \ldots, x_{ip}$ são medidas. Assim o modelo linear logístico é dado por

$$P(y_i = 1) = \frac{\text{Exp}\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\,\beta_j\right)}{1 + \text{Exp}\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\,\beta_j\right)}$$

e

$$P(y_i = 0) = \frac{1}{1 + \text{Exp}\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\,\beta_j\right)}$$

Vamos introduzir a variável indicadora $x_{i0} = 1$, $i = 1,2,...N$, para facilitar a notação. Usando notação matricial,

$$P(y_i = 1) = \theta_i = \frac{e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}}}{1 + e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}}} \tag{2.6}$$

$$P(y_i = 0) = 1 - \theta_i = \frac{1}{1 + e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}}} \tag{2.7}$$

onde, $\underset{\sim}{X}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})'$ é o vetor conhecido de variáveis explicativas e $\underset{\sim}{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ é o vetor dos parâmetros a serem estimados. As equações (2.6) e (2.7) parecem complicadas, entretanto o logaritmo da razão de θ_i por $1-\theta_i$ é uma função linear simples dos $\underset{\sim}{X}_i$'s,

$$\lambda_i = \ell n\left(\frac{\theta_i}{1-\theta_i}\right) = \sum_{j=0}^{p} x_{ij}\,\beta_j \tag{2.8}$$

λ_i é chamada de transformação logística da probabilidade θ_i ou

simplesmente de logit e a equação (2.8) é o modelo linear logísti_ co.

Estudaremos, agora, o modelo (2.6) através de sua fun_ção de verossimilhança. Em primeiro lugar, devemos notar que os dados são discretos. A função de verossimilhança é a probabilida_de de ocorrer as respostas obtidas dado os valores das variáveis explicativas. Logo, assumindo que as observações são independen_tes, a função de verossimilhança é dada por:

$$L(\underset{\sim}{\beta}/\underset{\sim}{Y};\underset{\sim}{X}) = \prod_{i=1}^{N} P(Y_i = y_i/\underset{\sim}{X}_i;\underset{\sim}{\beta}) \qquad (2.9)$$

onde $\underset{\sim}{Y}$ é um vetor de dimensão N constituído de 0's e 1's e y_i é uma componente deste vetor associada a i-ésima resposta. A proba_bilidade em (2.9) é dada pela expressão (2.6) quando houver um sucesso, e por (2.7) quando houver uma falha. Assim sendo, a expressão da verossimilhança do modelo é:

$$L(\underset{\sim}{\beta}/\underset{\sim}{Y};\underset{\sim}{X}) = \frac{\prod_{i=1}^{N} e^{(\underset{\sim}{X}_i'\underset{\sim}{\beta})y_i}}{\prod_{i=1}^{N} (1 + e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}})}$$

$$= \frac{\text{Exp}\left[\sum_{i=1}^{N} (\underset{\sim}{X}_i'\underset{\sim}{\beta})y_i\right]}{\prod_{i=1}^{N} (1 + e^{\underset{\sim}{X}_i\underset{\sim}{\beta}})} \qquad (2.10)$$

mas

$$\underset{\sim}{X}_i'\underset{\sim}{\beta} = \sum_{j=0}^{P} x_{ij}\,\beta_j \ ,$$

logo temos,

$$L(\underset{\sim}{\beta}/\underset{\sim}{Y};\underset{\sim}{X}) = \frac{\text{Exp}\left(\sum_{i=1}^{N} y_i \sum_{j=0}^{p} x_{ij}\,\beta_j\right)}{\prod_{i=1}^{N}(1+e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}})}$$

Trocando a ordem dos somatórios e fazendo

$$t_j = \sum_{i=1}^{N} x_{ij}\, y_i \tag{2.11}$$

tem-se,

$$L(\underset{\sim}{\beta}/\underset{\sim}{Y};\underset{\sim}{X}) = \frac{\text{Exp}\left(\sum_{j=0}^{p} \beta_j\, t_j\right)}{\prod_{i=1}^{N}(1+e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}})} \tag{2.12}$$

Como o denominador da expressão (2.12) é uma função exclusiva do vetor de parâmetros $\underset{\sim}{\beta}$,temos pelo teorema da fatorização de Fisher -Neyman (Bickel e Doksum, 1977) que T_j, $j = 0,1,\ldots,p$, são estatísticas suficientes para os parâmetros β_j, $j = 0,1,\ldots,p$. A variável aleatória T_j dada por (2.11) é simplesmente a soma de alguns dos termos da j-ésima coluna da matriz $\underset{\sim}{X}$. Os elementos incluídos na soma são aqueles que correspondem a uma resposta do tipo Y=1.

A função de logverossimilhança para o vetor de parâmetros $\underset{\sim}{\beta}$ é obtida aplicando o logaritmo na expressão (2.10)

$$\ell n\, L(\underset{\sim}{\beta}/\underset{\sim}{Y};\underset{\sim}{X}) = \sum_{i=1}^{N} \underset{\sim}{X}_i'\,\underset{\sim}{\beta}\, y_i - \sum_{i=1}^{N} \ell n(1+e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}})$$

$$= \sum_{i=1}^{N} [\underset{\sim}{X}_i'\,\underset{\sim}{\beta}\, y_i - \ell n(1+e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}})] \tag{2.13}$$

Então para obter o estimador de máxima verossimilhança de $\underset{\sim}{\beta}$ temos que derivar a expressão (2.12)e igualar a zero. Fazendo isto, $\hat{\underset{\sim}{\beta}}$ satisfaz ao seguinte sistema de equações

$$\sum_{i=1}^{N} \left[x_{ij} \left(y_i - \frac{e^{\underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}}}}{1 + e^{\underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}}}} \right) \right] = 0 \quad \text{para} \quad j = 0,1,\ldots,p$$

ou

$$\sum_{i=1}^{N} x_{ij}(y_i - \hat{\theta}_i) = 0 \qquad \text{para} \quad j = 0,1,\ldots,p \tag{2.14}$$

Escrevendo $s_i = y_i - \hat{\theta}_i$ e tomando a forma matricial, as equações de logverossimilhança ficam

$$\underset{\sim}{X}'\underset{\sim}{S} = \underset{\sim}{X}'(\underset{\sim}{Y} - \hat{\underset{\sim}{\theta}}) = 0 \tag{2.15}$$

As equações (2.15) embora muito similares às equações normais do modelo linear, são não lineares em $\hat{\underset{\sim}{\beta}}$, o que nos força a usar um processo numérico iterativo para determinar os valores de $\hat{\underset{\sim}{\beta}}$.

3. ESTIMAÇÃO DOS PARÂMETROS. O MÉTODO DE NEWTON-RAPHSON

Como foi dito na seção anterior, as equações (2.15) são não lineares em $\hat{\underset{\sim}{\beta}}$. Assim sendo, nós necessitaremos de um processo numérico para achar as estimativas dos parâmetros, que é o máximo da função de logverossimilhança (2.13).

Um dos métodos mais frequentemente usados para resolver equações deste tipo, pois em geral, converge rapidamente, é o método iterativo de Newton-Raphson. A Figura 2 abaixo dá uma descrição gráfica do método. Partindo de uma estimativa inicial x_1, nós prolongamos a tangente a curva neste ponto até interceptar o eixo das abcissas e tomamos este ponto, x_2, como a próxima aproximação. Este processo continua até que um valor x torne a função nula ou suficientemente próxima de zero.

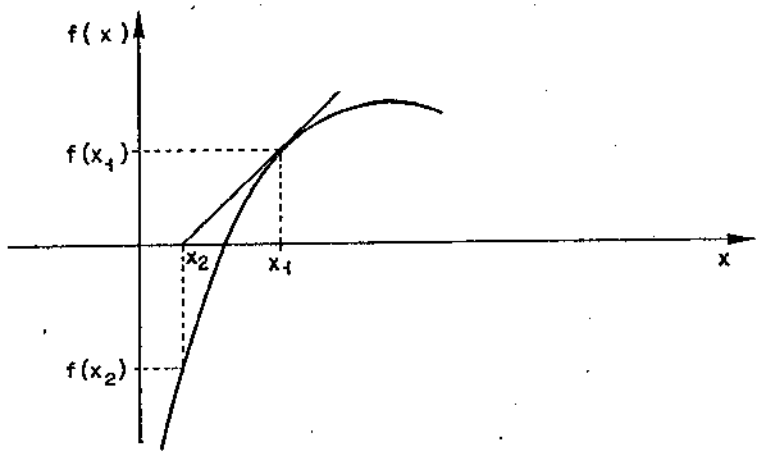

Fig. 2 - Interpretação geométrica do método iterativo de Newton-Raphson

No nosso caso, como queremos achar o máximo de uma função, devemos usar a derivada primeira, pois ela se anula no ponto de máximo e a derivada segunda para calcular as tangentes*. Isto nos leva ao seguinte esquema iterativo:

---

* Muitos trabalhos foram desenvolvidos em processos iterativos que dispensam o cálculo de derivadas segundas (Chambers,1973).No presente problema a derivada segunda é fácil de calcular e necessária para acharmos uma aproximação da matriz de variância-covariância das estimativas dos parâmetros como veremos adiante.

$$\hat{\underset{\sim}{\beta}}^{t+1} = \hat{\underset{\sim}{\beta}}^{t} + [I(\hat{\underset{\sim}{\beta}}^{t})]^{-1} S(\hat{\underset{\sim}{\beta}}^{t}) \qquad (3.16)$$

onde $S(\hat{\underset{\sim}{\beta}})$ e $I(\hat{\underset{\sim}{\beta}})$ são as funções "score" e informação, respectivamente.

A função "score" é dada por um vetor de dimensão $p + 1$ onde o j-ésimo elemento é

$$\frac{\partial \ell n L(\underset{\sim}{\beta}/\underset{\sim}{Y};\underset{\sim}{X})}{\partial \beta_j} = \sum_{i=1}^{N} x_{ij}(y_i - \theta_i) \quad , \quad j = 0,1,\ldots,p \qquad (3.17)$$

ou seja, é a expressão (2.14), e a função de informação é uma matriz p+1 x p+1, onde o elemento (ℓ,j) é:

$$\frac{\partial^2 \ell n L(\underset{\sim}{\beta}/\underset{\sim}{Y},\underset{\sim}{X})}{\partial \beta_j \, \partial \beta_\ell} = -\frac{\partial}{\partial \beta_\ell}\left[\sum_{i=1}^{N} x_{ij}(y_i - \theta_i)\right]$$

$$= -\frac{\partial}{\partial \beta_\ell}\left[\sum_{i=1}^{N} x_{ij}\, y_i - \sum_{i=1}^{N} x_{ij}\,\frac{e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}}}{1+e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}}}\right]$$

$$= \sum_{i=1}^{N} x_{ij}\left[\frac{e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}} x_{i\ell}(1+e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}}) - e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}} x_{i\ell} e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}}}{(1+e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}})^2}\right]$$

$$= \sum_{i=1}^{N} \frac{x_{ij}\, x_{i\ell}\, e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}}}{(1+e^{\underset{\sim}{X}_i'\underset{\sim}{\beta}})^2}$$

$$= \sum_{i=1}^{N} x_{ij}\, x_{i\ell}\, \theta_i(1-\theta_i) \qquad (3.18)$$

para $j = 0,1,\ldots,p$ e $\ell = 0,1,\ldots,p$

Com as expressões (3.16), (3.17) e (3.18) podemos escrever um programa para calcular as estimativas de máxima verossimilhança, ou seja, as estimativas dos parâmetros $\underset{\sim}{\beta}$. Iniciando o processo com $\hat{\underset{\sim}{\beta}}^{o} = 0$ ele converge em geral, em cinco ou seis iterações. Existem pacotes estatísticos, como o BMDP, com programas sobre regressão logística onde o método descrito é usado.

Uma vantagem em se usar o método acima, é que no passo final do processo iterativo obtemos também a inversa da função de informação, que é assintoticamente a matriz de variância-covariância de $\hat{\underset{\sim}{\beta}}$, que nos possibilita fazer inferências sobre os parâmetros baseados na teoria normal.

4. TESTES DE HIPÓTESES

Usualmente no modelo de regressão logística, como no modelo de regressão linear, são feitos dois conjuntos de testes de hipóteses com finalidades distintas. O primeiro é feito quando da escolha do modelo, onde o objetivo é testar se uma variável independente ou um conjunto delas tem coeficiente igual a zero. Depois de escolhido o modelo um outro conjunto de testes de hipóteses pode ser utilizado na verificação da adequação global do modelo. Vamos analisar em detalhe as duas situações.

Na escolha do modelo, usamos o teste da razão da verossimilhança (Bickel e Doksum, 1977) para a hipótese de que os coeficientes β_i's, correspondentes às q variáveis retiradas do modelo, são iguais a zero. Este teste é baseado na estatística

$$\chi^2_q = 2[\ell n\ L(\hat{\underset{\sim}{\beta}}/\underset{\sim}{X};\underset{\sim}{Y}) - \ell n(\underset{\sim}{\beta}^*/\underset{\sim}{X};\underset{\sim}{Y})] \qquad (4.19)$$

onde $\hat{\underset{\sim}{\beta}}$ é o vetor de estimativas dos parâmetros no modelo logístico (2.6) com todas as variáveis e $\underset{\sim}{\beta}^*$ é o vetor de estimativas dos parâmetros para aquelas variáveis que continuam no modelo quando q variáveis são retiradas.

Sob a hipótese que os coeficientes das variáveis retiradas são iguais a zero, χ^2_q tem assintoticamente uma distribuição qui-quadrado com q graus de liberdade . Valores altos de χ^2_q indicam que uma ou mais das q variáveis retiradas têm coeficiente de regressão diferente de zero.

A estatística (4.19) pode ser usada para testar se uma determinada variável, digamos x_p, mostra uma associação significativa como fator de risco para a variável resposta na presença das demais variáveis $x_1, x_2, \ldots, x_{p-1}$. Isto será usado na próxima seção, quando discutiremos a seleção de variáveis. Um teste exato de β_p igual a zero, quando consideramos os demais parâmetros como flutuação aleatória do modelo, foi dado por Cox (1970).

Outra possibilidade é construir intervalos de confiança para os parâmetros. Como já foi dito, a inversa da matriz de

informação (3.18) é assintoticamente a matriz de variância-covância da estimativa dos parâmetros. Então, uma estimativa para grandes amostras do erro padrão de $\hat{\beta}_i$ é

$$EP(\hat{\beta}_i) = \sqrt{\hat{\sigma}_{ii}}$$

onde $\hat{\sigma}_{ii}$ é o i-ésimo elemento da diagonal da inversa da matriz de informação. Um intervalo de confiança de aproximadamente nível $1 - \alpha$ para β_i é dado por

$$\hat{\beta}_i \pm z_{\alpha/2}\sqrt{\hat{\sigma}_{ii}}$$

O segundo conjunto de testes de hipóteses tem por objetivo julgar a adequação do modelo ajustado e é baseado na comparação das probabilidades estimadas e os valores observados para cada caso. Um teste para a adequação do modelo pode ser feito usando a seguinte estatística:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{\theta}_i)^2}{\hat{\theta}_i(1 - \hat{\theta}_i)}$$

Sob a hipótese nula que o modelo se ajusta bem aos valores observados, χ^2 tem assintoticamente uma distribuição qui-quadrado com N- p (número de parâmetros estimados) graus de liberdade. Valores altos de χ^2 indicam inadequação do modelo.

Uma forma alternativa de testar a adequação do modelo é usar o desvio (Pregibon, 1981) que é dado por:

$$D = -2[\ell n\ L(\hat{\beta}/X;Y) - \ell n\ L(\tilde{\beta}/Y)]$$

onde $\ell n\, L(\tilde{\theta}/Y)$ refere-se ao máximo da função de logverossimilhança quando cada ponto é ajustado exatamente. O desvio sob a hipótese nula tem também, assintoticamente, a mesma distribuição de χ^2. Este teste mede o desvio entre o modelo ajustado e o modelo saturado.

5. INTERAÇÃO ENTRE AS VARIÁVEIS E CONSIDERAÇÕES FINAIS

Na análise que fizemos nas seções anteriores do modelo de regressão logística nós supomos que não existia efeitos de interação entre as variáveis explicativas. Na realidade, isto nem sempre é verdade, ou seja, podem existir tais efeitos entre as variáveis. Uma forma de investigar a existência ou não do efeito de interação entre duas variáveis, será introduzir um termo da forma $\gamma(x_i, x_j)$ na equação do modelo (2.8). Estimando γ por máxima verossimilhança e testando a significância do parâmetro nós poderemos afirmar se existe ou não o efeito de interação entre as variáveis x_i e x_j. Por exemplo, em um modelo com sete variáveis explicativas e nós estamos desconfiando de um possível efeito de interação entre a segunda e a terceira, o modelo será alternativamente especificado como

$$\lambda_i = \ell n \left(\frac{\theta_i}{1-\theta_i} \right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_7 x_{i7} + \gamma(x_2 x_3)$$

Efeitos de interação mais complicados que $\gamma(x_2 x_3)$ podem também ser especificados e testados. Em geral, entretanto, começa-se com a forma mais simples, a menos que existam evidências externas sugerindo outra alternativa.

Uma outra forma de investigar interações é conhecer as relações univariadas para interpretar os resultados da análise multivariada. Por exemplo, se nível de colesterol, pressão sanguínea e peso entraram como variáveis explicativas em um modelo de regressão logística onde a presença ou não de doença coronária é a variável resposta o coeficiente para peso, que pode ser estatisticamente significante por si mesmo, pode aproximar a zero na equação multivariada. Isto não significa que o peso não é importante, mas sugere que seu efeito é medido pelos níveis de pressão sanguínea e colesterol. Uma forma de usar relações univariadas como uma aproximação efetiva para este problema, é baseado no procedimento "stepwise", que adiciona ou retira variáveis da equação para determinar o impacto nos coeficientes das demais variáveis. Este impacto é medido pelo teste da razão de verossimilhança descrito na seção 1.4, e um pacote estatístico que tem esta versão "stepwise" é o BMDP (Programa LR).

O modelo logístico (2.8) implica em uma dependência linear do logaritmo da razão das probabilidades da variável resposta em cada uma das variáveis explicativas. Estimativas dos parâmetros do modelo são feitas sob a suposição que as variáveis tem uma escala de medida. Em muitas situações a relação linear (2.8) é mais razoável se usarmos transformações das variáveis explicativas originais. Por exemplo, se x_{ij} representa idade em anos no modelo logístico $\lambda_i = \ell n[\theta_i/(1-\theta_i)] = \beta_0 + \beta_1 x_{i1}$, implica que cada ano acrescido na idade é associado com um aumento de β_1 unidade no λ_i. Se, entretanto, λ_i não for uma função linear da

idade, nós podemos tentar transformar a variável x_{ij}, usando, por exemplo, $\sqrt{x_{i1}}$ ou $\ell n\ x_{i1}$. Alternativamente, podemos introduzir parâmetros adicionais, tais como $\beta_1 x_{i1} + \beta_2 x_{i1}^2$.

Um outro problema que surge frequentemente com as variáveis explicativas em um modelo de regressão logística é a forma de interpretação dos parâmetros. Variáveis discretas, tais como , religião, raça e sexo não têm escala ordenada de medida. Desta forma, se codificarmos, de forma arbitrária, a variável raça como:

$$X_{ij} = \begin{cases} 0, \text{ se o i-ésimo indivíduo for branco} \\ 1, \text{ se o i-ésimo indivíduo for negro} \\ 2, \text{ se o i-ésimo indivíduo for amarelo} \\ 3, \text{ se o i-ésimo indivíduo for de outra raça,} \end{cases}$$

o parâmetro estimado não tem interpretação, pois dizer que o $\ell n[\theta_i/(1-\theta_i)]$ é acrescido de uma unidade se passamos de um indivíduo da raça branca para um da negra não faz sentido algum.

Variáveis indicadoras tomando os valores um ou zero para designar o presença ou ausência do atributo, devem ser usadas para representar corretamente os efeitos de tais variáveis em um modelo de regressão logística. Desta forma, a variável raça tendo quatro categorias necessita usar três variáveis indicadoras. Primeiramente, nós designamos uma das categorias como referência e para cada uma das demais introduz-se uma variável R_j, que é codificada como um (presente) e zero (ausente). No exemplo, tomando a raça branca como grupo de referência, nós introduzimos três variáveis indicadoras para raça: R_1, codificada como um (preta) ou zero (não preta); R_2, codificada como um (amarela) ou zero (não ama

rela) e R_3 codificada como um se o indivíduo não é branco nem preto nem amarelo e zero caso contrário. A variável x_{ij} é então trocada por três variáveis indicadoras R_1, R_2 e R_3 e a saída dos dados para a análise da regressão deve ser recodificada para refletir esta mudança. Em geral, se uma variável discreta tem k categorias, será necessário usar k-1 variáveis indicadoras e o coeficiente β_j associado com cada uma das variáveis indicadoras representa a mudança no $\ell n[\theta_i/(1-\theta_i)]$ para esta categoria relativa à categoria de referência.

CAPÍTULO 2

DIAGNÓSTICOS

Uma aplicação de modelos de regressão logística é na análise de dados obtidos em estudos observacionais, como é o caso do exemplo apresentado no próximo capítulo.Em estudos deste tipo, surgem comumente entre as observações "outliers"* e pontos extremos. "Outlier" é o ponto que tem resíduo muito maior em valor absoluto que os outros e ponto extremo é um ponto deslocado dos demais no espaço de delineamento. "Outlier" está associado com a resposta e ponto extremo com as variáveis explicativas.

O método usual de estimar os parâmetros em um modelo de regressão logística é, como foi visto no Capítulo 1, o de máxima verossimilhança, e este método é sensível a dados "ruins" como os citados acima. Desta forma, técnicas de diagnóstico para identificar tais observações são de grandes importância.

No modelo de regressão linear é conhecido o efeito de "outliers" e pontos extremos no ajuste de mínimos quadrados (Cook e Weisberg, 1982). Uma revisão das técnicas de diagnóstico usadas na análise dos modelos de regressão linear é feita na seção 2.1.

A saída usual de um programa de computador de um modelo de regressão logística e as primeiras medidas de diagnóstico que são os resíduos e a matriz de projeção aparecem na seção 2.2

Medidas de diagnóstico que quantificam o efeito de cada uma das observações sobre o modelo ajustado são descritas no res

---

* Foram propostas algumas traduções para o português da palavra "outlier" (ponto aberrante, ponto anômalo, ponto remoto, etc.). Entretanto,como nenhuma delas é de uso geral, manterei o termo em inglês, que é conhecido por todos.

tante do capítulo. Na seção 2.3.estudamos a influência de cada observação sobre os parâmetros estimados e na seção 2.4 sobre as estatísticas de adequação do modelo.

1. REVISÃO DOS MÉTODOS DE DIAGNÓSTICOS PARA O MODELO DE REGRESSÃO LINEAR

Diagnósticos são técnicas usadas para identificar aspectos de um conjunto de observações que não condizem com as suposições feitas no processo de modelagem. Tais técnicas são úteis para reconhecer importantes fenômenos que de outra forma não seriam notados. Detecção de "outliers" e pontos extremos são exemplos disto.

Com a crescente disponibilidade de recursos computacionais, uma enorme quantidade de ferramentas de diagnósticos foram propostas recentemente. Os dois elementos básicos usados para construir ferramentas de diagnósticos são os resíduos

$$e_i = Y_i - \hat{Y}_i \qquad (1.1)$$

e a matriz

$$\underset{\sim}{H} = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'$$

comumente chamada de matriz chápeu por transformar o vetor de

observações $\underset{\sim}{Y}$ no vetor de valores estimados $\hat{\underset{\sim}{Y}}$.

Os resíduos, como mostra a expressão (1.1), são as diferenças entre os valores observados e os ajustados pela equação de regressão. O seu uso inclui a detecção de possíveis "outliers". Os resíduos associados a pontos extremos são tipicamente pequenos o que os torna inadequados para a detecção daqueles.

A matriz $\underset{\sim}{H}$ é simétrica e idempotente, e representa a transformação linear que projeta ortogonalmente qualquer vetor de dimensão N no espaço gerado pelas colunas da matriz $\underset{\sim}{X}$. Os elementos da diagonal principal da matriz $\underset{\sim}{H}$ são úteis na detecção de pontos extremos. Uma vez que a matriz $\underset{\sim}{H}$ é idempotente e simétrica, temos que

$$\text{TRAÇO}(\underset{\sim}{H}) = \text{POSTO}(\underset{\sim}{H}) = \sum_{i=1}^{N} h_{ii} = p+1 \quad \text{e} \quad 0 \leq h_{ii} \leq 1$$

ou seja, a soma dos elementos da diagonal da matriz $\underset{\sim}{H}$ é igual ao número de parâmetros do modelo ajustado. Se as observações fossem igualmente influentes, pelo menos sobre o seu próprio valor ajustado, o valor médio dos elementos h_{ii} seria

$$h_{ii} = (p+1)/N$$

Hoaglin e Welsch (1978) sugerem usar $h_{ii} \geq 2(p+1)/N$ como guia para determinar se um ponto é ou não extremo.

As quantidades e_i e h_{ii} são úteis para detectar tais pontos, mas não para determinar seu impacto em vários aspectos do ajuste. Por exemplo, estimativas dos parâmetros, valores ajustados,

medidas de adequação do modelo. Uma forma de quantificar o efeito destes pontos é investigar o impacto de retiradas de observações individuais em alguns aspectos do modelo.

Retirando-se uma única observação, a j-ésima, digamos, a expressão para a mudança nas estimativas de mínimos quadrados dos parâmetros é dada por:

$$\Delta_j\hat{\underset{\sim}{\beta}} = \hat{\underset{\sim}{\beta}}(1) - \hat{\underset{\sim}{\beta}}(0) = \frac{(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\ \underset{\sim}{X}_j\ e_j}{1 - h_{jj}}$$

onde (1) e (0) representam respectivamente a presença e ausência da j-ésima observação.

Uma medida escalar c_j do efeito da retirada da j-ésima observação sobre todos os coeficientes simultaneamente é

$$c_j = (\Delta_j\hat{\underset{\sim}{\beta}})'\ \underset{\sim}{X}'\underset{\sim}{X}\ (\Delta_j\hat{\underset{\sim}{\beta}}) = \frac{e_j^2\ h_{jj}}{(1-h_{jj})^2}$$

esta medida, comumente chamada de D-Cook, dá a distância quadrada de $\hat{\beta}(1)$ a $\hat{\beta}(0)$ relativa a geometria fixa X'X. Padronizando-se aproximadamente c_j, pode-se ter várias interpretações. Por exemplo, dividindo c_j por ps^2, onde s^2 é o quadrado médio residual, é o deslocamento da região de confiança simultânea para todos os parâmetros do modelo devido a retirada de j-ésima obervação. Observações que produzem grandes valores de c_j influenciam no ajuste total do modelo. Uma discussão geral sobre esta medida é encontrada em Cook e Weisberg (1982).

Um outro diagnóstico sumário é a variação na soma de quadrados residual (SQR) devido a retirada da j-ésima observação:

$$\Delta_j SQR = SQR(1) - SQR(0) = \frac{e_j^2}{1-h_{jj}}$$

A grande vantagem em se usar estas três últimas medidas de diagnóstico é que elas não requerem o ajuste do novo modelo de regressão com cada uma das observações retiradas. Os próprios valores achados para o modelo com todas as observações nos permitem calcular tais medidas tornando rápida e econômica a parte computacional.

Essas técnicas de diagnósticos apresentadas foram estendidas e adaptadas para o modelo de regressão logística por Pregibon (1981). Uma discussão detalhada de tais técnicas é feita nas próximas seções.

2. RESÍDUOS E MATRIZ DE PROJEÇÃO PARA O MODELO DE REGRESSÃO LOGÍSTICA

Após o ajuste por máxima verossimilhança do modelo de regressão logística, vários resultados deste processo são disponíveis. Tipicamente, as quantidades de maior importância são as seguintes:

(a) o vetor de parâmetros estimados, $\hat{\underset{\sim}{\beta}}$;

(b) os erros padrão dos parâmetros estimados, $EP(\hat{\beta}_j)$;

(c) a covariância entre os parâmetros estimados, $Cov(\hat{\beta}_i, \hat{\beta}_j)$;

(d) a estatística qui-quadrado de adequação do modelo,

$$\chi^2 = \sum_{i=1}^{N} (y_i - n_i\hat{\theta}_i)^2 / n_i\hat{\theta}_i(1-\hat{\theta}_i)^*;$$

(e) os componentes individuais χ^2, $\chi_i = (y_i - n_i\hat{\theta}_i)/\sqrt{n_i\hat{\theta}_i(1-\hat{\theta}_i)}$;

(f) o desvio $D = -2 \sum_{i=1}^{N} [\ell n\ L(\hat{\beta}/X_i; y_i) - \ell n\ L(\hat{\theta}_i/y_i)]$;

A partir dessas quantidades do ajuste de máxima verossimilhança podem ser desenvolvidas medidas de diagnósticos para identificar observações que não estão bem explicadas pelo modelo.

Inicialmente, como no modelo de regressão linear, as medidas de diagnóstico para identificar "outliers" e pontos extremos no modelo de regressão logística são também um vetor de resíduos e uma matriz de projeção. No modelo de regressão linear, os resíduos são simplesmente as diferenças entre os valores observados e os estimados. No modelo de regressão logística os resíduos podem ser definidos de várias maneiras. Por exemplo, uma analogia com mínimos quadrados em regressão linear, sugere os componentes individuais do χ^2. Pregibon (1981) usa um conjunto de resíduos baseados nos componentes individuais do desvio, que são definidos como:

---

* Nesta expressão e nas seguintes, n_i representa o número de observações existentes para a i-ésima combinação dos valores das variáveis explicativas. Uma explicação mais detalhada será dada no Capítulo 3.

$$d_i = \pm\sqrt{2}[\ell n\ L(\hat{\theta}_i/y_i) - \ell n\ L(\hat{\beta}/\underset{\sim}{X}_i;\underset{\sim}{Y}_i)$$

onde o sinal mais é usado se o $\text{logit}(y_i/n_i) > \underset{\sim}{X}_i'\hat{\underset{\sim}{\beta}}$ e o menos caso contrário. No caso em que temos y_i igual a n_i ou zero, os valores de d_i são

$$d_i = \sqrt{2}\ n_i\ \ell n\ \hat{\theta}_i \quad , \text{ se } y_i = n_i$$

$$-\sqrt{2}\ n_i\ \ell n(1-\hat{\theta}_i), \text{ se } y_i = 0$$

Ambos χ^2 e D são medidas usadas para testar a adequação do modelo. A primeira mede os desvios relativos entre os valores observados e os ajustados e a segunda mede os desvios entre os máximos das funções de logverossimilhança dos valores observados e dos ajustados. Portanto, nos dois casos, grandes componentes individuais indicam observações mal ajustadas pelo modelo.

O análogo da matriz de projeção para o modelo de regressão logística, que também representaremos por $\underset{\sim}{H}$, é dada por:

$$\underset{\sim}{H} = \underset{\sim}{V}^{1/2}\ \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{V}\underset{\sim}{X})^{-1}\ \underset{\sim}{X}'\underset{\sim}{V}^{1/2} \qquad (2.2)$$

onde $\underset{\sim}{V}$ é uma matriz diagonal em que $v_{ii} = n_i\hat{\theta}_i(1-\hat{\theta}_i)$, ou seja, a variância estimada da i-ésima observação. Para entender a idéia de como foi obtida a expressão (2.2), basta tomarmos o modelo de regressão linear com $\underset{\sim}{V}^{1/2}\underset{\sim}{X}$ como a matriz de delineamento. Neste caso, $\underset{\sim}{H}$ também é simétrica e idempotente. Isto sugere que valores altos de h_{ii} indicam pontos extremos no espaço de delineamento.

3. EFEITO DA RETIRADA DE UMA OBSERVAÇÃO SOBRE OS PARÂMETROS DO MODELO

Os resíduos e a matriz de projeção são úteis para detectar quais observações não estão bem explicadas pelo modelo ou dominam algum aspecto do ajuste. Estas quantidades, entretanto, não podem medir adequadamente o efeito das observações sobre os componentes do modelo ajustado. Nesta seção, nós iremos apresentar medidas de diagnóstico para quantificar o impacto sobre os parâmetros ajustados quando retiramos uma observação.

As estimativas dos parâmetros no modelo de regressão logística são obtidas usando um processo iterativo (seção 1.3). Desta forma, não temos expressões simplificadas como no modelo de regressão linear, para a variação destas estimativas quando retiramos uma observação. A forma de fazer isto é, então, ajustar um modelo usando o processo descrito para cada grupo formado pela retirada de uma observação e comparar com o modelo ajustado com todas as observações presentes. Seria necessário ajustar tantos modelos quanto o número de observações existentes, o que acarretaria num custo computacional elevado, principalmente quando temos muitas observações.

Uma forma alternativa que dá bons resultados, pois diminui consideravelmente o tempo computacional e simplifica os cálculos é iniciar o processo iterativo de máxima verossimilhança com os valores obtidos para o modelo com todas as observações e

terminar o processo após um passo*. A equação corresponde a do modelo de regressão linear com $\underset{\sim}{V}^{1/2}\underset{\sim}{X}$ como a matriz de delineamento. Desta forma, temos:

$$\Delta_j^1\hat{\underset{\sim}{\beta}} = \hat{\underset{\sim}{\beta}}(1) - \hat{\underset{\sim}{\beta}}(0) = \frac{(\underset{\sim}{X}'\underset{\sim}{V}\underset{\sim}{X})^{-1}\,\underset{\sim}{x}_j'\,(y_j - n_j\,\hat{\theta}_j)}{1 - h_{jj}}$$

onde $\Delta_j^1\hat{\underset{\sim}{\beta}}$ é a variação a um passo no vetor de parâmetros estimados quando retiramos a j-ésima observação.

Gráficos de $\Delta_j^1\hat{\beta}_i/EP(\hat{\beta}_i)$ versus j são úteis para identificar observações que estão causando instabilidades nos coeficientes selecionados. Se a variação nos coeficientes é desprezível, então a observação exerce pouca influência nos coeficientes e consequentemente no ajuste. Por outro lado, grandes variações nos coeficientes implicam em observação influente.

Em ajustes do modelo de regressão logística onde o número de variáveis explicativas é grande, analisar os gráficos de cada coeficiente para determinar se a observação em questão tem uma grande influência sobre o ajuste é uma tarefa árdua e que demanda um longo tempo de análise. A tarefa complica-se mais, quando além de um grande número de variáveis explicativas temos também um grande número de observações.

---

* Pregibon (1981) discute a precisão desta aproximação a um passo e conclui que ela tende a subestimar o valor obtido pela iteração completa, mas que isto se torna sem importância quando se quer identificar casos influentes.

Uma medida de diagnóstico que sintetiza a influência de uma observação sobre todos os coeficientes pode ser obtida através da expressão

$$-2[\ell n\ L(\underset{\sim}{\beta}/\underset{\sim}{X},\underset{\sim}{Y}) - \ell n(\hat{\underset{\sim}{\beta}}/\underset{\sim}{X},\underset{\sim}{Y})] = c \qquad (3.3)$$

Esta equação descreve o contorno de uma região de confiança assintótica para o vetor de parâmetros $\underset{\sim}{\beta}$. Substituindo nesta equação o vetor $\underset{\sim}{\beta}$ por $\hat{\underset{\sim}{\beta}}(0)$, o vetor de parâmetros estimados sem a j-ésima observação, temos uma medida escalar do deslocamento desta região de confiança devido a retirada da j-ésima observação.

A equação (3.3) corresponde novamente àquela do modelo de regressão linear com $\underset{\sim}{V}^{1/2}\underset{\sim}{X}$ como a matriz de delineamento, ou seja,

$$[\hat{\underset{\sim}{\beta}}(0) - \hat{\underset{\sim}{\beta}}(1)]'\ \underset{\sim}{X}'\underset{\sim}{V}\underset{\sim}{X}[\hat{\underset{\sim}{\beta}}(0) - \hat{\underset{\sim}{\beta}}(1)] = c_j$$

Se desenvolvermos a expressão acima a um passo, obtemos uma medida de diagnóstico do deslocamento da região de confiança, que é a seguinte:

$$c_j^1 = \frac{\chi_j^2\ h_{jj}}{(1 - h_{jj})^2}$$

Note que c_j^1 é uma função exclusiva dos componentes do qui-quadrado e dos elementos da diagonal da matriz de projeção.

Resultados similares ao dado acima são obtidos considerando a situação contrária, isto é, tomando o contorno de uma re

gião de confiança assintótica para o vetor de parâmetros sem a j-ésima observação, que é dado por

$$-2[\ell n\, L(\underset{\sim}{\beta}/\underset{\sim}{X},\underset{\sim}{Y}) - \ell n(\hat{\underset{\sim}{\beta}}(0)/\underset{\sim}{X};\underset{\sim}{Y}) = c$$

Substituindo o vetor $\underset{\sim}{\beta}$ por $\hat{\underset{\sim}{\beta}}(1)$, temos uma medida escalar do deslocamento desta região de confiança devido à inclusão da j-ésima observação. Cálculos similares desenvolvidos para c_j^1 nos levam à seguinte expressão para a medida de diagnóstico do deslocamento da região de confiança,

$$\bar{c}_j^1 = \frac{\chi_j^2\, h_{jj}}{1 - h_{jj}}$$

Em comparação com c_j^1, $\bar{c}_j^1$ terá sempre um valor menor. Por este motivo, c_j^1 será usualmente preferido para medir as variações totais nos coeficientes devido à j-ésima observação.

Novamente, gráficos de c_j^1 e $\bar{c}_j^1$ versus j dão informações úteis sobre a influência da j-ésima observação sobre o ajuste.

4. EFEITO DA RETIRADA DE UMA OBSERVAÇÃO SOBRE AS ESTATÍSTICAS DE ADEQUAÇÃO DO MODELO

As medidas de diagnóstico são disponíveis em um outro

aspecto fundamental do ajuste que são suas próprias estatísticas de adequação do modelo. Observações que influenciam fortemente o modelo devem induzir grandes variações na qualidade do ajuste quando medidas por estas estatísticas. Mudanças nas estatísticas de adequação do modelo podem tomar duas formas:

(i) se a j-ésima observação não está bem ajustada pelo modelo, mudanças em D e χ^2 causadas pela retirada desta observação estão isoladas nos componentes individuais d_j e χ_j.

(ii) se a j-ésima linha de $\underset{\sim}{X}$ é um ponto extremo do espaço de delineamento, as variações em D e χ^2 serão o resultado da variação de todos os componentes individuais.

Um valor alto para a variação nestas estatísticas de adequação do modelo não indicará se o caso em consideração é (i) ou (ii), mas informações de outras medidas de diagnóstico (especialmente de h_{jj}) esclarecem eventualmente a situação.

Comecemos determinando a variação sobre o desvio devido a retirada da j-ésima observação. A expressão do desvio nesta situação é dada por:

$$D_0(\hat{\underset{\sim}{\beta}}(0);\underset{\sim}{X};\underset{\sim}{Y}) = 2[\ell n\ L(\hat{\underset{\sim}{\theta}}(0)/\underset{\sim}{Y}) - \ell n\ L(\hat{\underset{\sim}{\beta}}(0)/\underset{\sim}{X},\underset{\sim}{Y}]$$

Desenvolvendo a expressão acima a um passo, obtemos

$$\Delta_j D = D_1(\hat{\beta}(1)/\underset{\sim}{X};\underset{\sim}{Y}) - D_0(\hat{\beta}(0);\underset{\sim}{X};\underset{\sim}{Y}) = d_j^2 + \frac{\chi_j^2 \, h_{jj}}{1 - h_{jj}}$$

A estatística χ^2 é similar a SQR na regressão linear pois ambas são somas de quadrado da diferença entre valores observados e ajustados. A aproximação a um passo para a variação do χ^2 devido a retirada da j-ésima observação é dada por:

$$\Delta_j \chi^2 = \chi^2(1) - \chi^2(0) = \frac{\chi_j^2}{1 - h_{jj}}$$

É importante notar que $\Delta_j \chi^2$ é o análogo na regressão linear logística de Δ_jSQR (seção 1.2) na regressão linear.

CAPÍTULO 3

APRESENTAÇÃO DO ARQUIVO DE DADOS E SEUS RESULTADOS

Nos dois primeiros capítulos deste trabalho,discutimos as bases técnicas do processo de modelagem da dependência de uma probabilidade com uma série de variáveis explicativas e técnicas de diagnósticos para julgar a adequação deste modelo. O objetivo do presente capítulo é ilustrar o uso destas técnicas num estudo observacional cuja finalidade é identificar fatores de risco e quantificar seu impacto no ajuste.

A seção 3.1 dá uma idéia detalhada do arquivo de dados considerado. A seção 3.2 descreve o processo de ajuste do modelo de regressão logística. O uso dos diagnósticos na avaliação da qualidade do modelo ajustado está ilustrado na seção 3.3. Finalmente, na seção 3.4, apresentamos as conclusões que a análise do modelo ajustado nos permite tirar.

3.1. O ARQUIVO DE DADOS

Dados relativos a cerca de 41000 partos foram coletados durante dois anos (1969-70), no hospital Barros Luco, em Santiago (Chile), instituição pública mantida pelo governo chileno, pela equipe do Dr. Anibal Faundes Lathan . Os pacientes que utilizam os serviços de tal hospital são provenientes em geral de camadas menos privilegiadas da população. Os dados foram coletados dentro de um programa cujos objetivos gerais eram a identificação de fatores de risco de parto e o desenvolvimento de proce

dimentos simples e eficientes para o seu tratamento.

Entre as variáveis que formam cada caso, podemos distinguir três grandes grupos: variáveis históricas,do processo e respostas. As variáveis históricas caracterizam o período pré-parto, as do processo são as do parto propriamente dito e as respostas são relativas ao recém-nascido. Elas aparecem nesta ordem na sequência natural da codificação com exceção das três últimas,que são variáveis históricas.

As variáveis históricas são dados pessoais da mãe, como idade, estado civil, número de abortos, etc., e informações médicas relativas ao período de gravidez como amenorreia (número de semanas sem menstrução), controle do pré-natal, infecções, etc.

As variáveis do processo formam o maior grupo. As primeiras variáveis caracterizam o início do parto como estado do líquido ovular, tipo de rotura de membrana, analgesia (anestesia)para o trabalho de parto, etc., até as últimas, que indicam o período final, como duração do período expulsivo, analgesia para exposição ou intervenção e características do cordão umbilical.

O terceiro grupo, as variáveis respostas, constituem -se de dados relativos ao recém-nascido, como peso, apgar 1' (nota atribuída ao recém-nascido a um minuto de vida pelo médico, baseada em sua saúde naquele instante) e estado do recém-nascido na alta da mãe.

As variáveis, num total de 27, são discretas no caso das qualitativas, ou foram discretizadas. No Apêndice I, apresentamos

todas estas variáveis e suas respectivas categorias.

Duas variáveis deste arquivo não têm uma escala ordenada. A primeira, forma de término, será substituída por três variáveis indicadoras, como discutido na seção 1.5. A segunda, duração do período expulsivo, tem oito categorias ordenadas e uma nona, cesárea, que não se ordena com as demais. Esta variável será substituída por duas, onde a primeira é uma variável indicadora, um (cesárea) e zero (sem cesárea), e a segunda que leva em consideração as categorias previamente ordenadas e atribui zero a categoria cesárea.

Um subarquivo destes dados de grande interesse dada a alta frequência de resultados adversos é o constituído de mães primíparas (primeiro parto) com apresentação pélvica, ou seja, o recém-nascido está na posição de cócoras ao invés da usual*. Assim, o objetivo fundamental da análise é identificar os fatores de risco mais importantes nestes casos e ajustar um modelo que atribua pesos a estes fatores para estimar a probabilidade de "sucesso".

Várias respostas dicotômicas são de grande interesse médico: morte perinatal, morte materna, apgar baixo, presença de infecções, etc. Nós nos preocupamos aqui, somente com o apgar um minuto que reflete o estado geral do recém-nascido. Esta variável foi recodificada da seguinte forma:

---

* A intenção inicial deste trabalho era trabalhar com todos os 41000 casos. Entretanto, razões de tempo e espaço fizeram com que nos limitássemos a uma análise detalhada deste subarquivo.

0 , se o apgar for ≤ 6

1 , se o apgar for ≥ 7

A razão para esta divisão é que os médicos distinguem as condições de saúde do recém-nascido pelas duas categorias acima. A categoria zero caracteriza a resposta adversa e a outra, boa saúde.

3.2. O MODELO DE REGRESSÃO LOGÍSTICA AJUSTADO

Nesta seção vamos ajustar um modelo simples que identifique os fatores de risco mais importantes. Não faz sentido tentar detectar todos, pois existem fontes não controladas de variação como habilidade da equipe médica, estado psicológico da mãe, etc., que confundem facilmente efeitos pouco importantes.

O subarquivo de primíparas com apresentação pélvica é constituído de 503 casos. Em vários casos tinhamos variáveis com falta de informação. Nossa posição foi retirar todos os casos que tinham alguma falta de informação, obtendo, desta forma, um arquivo com 193 casos.

O primeiro passo foi passar este arquivo pelo processo "stepwise" do programa LR do BMDP. Um resumo deste processo é mostrado a seguir.

Passo Nº	Termo Incluído	Grau de Liberdade	Termo Removido	Logveros-similhança	Melhora		Adequação de Ajuste	
					Qui-quadrado	Valor P	Qui-quadrado	Valor P
0				-105,988			211,976	0,142
1	Tipo Term.	1		- 90,681	30,614	0,000	181,363	0,661
2	Dur. Dilat.	1		- 82,131	17,100	0,000	164,262	0,903
3	Est. Liq.	1		- 76,452	11,359	0,001	152,904	0,971
4	Peso	1		- 70,249	12,406	0,000	140,498	0,995
5		1	Tipo term.	- 71,235	1,972	0,160	142,470	0,994
6	Carac.Cord.	1		- 66,155	10,159	0,001	132,310	0,999
7	Amenor.	1		- 63,183	4,685	0,030	127,626	1,000

Note que a variável tipo de término entrou e foi removida do modelo e no final ficamos com as seguintes variáveis: duração para dilatação, estado do líquido ovular, peso do recém-nascido, característica do cordão e amenorreia.

O arquivo original de 503 casos foi retomado e dele retiramos os casos com falta de informação somente para as variáveis citadas acima, obtendo, assim, um arquivo com 306 casos que apresentaram, entretanto, somente 222 combinações distintas das variáveis explicativas. Deste arquivo obtivemos os seguintes valores estimados para os coeficentes dos fatores de risco:

Termo	Coeficiente	Erro padrão
Amenor.	0,175	0,082
Est. liq.	-0,492	0,124
Dur. Dilat.	-0,496	0,092
Carac. Cord.	-0,353	0,154
Peso	0,462	0,152
Constante	1,012	0,766

e o modelo é então dado por

$$\lambda_i = \ell n \left(\frac{\hat{\theta}_i}{1-\hat{\theta}_i} \right) = 1,012 + 0,175\, x_{i1} - 0,492\, x_{i2} - 0,496\, x_{i3} - 0,353\, x_{i4} + 0,462\, x_{i5}$$

Cada coeficiente indica a quantidade que será acrescida ao logit (λ_i) se aumentarmos uma unidade no valor de alguma variável explicativa. Por exemplo, o logit cresce 0,462 para cada acréscimo

de 500 g no peso do recém-nascido ou decresce 0,462 para cada a créscimo de 3 horas na duração para a dilatação. O aumento de logit está associado com o aumento da probabilidade de sucesso (apgar ≥7). Então, aumentar o peso do recém-nascido significa aumentar esta probabilidade e o mesmo acontece com a variável amenorréia. O inverso é verdade para as outras variáveis.

Esta relação entre o logit e a probabilidade de sucesso não é linear, portanto um acréscimo de uma unidade no logit não aumentará esta probabilidade em uma unidade.

O desvio para o ajuste é 215,39 com 216 graus de liberdade e a correspondente estatística qui-quadrada é 235,0, ambas com probabilidade de significância em torno de 0,5 indicando que o modelo ajustado é bastante razoável. O histograma das probabilidades preditas de sucesso para o grupo sucesso e para o grupo falha são mostradas nas Figs. 1 e 2, respectivamente. Nota-se que as probabilidades de sucesso estimadas nos casos onde o valor observado foi favorável concentram-se, como seria de esperar, próximos de um. Entretanto, as probabilidades estimadas nos casos onde ocorreram falhas, apresentam um comportamento mais irregular distribuindo-se por todo o intervalo (0,1).

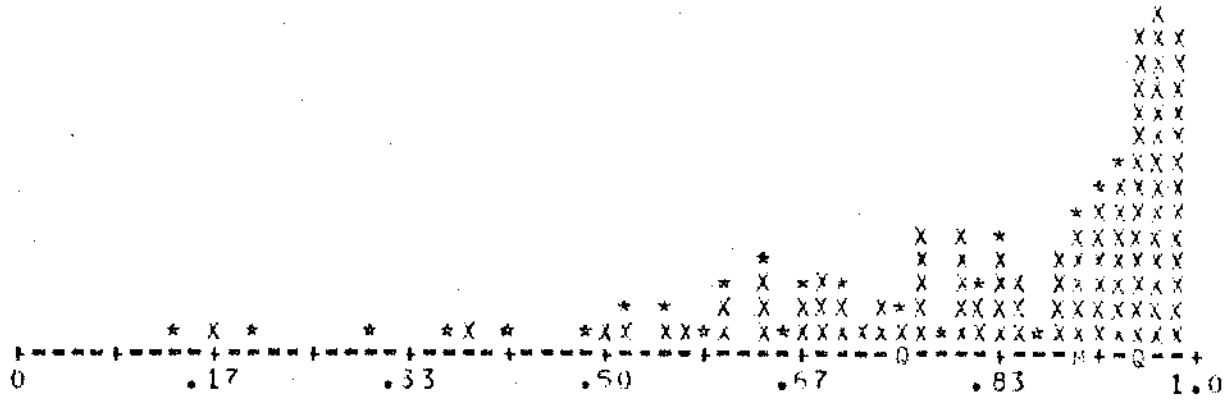

Fig. 1 - Histograma das probabilidades estimadas para o grupo sucesso. Cada "x" representa 2 respostas, "*" representa mais de 2, "M" indica a mediana e "Q" indica os quartis.

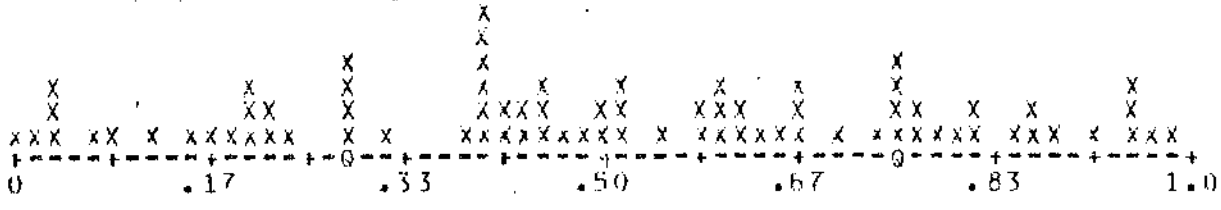

Fig. 2 - Histograma das probabilidades estimadas para o grupo falha. Cada "x" representa 2 respostas, "M" indica a mediana e "Q" indica os quartis.

Uma forma interessante de usar as probabilidades estimadas de sucesso na previsão da variável é classificar os casos, segundo pontos de cortes arbitrários, em grupos da seguinte forma:

Um caso estará no grupo falha se sua probabilidade estimada de sucesso for menor que o ponto de corte.

Um caso estará no grupo sucesso se sua probabilidade estimada de sucesso for maior que este mesmo ponto de corte.

Cada ponto de corte então classifica os valores observados e os estimados na seguinte tabela:

		Predito	
		Sucesso	Falha
Observado	Sucesso	A	B
	Falha	C	D

As frequências A e D são as corretas e as B e C são as incorretas. As percentagens das predições corretas de sucesso (A/A+B) de falha (D/C+D) e total (A+D/A+B+C+D) aparecem na Fig.3 para vários pontos de corte. Observa-se que as maiores porcentagens corretas total estão na faixa de 0,4 a 0,65 do ponto de corte. No ponto de corte 0,74, as porcentagens corretas de sucesso e falha são praticamente iguais, em torno de 75%.

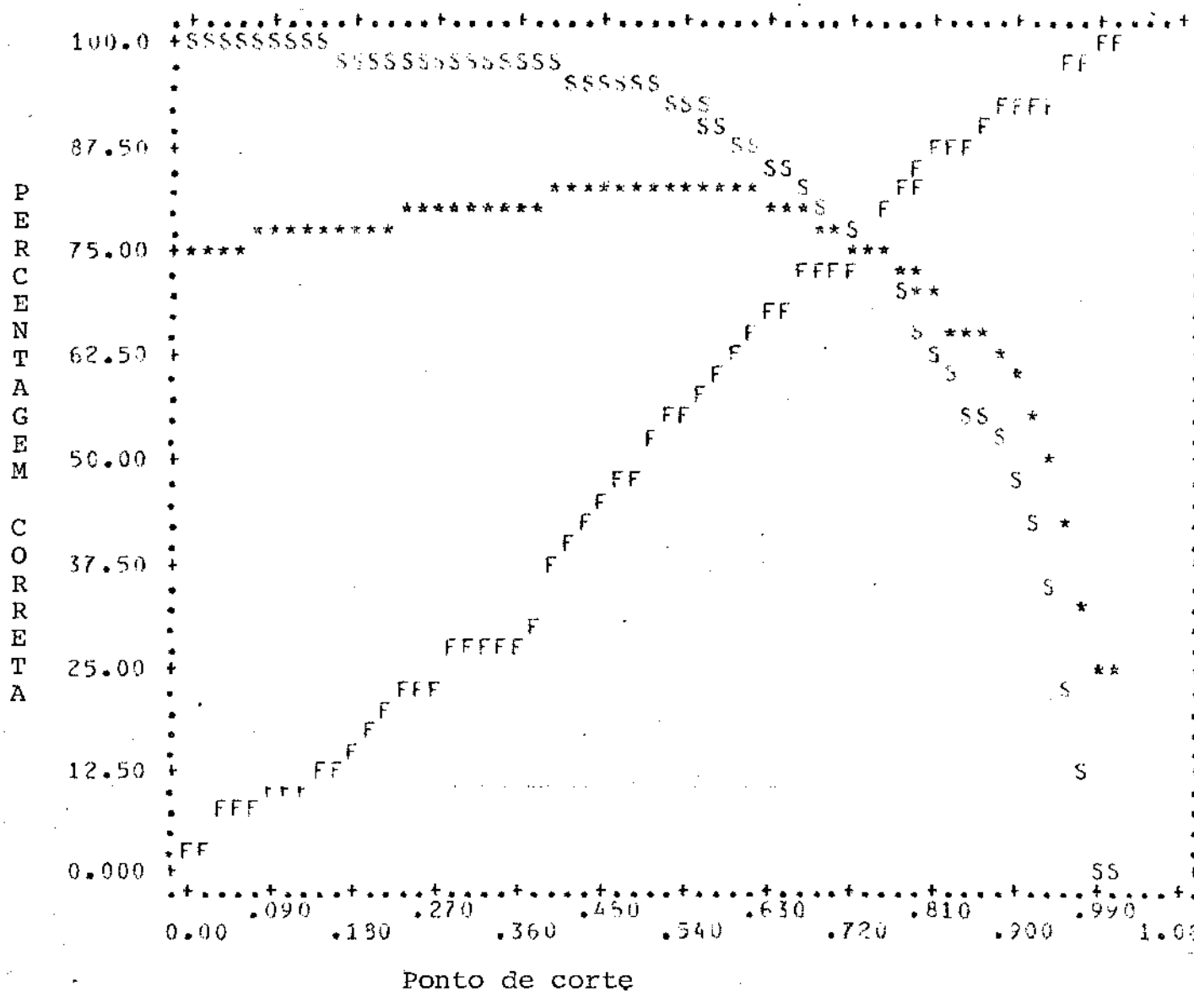

Fig. 3 - Classificação da percentagem correta como função do ponto de corte, onde "S" é a de sucesso, "F" é a de falha e "*" é a total.

3.3. MEDIDAS DE DIAGNÓSTICO

Na seção anterior o modelo de regressão logística foi ajustado para o arquivo de primíparas com apresentação pélvica . Nesta seção iremos criticar tal ajuste usando as técnicas de diagnóstico apresentadas no Capítulo 2. Os resultados de tais técnicas serão apresentados em gráficos abaixo.

É importante notar que na realidade estamos tratando com 222 casos e não 306, pois alguns casos são na verdade observações múltiplas. Todas observações que apresentem a mesma combinação das variáveis explicativas formam um único caso. Por exemplo, em um estudo onde as variáveis são idade, discretizada da forma apresentada, e raça, uma mulher branca com 21 anos e outra branca com 22 anos têm a mesma combinação de variáveis explicativas e portanto formam um único caso.

Inicialmente os resíduos que são os componentes do qui-quadrado e os do desvio são apresentados nas Figs. 1 e 2, respectivamente. No gráfico dos componentes do qui-quadrado dois pontos, as observações de número 37 e 151, se destacam muito dos demais. Outros quatro, as observações de número 27, 53, 133 e 146, tem um ligeiro destaque. O valor do componente do qui-quadrado para tais pontos são os seguintes:

$X_{37} = -4,15$ $\qquad$ $X_{27} = -3,03$

$X_{151} = -4,97$ $\qquad$ $X_{53} = -2,69$

$X_{133} = 2,54$

$X_{146} = -2,73$

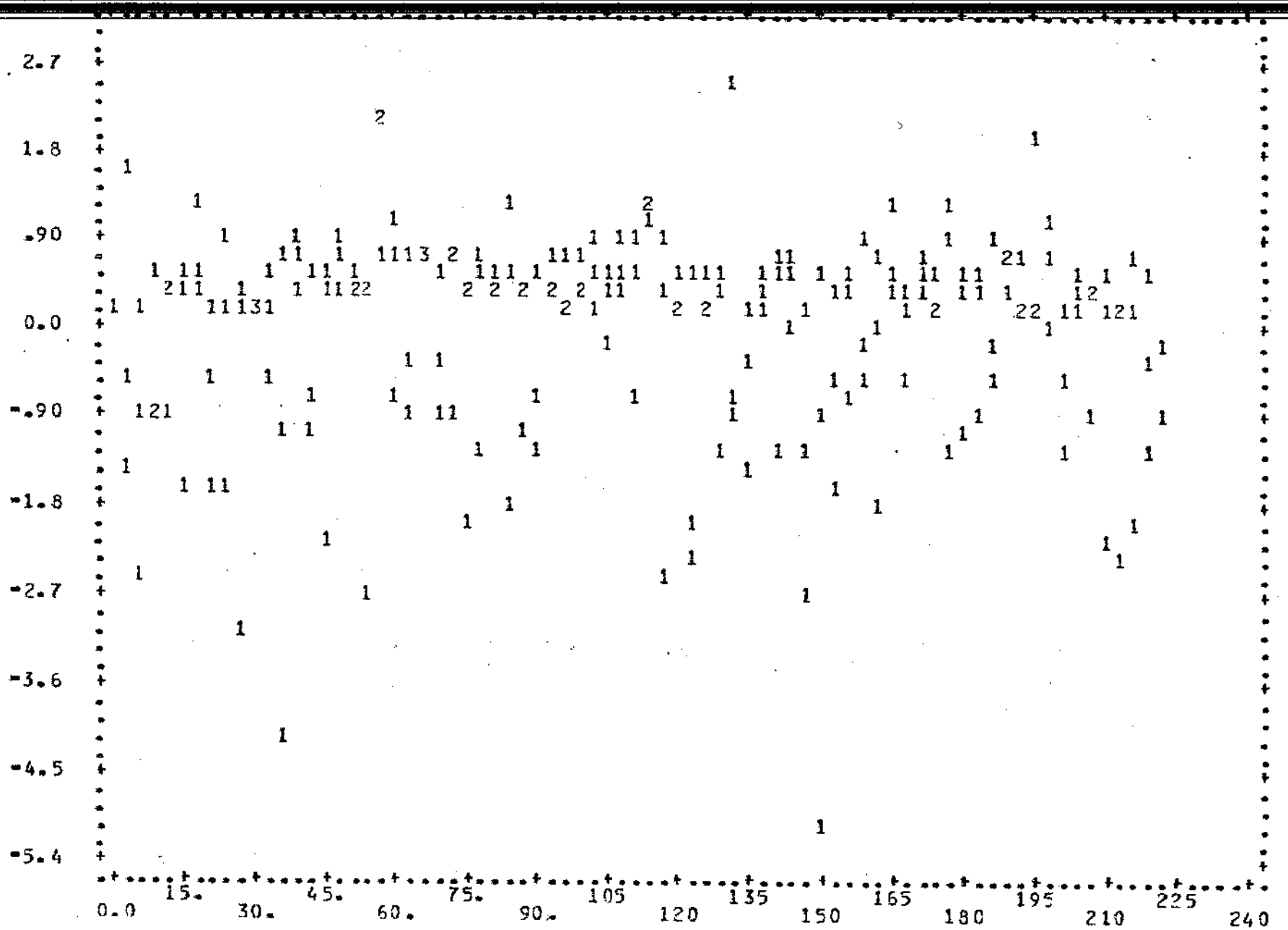

Fig. 1 - Componentes do Qui-Quadrado (χ_j) vs j

No gráfico dos componentes dos desvios existe um maior espalhamento dos pontos e nenhum tem grande destaque. Entretanto, dentre eles podemos notar seis, as observações de nºs 27,37,124,133, 146 e 151 que são os que apresentam maiores componentes, cujos valores são os seguintes:

$d_{27} = -2,10$ $\qquad$ $d_{133} = 2,01$

$d_{37} = -2,08$ $\qquad$ $d_{146} = -2,07$

$d_{124} = -2,11$ $\qquad$ $d_{151} = -2,55$

Observe que cinco pontos de destaque são comuns aos dois componentes e o 151 é o maior nos dois casos.

O gráfico dos elementos da matriz de projeção está na Fig. 3, destes pontos seis, as observações de número 4,9,59,65 , 91 e 208, sobressaem dos demais. Os valores destes pontos são os seguintes:

$h_{44} = 0,124$ $\qquad$ $h_{6565} = 0,136$

$h_{99} = 0,135$ $\qquad$ $h_{9191} = 0,143$

$h_{5959} = 0,120$

As medidas de diagnóstico que vêm a seguir quantificam o efeito da retirada de um caso sobre determinados aspectos do ajuste. Os seis primeiros gráficos, nas Figs. 4,5,6,7,8 e 9 medem este efeito sobre os coeficientes estimados.

Um ponto, o de número 27, em particular, causa um gran

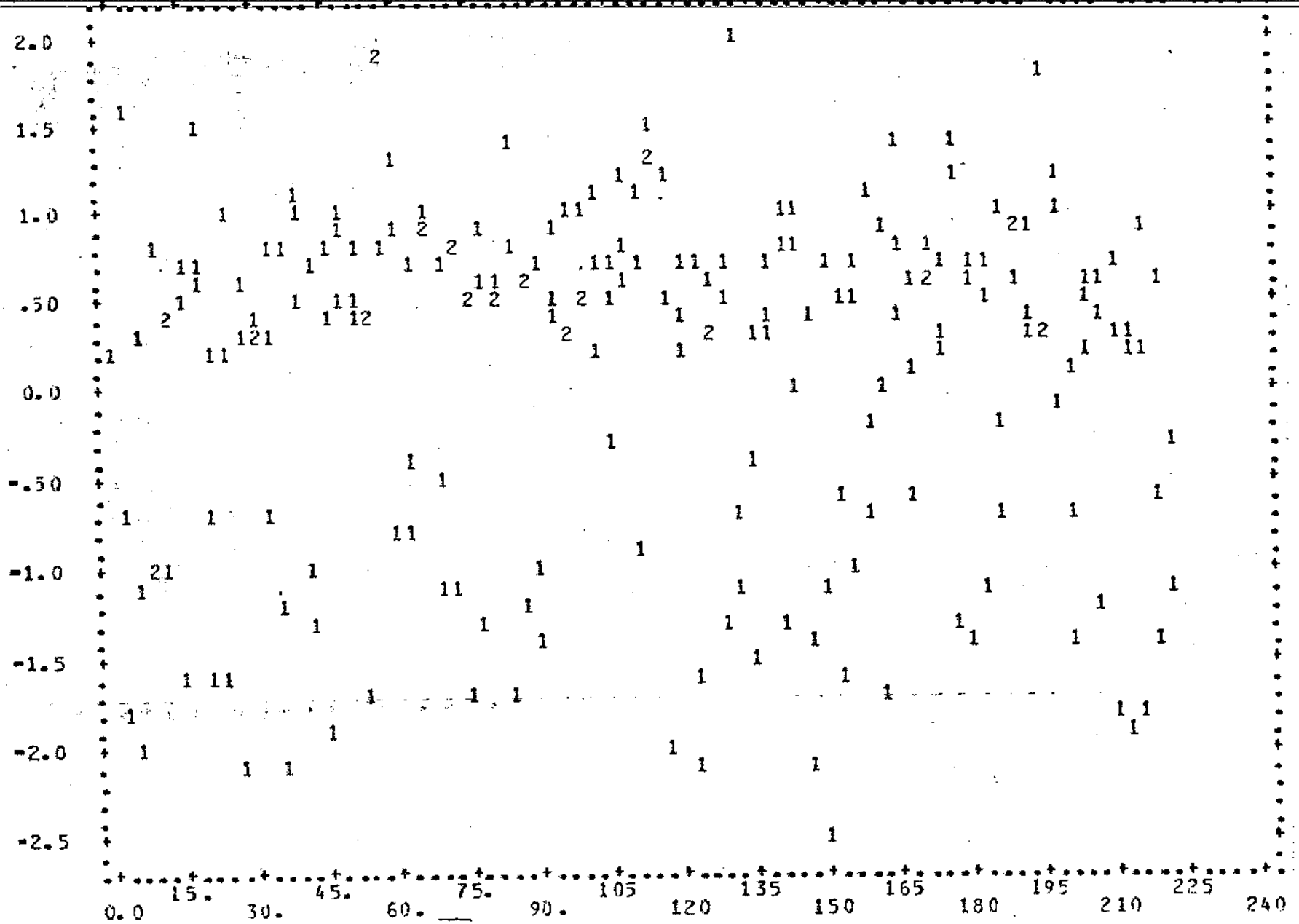

Fig. 2 - Componentes do desvio (d_j) vs j

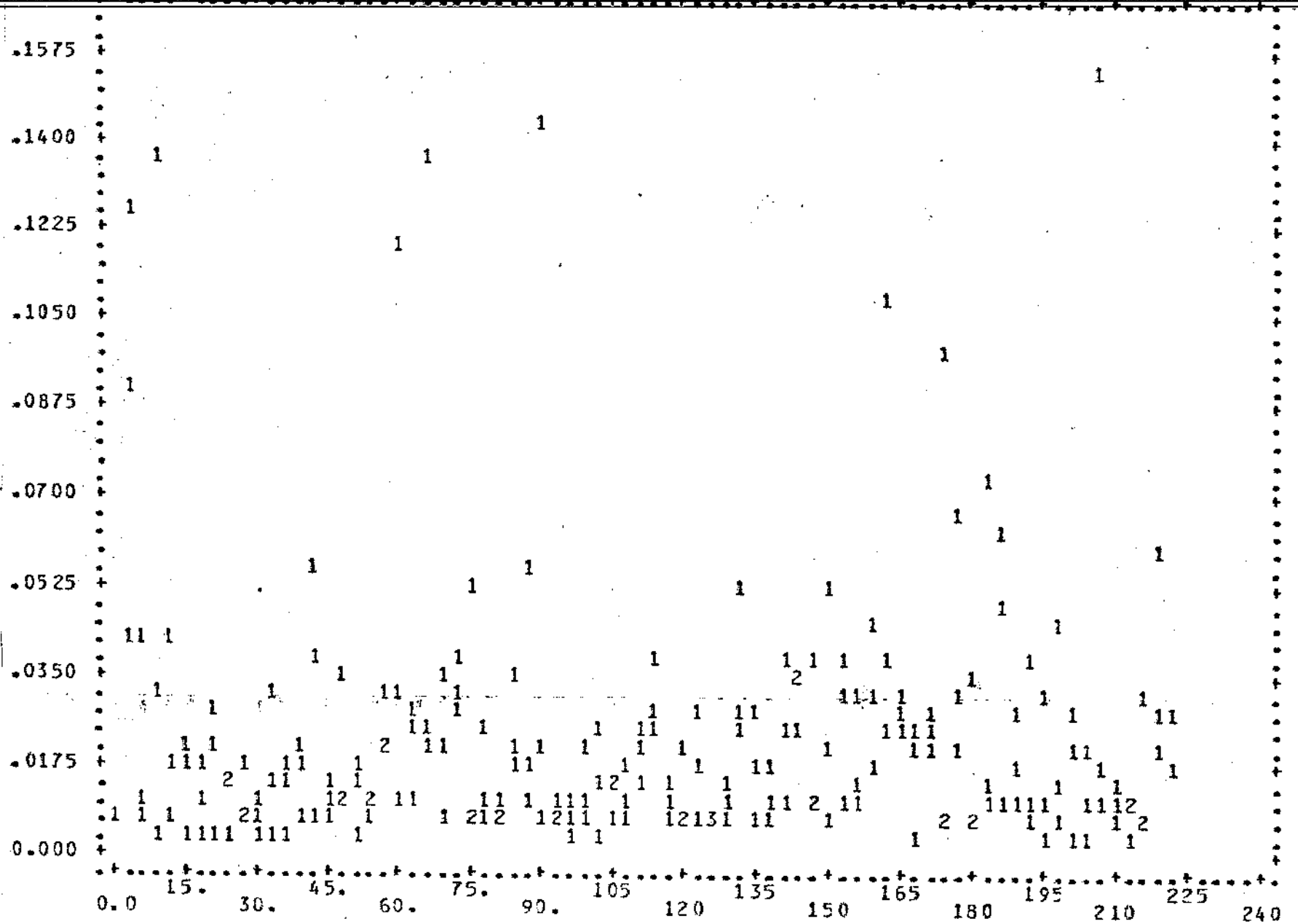

Fig. 3 - Elementos da diagonal da matriz de projeção (h_{jj}) vs j

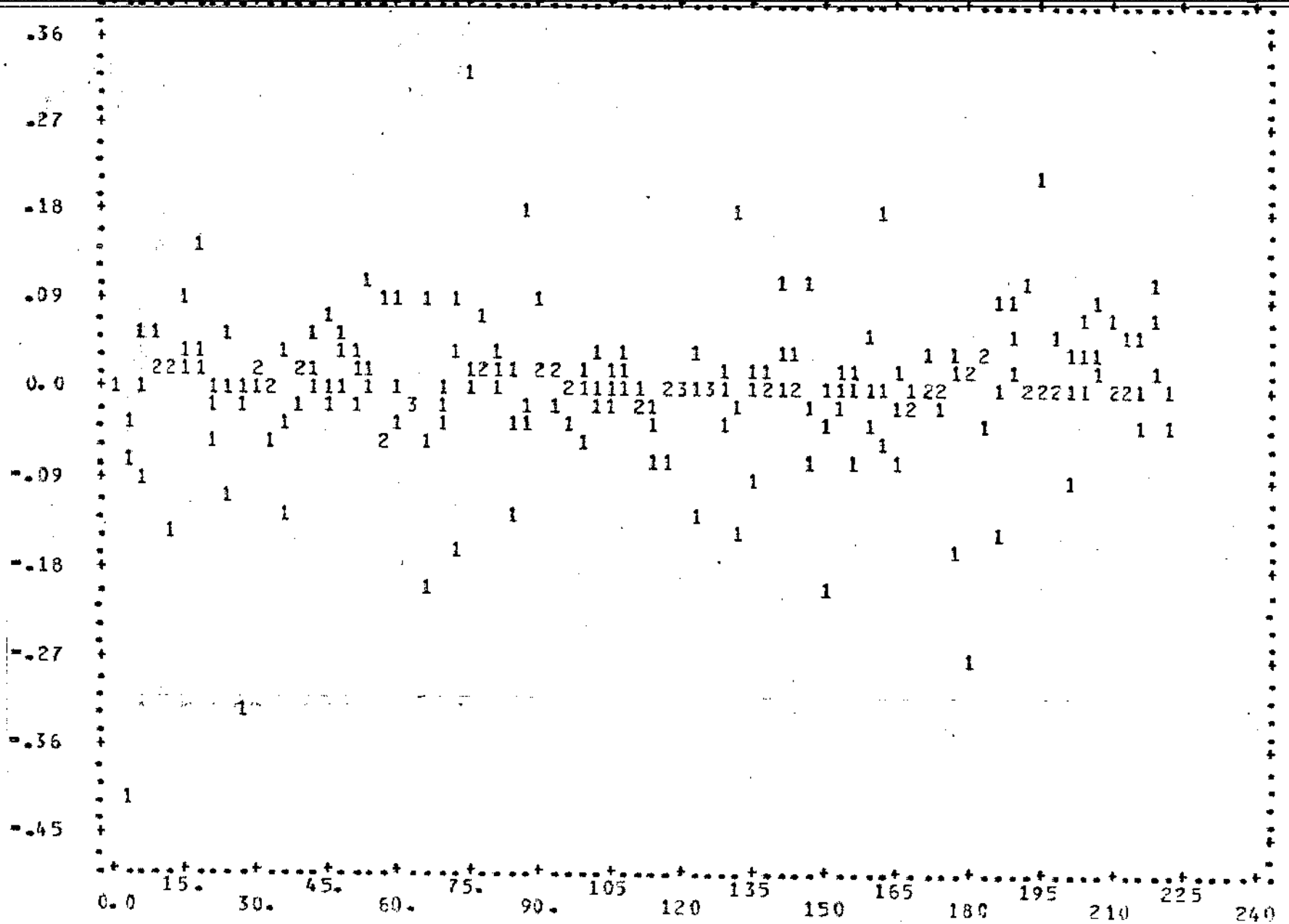

Fig. 4 - Variação padronizada em $\hat{\beta}_0$ ($\Delta_j \hat{\beta}_0$) vs j

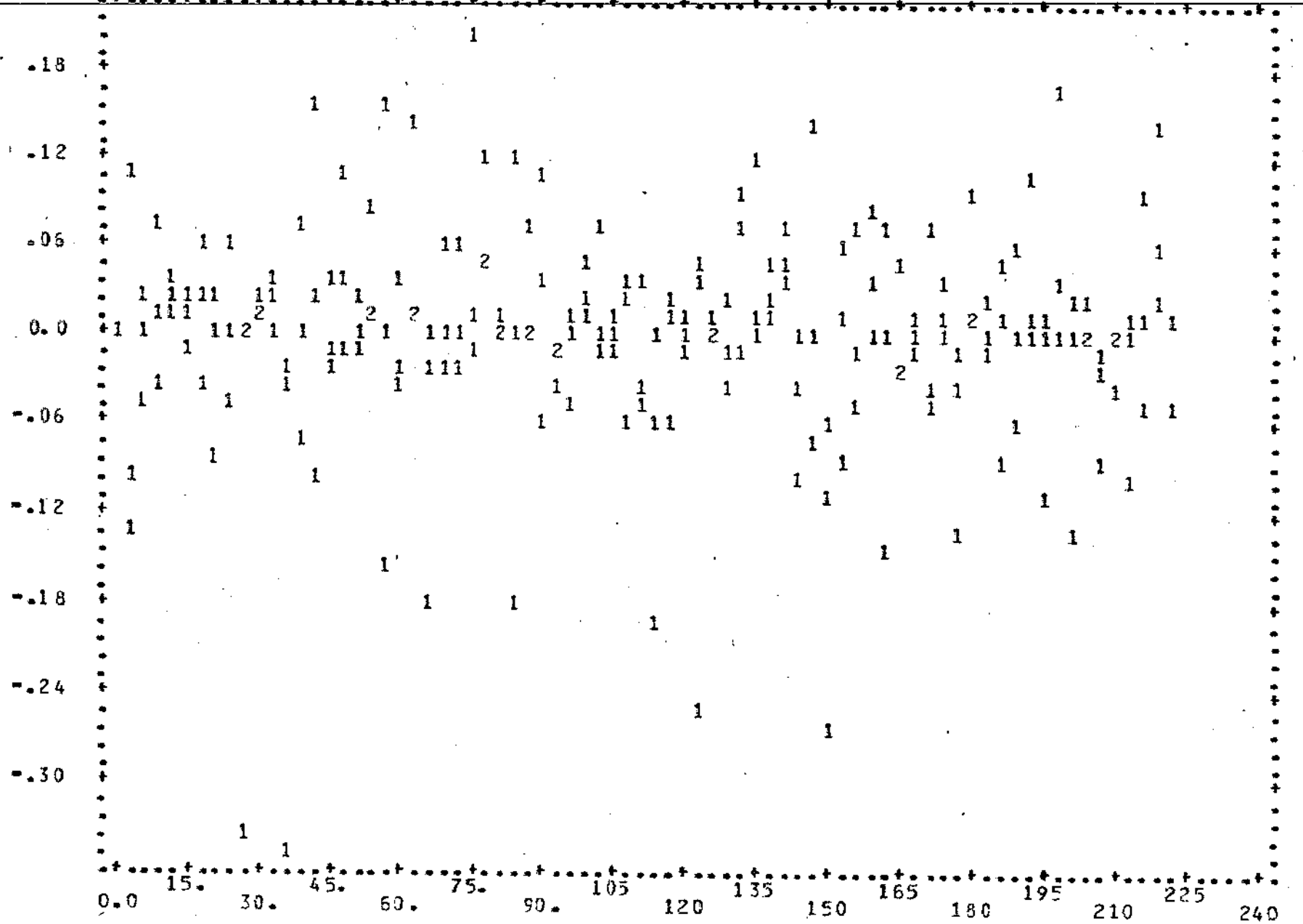

Fig. 5 - Variação padronizada em $\hat{\beta}_1$ ($\Delta_j \hat{\beta}_1$) vs j

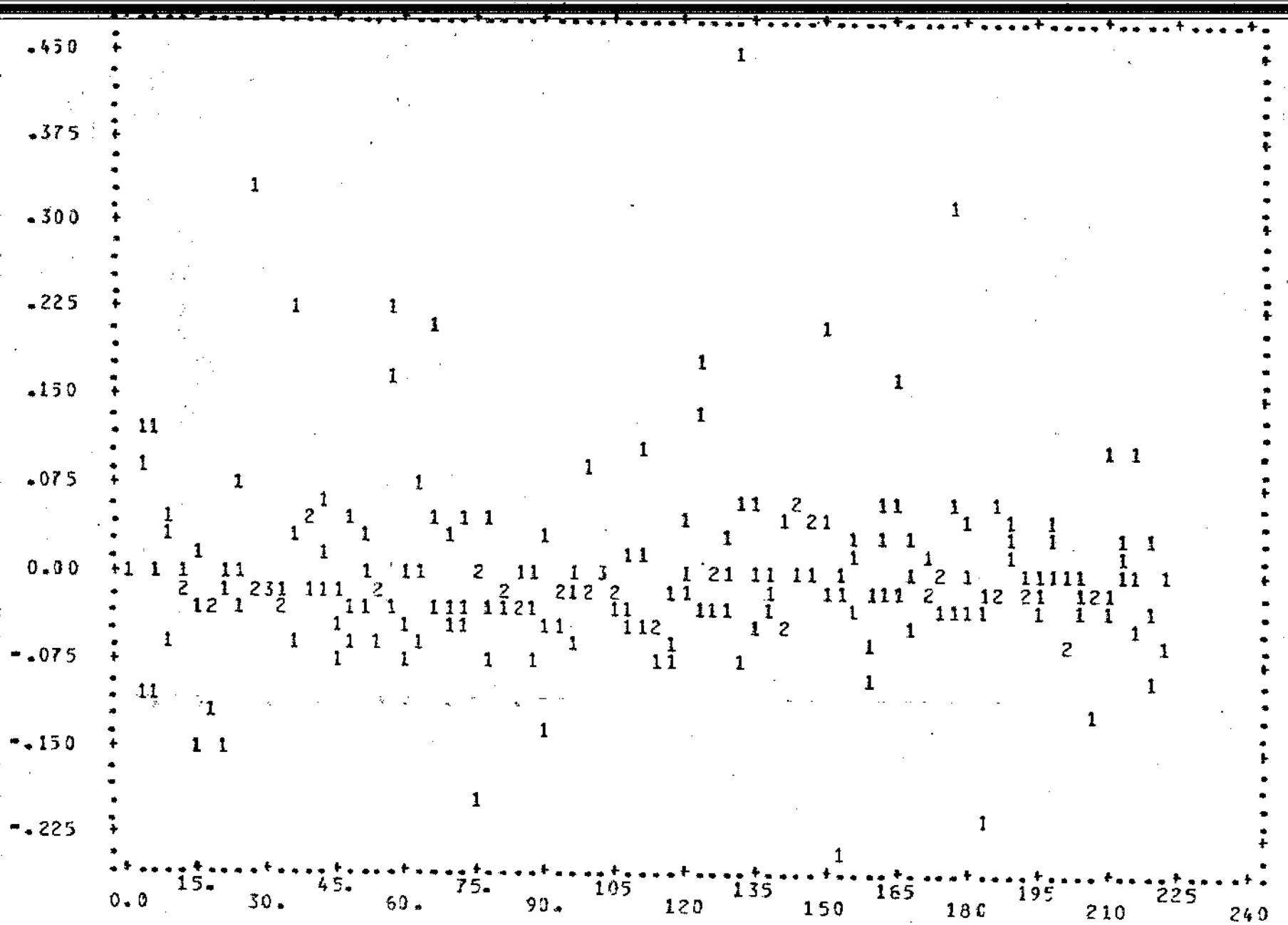

Fig. 6 - Variação padronizada em $\hat{\beta}_2$ ($\Delta_j \hat{\beta}_2$) vs j

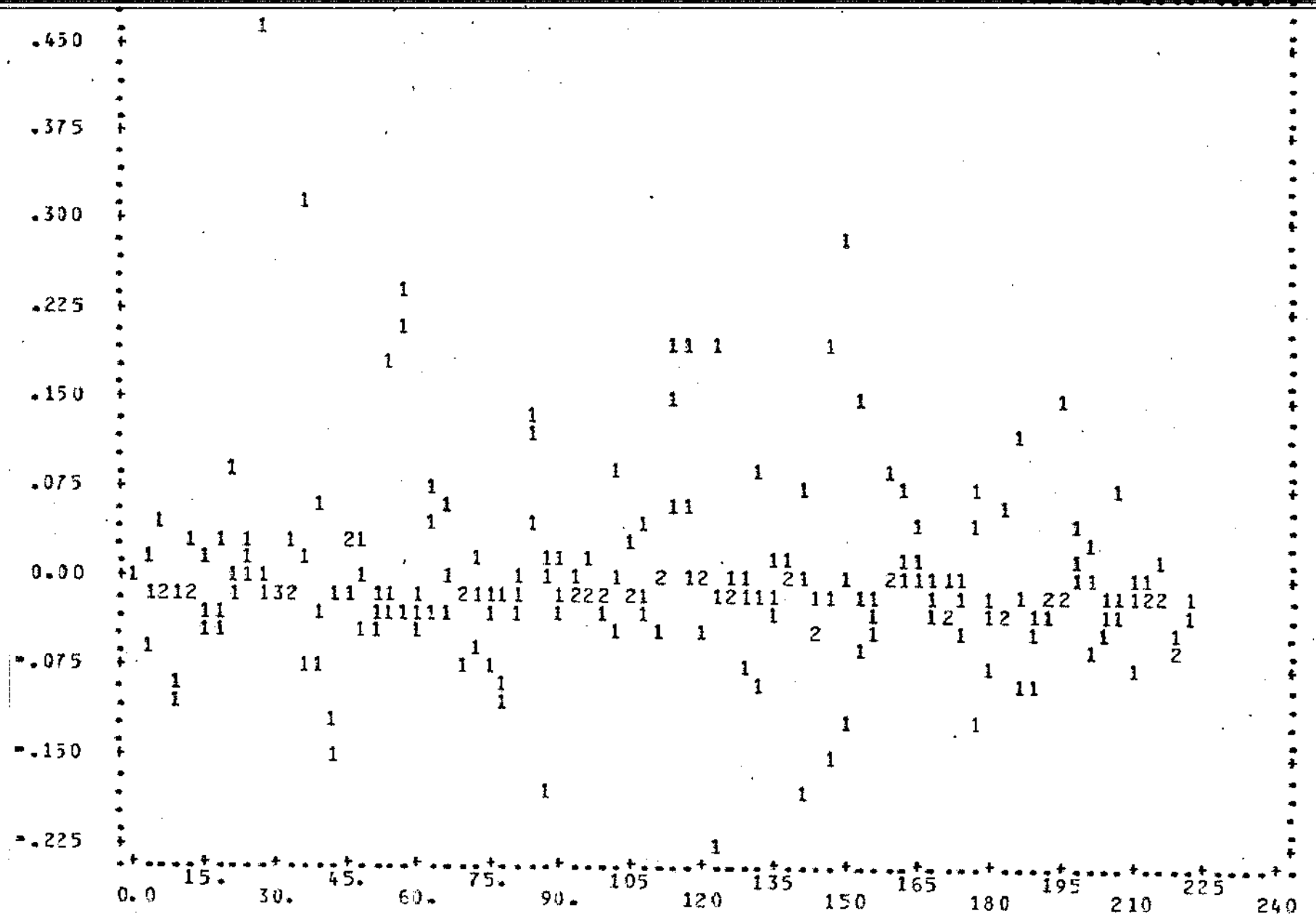

Fig. 7 - Variação padronizada em $\hat{\beta}_3$ ($\Delta_j \hat{\beta}_3$) vs j

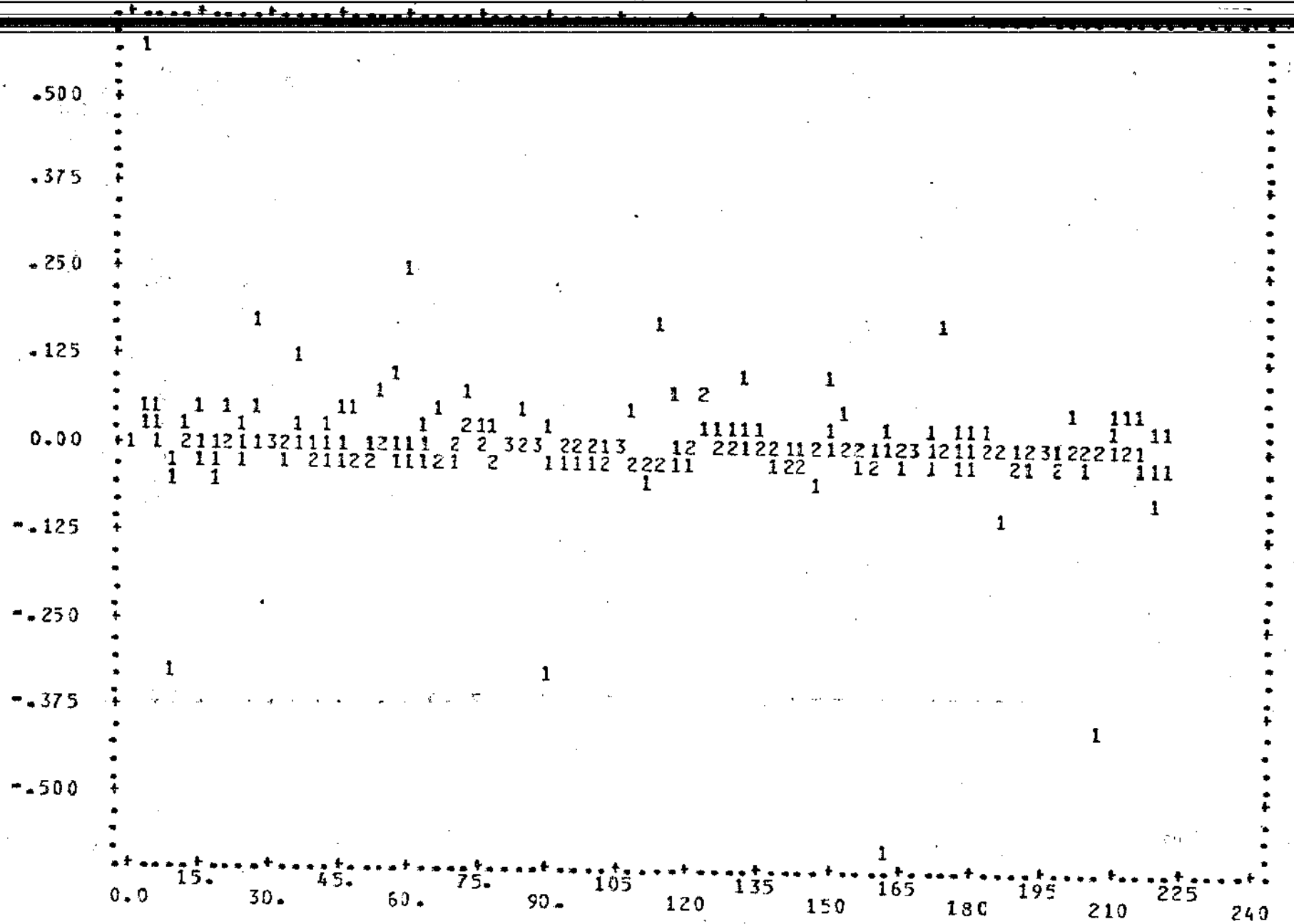

Fig. 8 - Variação padronizada em $\hat{\beta}_4$ ($\Delta_j \hat{\beta}_4$) vs j

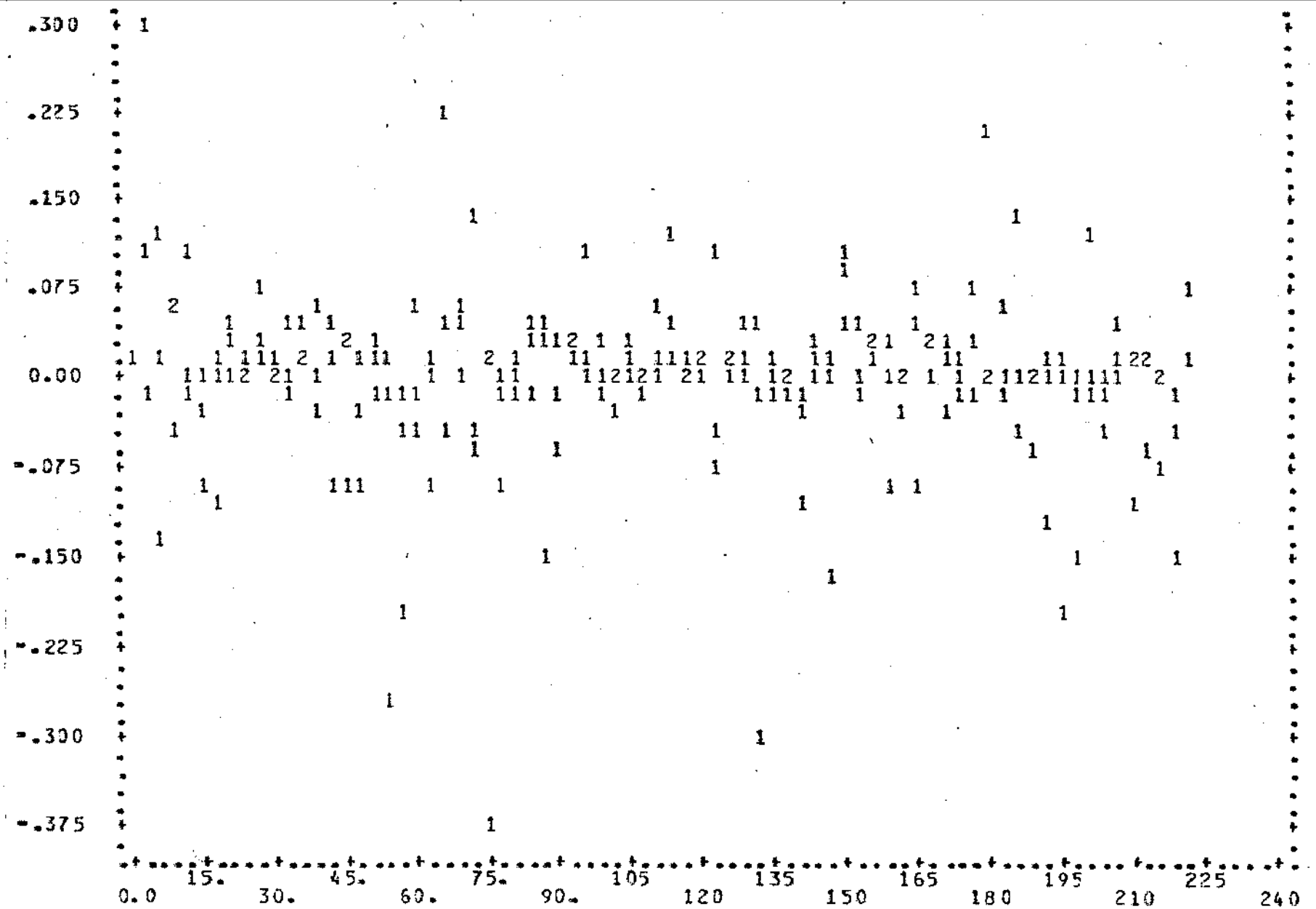

Fig. 9 - Variação padronizada em $\hat{\beta}_5$ (Δ_j $\hat{\beta}_5$) vs j

	3	4	9	27	37	53	74	91	124	133	151	161	176	179	208
β_0	−0,42			−0,32			0,33							−0,27	
β_1				−0,33	−0,34				−0,25		−0,27				
β_2				0,33						0,45			0,31		
β_3				0,47	0,31						0,29				
β_4		0,57	−0,33					−0,33				−0,57			−0,40
β_5	0,30					−0,27	0,38			−0,30					

de impacto sobre quatro coeficientes estimados. Este ponto se sobressae por causar impacto sobre o maior número de coeficientes. A tabela abaixo resume os pontos mais marcantes.

	3	4	9	27	37	53	74	91	124	133	151	161	176	179	208
β_0	-0,42			-0,32			0,33							-0,27	
β_1				-0,33	-0,34				-0,25		-0,27				
β_2				0,33						0,45			0,31		
β_3				0,47	0,31						0,29				
β_4		0,57	-0,33					-0,33				-0,57			-0,40
β_5	0,30					-0,27	0,38			-0,30					

Outros cinco pontos, os de número 3,37,74,133 e 151 se destacam por causar impacto sobre dois coeficientes. Observa-se, também, a grande variação causada sobre β_4 pelos pontos de número 4 e 161.

Outras duas medidas, c_i e $\overline{c}_i$, que quantificam o efeito da retirada de cada observação simutaneamente sobre todos os coeficentes estimados são apresentadas nas Figs. 10 e 11, respectivamente. Os pontos de número 4,56,74,133,151,161 e 208 se destacam em ambas as medidas, cujos valores são:

Pontos	c	$\overline{c}$
4	0,39	0,34
56	0,17	0,16
74	0,21	0,20
133	0,37	0,35
151	0,17	0,17
161	0,41	0,36
208	0,21	0,18

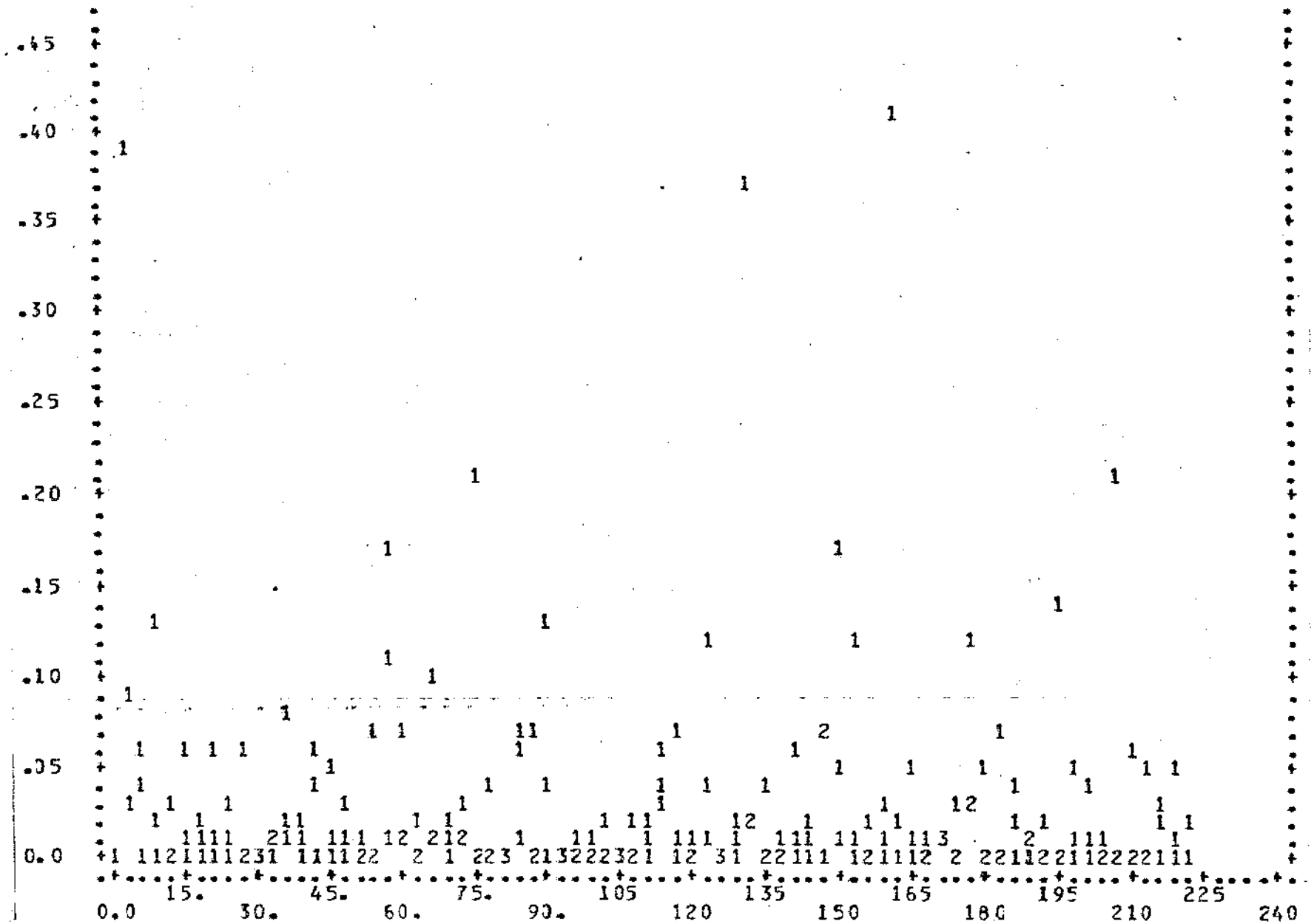

Fig. 10 - Deslocamento da região de confiança (ε_j) vs j

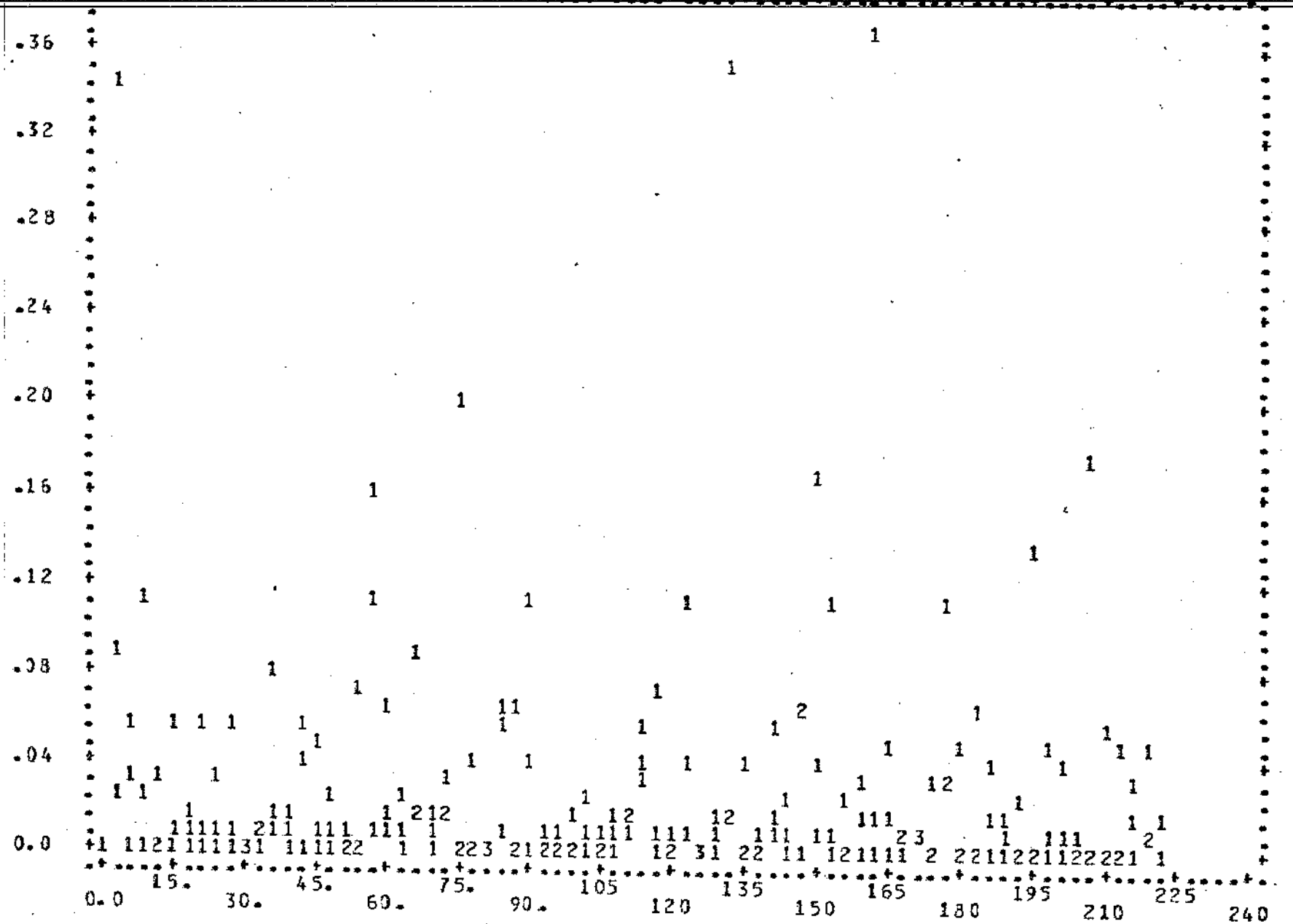

Fig. 11 - Deslocamento da região de confiança ($\bar{c}_j$) vs j

Finalizando as medidas de diagnósticos, o efeito da variação nas estatísticas de adequação do modelo χ^2 e D, devido a retirada de uma observação aparece nas Figs. 12 e 13, respectivamente. Dois pontos, os de números 37 e 151, se destacam muito dos demais e outros 7 também têm um ligeiro destaque. Os valores destes pontos são os seguintes:

Pontos	χ^2	D	Pontos	χ^2	D
37	17,35	4,41	146	7,55	4,34
151	24,90	6,66	124	4,26	4,61
7	6,30	4,01	53	7,29	3,13
27	9,28	4,47			
117	6,58	4,10			
133	6,82	4,37			

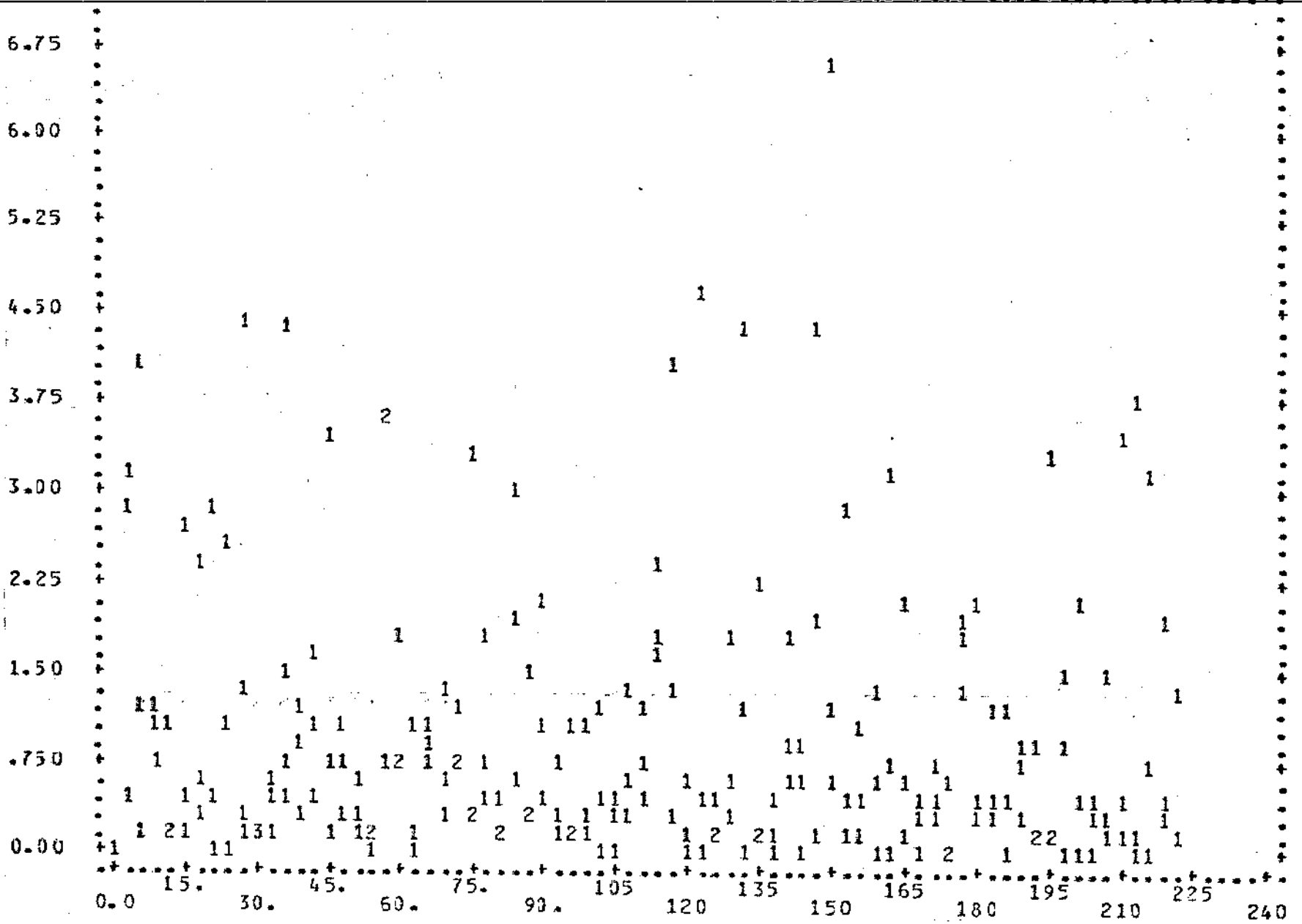

Fig. 13 - Variação no desvio ($\Delta_j D$) vs j

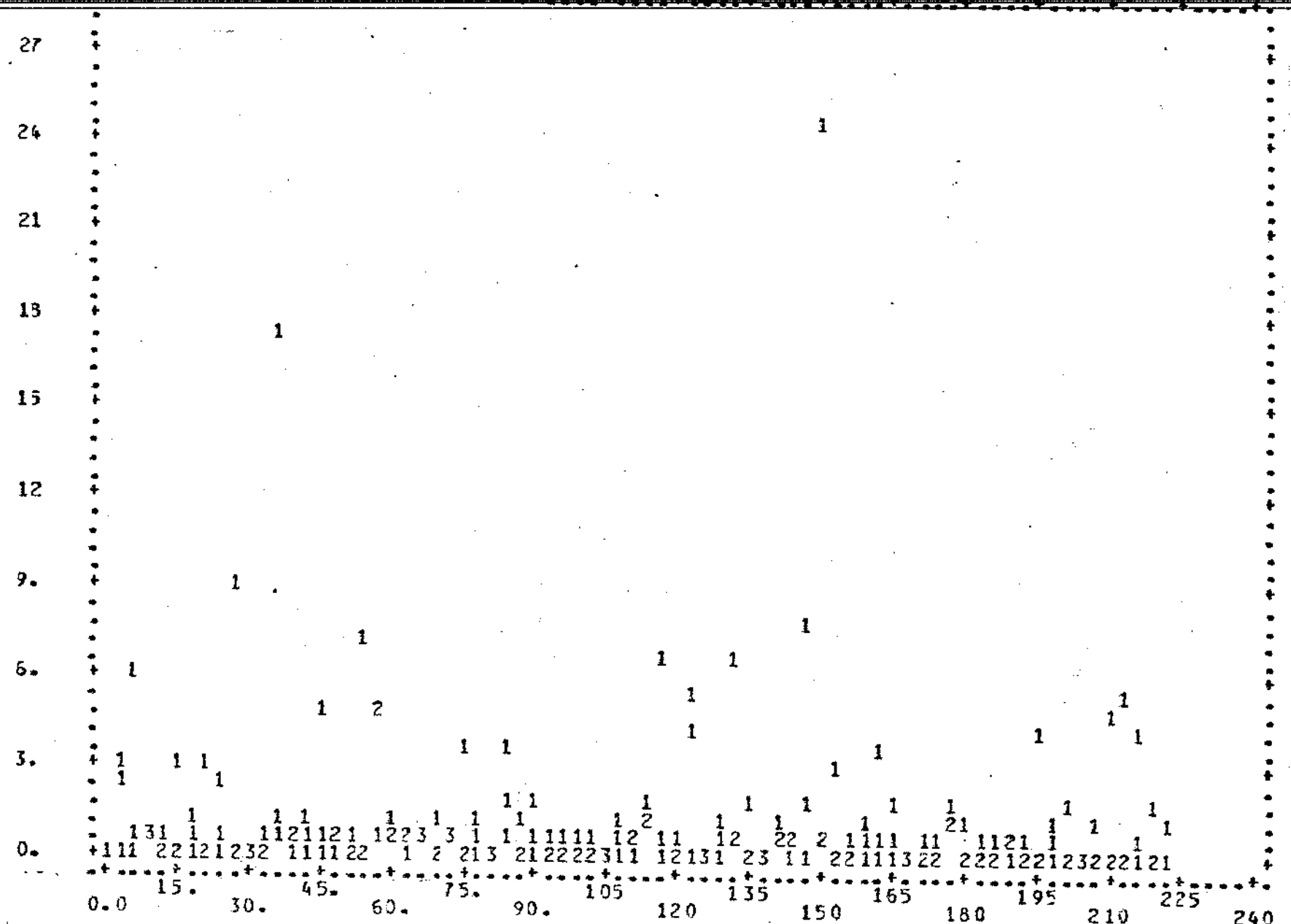

Fig. 12 - Variação na estatística Qui-Quadrado ($\Delta_j \chi^2$) vs j

3.4. DISCUSSÃO FINAL

O modelo ajustado para este estudo está explicando de forma bastante razoável os valores observados. Indicações disso são as estatísticas de adequação do modelo que deram valores relativamente altos para as probabilidades de significância. A qualidade do ajuste pode ser considerada particularmente boa se considerarmos que estamos analisando um estudo observacional, conduzido sem o rigor de um experimento controlado. Além disso os dados foram levantados por diferentes profissionais, usando critérios cuja uniformidade é essencialmente limitada. Finalmente, é de se notar que as observações foram obtidas durante um período relativamente longo, de quase dois anos.

Os fatores de risco incluídos no modelo confirmam em linhas gerais hipóteses anteriormente levantadas clinicamente.

A anemorreia é o número de semanas que a gestante ficou sem mesntruação. A partir dela temos, então, uma indicação da idade do feto e nesse sentido mede o grau de desenvolvimento do recém-nascido. É de se esperar que quanto maior o valor desta variável maior em geral será o valor do APGAR, o que é comprovado pelo modelo ajustado.

O peso do recém-nascido indica como se deu o processo de desenvolvimento do feto no útero da mãe. Alto peso* está frequentemente associado a boa saúde ou APGAR alto, pode-se obser

---

* Até um certo ponto, valores extremamente alto do peso do bebê estão associados com os diabetes.

var que o coeficiente para a variável é, além de positivo, grande comparado com os demais.

A duração para a dilatação mede, é claro, o tempo necessário para haver dilatação do colo útero. Um tempo elevado acarreta numa redução de oxigênio no sangue. Isto faz com que o intestino sofra contração e comece a surgir mecônio (fezes do feto) no líquido ovular o que implica numa série de consequências adversas. O coeficiente desta variável é negativo concordando com o que foi exposto acima.

O estado do líquido ovular caracteriza possíveis agravamentos na saúde do recém-nascido. A presença de mecônio foi discutida acima. Outras situações mais graves são as presenças de pus, hemorragia e infecção. Observa-se que os valores das categorias desta variável crescem com a gravidade do estado do recém-nascido o que concorda com o sinal do coeficiente estimado.

As características do cordão umbilical são as posições do mesmo com relação ao corpo do recém-nascido. Várias posições levam ao sufocamento e consequentemente a uma abrupta redução dos níveis de oxigênio no sangue do bebê.

Por outro lado, dos resultados pelas medidas de diagnóstico na seção 3.3, quatro dos casos se sobressaem em quase todos os gráficos, são eles os de número 27, 37, 133 e 151. Uma descrição destes casos, incluindo os valores assumidos pelas variáveis explicativas correspondentes, segue abaixo.

CASO	$\hat{\theta}i$	ni	Yi	Xi_1	Xi_2	Xi_3	Xi_4	Xi_5
27	0,954	7	5	6	1	1	1	5
37	0,982	3	2	9	1	1	1	6
133	0,134	1	1	2	6	1	2	2
151	0,961	1	0	7	1	1	1	5

onde,

Xi_1 é amenorreia,

Xi_2 é estado do líquido ovular,

Xi_3 é duração para dilatação,

Xi_4 é características do cordão,

Xi_5 é peso do recém-nascido.

Nos casos de número 27, 37 e 151 temos amenorreia e peso com valores altos e as outras três variáveis com valores baixos o que caracteriza pelo ajuste alta probabilidade de APGAR $\geq$ 7 e isto é comprovado pelos valores encontrados para $\hat{\theta}i$. No entanto, no caso 27, duas gestantes apresentam APGAR $\leq$ 6. No 37 e no 151 uma gestante tem APGAR $\leq$ 6. No outro caso, o 133, acontece a situação contrária, ou seja, amenorreia e peso com valores baixos e a duração para dilatação com valor alto e o valor observado para a gestante foi APGAR $\geq$ 7. Estes quatro casos são "OUTLIERS" dentro do ajuste.

Uma possível explicação para alguns destes casos é que os valores do APGAR estão próximos do ponto de corte. Isto acon

tece para uma gestante do caso 27 e do caso 151 em que os valores do APGAR são seis e no caso 133 que é sete.

Com relação a pontos extremos, vimos que seis casos mereceram destaque no gráfico de h_{ii} versus i. Entretanto, nenhum destes casos teve realce nas outras medidas de diagnóstico, significando que retiradas únicas destes casos não causaram impacto no ajuste e consequentemente suas influências são pequenas. Uma provável explicação para este fato, é estarmos lidando com variáveis categóricas que assumem somente um dígito num estudo com muitos casos. Isto faz com que haja um espalhamento por quase todas as categorias das variáveis presentes no ajuste.

Mesmo o ajuste estando adequado, os quatro casos mencionados merecem uma atenção especial. Um exame cuidadoso deve ser feito nos mesmos.

Dentre as variáveis que não foram identificadas como fatores de risco, aparecem algumas que segundo considerações clínicas seriam importantes para esta resposta. Por exemplo, idade da mãe. Entretanto, como estamos lidando com primíparas (primeiro parto) a idade destas mulheres concentra-se numa faixa estreita em torno de 20 anos, o que explica a ausência desta variável como fator de risco. Outras devem estar sendo explicadas pelas que estão presentes, assim como discutido na seção 1.5.

A não inclusão da forma de término do parto como uma das variáveis explicativas é razoavelmente surpreendente do ponto de vista médico. Acredita-se que na maioria dos casos de partos pél

vicos teríamos um prognóstico significativamente melhor caso se opte pela cesária. O efeito da forma de término é entretanto medido de maneira indireta pelas variáveis duração da dilatação e características do cordão. Ambas as variáveis tem coeficientes estimados negativos e assumem os valores mínimos quando a forma de término é cesária.

APÊNDICE I

CONFIGURAÇÃO DAS VARIÁVEIS

1. IDADE

1 - 14 ou (-)	4 - 25 - 29	7 - 40 - 44
2 - 15 - 19	5 - 30 - 34	& - Ignorado
3 - 20 - 24	6 - 35 - 39	

2. ESTADO CIVIL

0 - Com companheiro	& - Ignorado
1 - Sem companheiro	

3. TOTAL DE ABORTOS

0 - 4	& - Ignorado

4. ANTECEDENTES OBTÉTRICOS

0 - Com	& - Ignorado
1 - Sem	

5. ANTECEDENTES MÓRBIDOS

0 - Com	& - Ignorado
1 - Sem	

6. AMENORRÉIA (em semanas)

0 - 34 ou (-)
1 - 35
2 - 36
3 - 37
4 - 38
5 - 39
6 - 40
7 - 41
8 - 42
9 - 43 ou (+)
& - Ignorado

7. PRÉ-NATAL

1 - Controle bom
2 - Controle mau
3 - Sem controle
& - Ignorado

8. TIPO DE ROTURA DA MEMBRANA

1 - Rotura Tempestiva ou Cesárea
2 - Rotura precoce
3 - Rotura prematura ou alta
& - Ignorado

9. ESTADO DO LÍQUIDO OVULAR

1 - Claro
2 - Ligeiramente mecônio
3 - Tingido mecônio
4 - Espesso mecônio
5 - Mecônio antigo
6 - Hemorrágico ou infeccioso ou purulento
& - Ignorado

10. QUANTIDADE DO LÍQUIDO OVULAR

1 - Normal
2 - Polidramnio
3 - Oligodramnio
& - Ignorado

11. Nº DE TOQUES DESDE RÓTURA DA MEMBRANA

1 - 0 ou 1 ou Cesárea
2 - 9
& - Ignorado

12. ANALGESIA TRABALHO DE PARTO

1 - Sem
2 - Demerol ou Opiceos ou Morfina
3 - Epidoral ou Caudal
4 - Raquídea
& - Ignorado

13. FORMA DE INÍCIO

0 - Espontâneo ou Indução Ocitócica
1 - Cesárea eletiva
& - Ignorado

14. FORMA DE TÉRMINO

1 - Espontâneo
2 - Assist. Pélvico
3 - Ext. Pélvica ou Fórcipe
4 - Cesárea
& - Ignorado

15. APRESENTAÇÃO

1 - Pélvica completa
2 - Pélvica incompleta
3 - Pélvica alta

15. POSIÇÃO

1 - EA ou DA
2 - ET ou DT
3 - EP ou DP
4 - Cesárea
& - Ignorado

17. DURAÇÃO PARA DILATAÇÃO

1 - Sem dilatação
2 - 3 hs. ou (-)
3 - 3 - 6
4 - 6 - 9
5 - 9 - 12
6 - 12 - 18
7 - 18 ou (+)
& - Ignorado

18. DURAÇÃO PERÍODO EXPULSIVO (em minutos)

0 - Cesárea
1 - 5' ou (-)
2 - 5' - 10'
3 - 10' - 20'
4 - 20' - 30'
5 - 30' - 45'
6 - 45' - 60'
7 - 60' - 75'
8 - 75' ou (+)
& - Ignorado

19. TIPO DE TÉRMINO

0 - Sem cesárea
1 - Cesárea
& - Ignorado

20. ANALGESIA PARA EXPOSIÇÃO OU INTERVENÇÃO

1 - Nenhuma
2 - Pudenda ou local
3 - Epidural ou caudal
4 - Raquidiana
5 - Pentothal
6 - Éter
& - Ignorado

21. CARACTERÍSTICAS DO CORDÃO

1 - Normal
2 - Circular redutível
3 - Circular irredutível
4 - Circular curto
5 - Nó verdadeiro
6 - Procubito ou procidência
& - Ignorado

22. PESO DO RECÉM-NASCIDO

1 - 1 kg ou (-)
2 - 1 - 1,5
3 - 1,5 - 2,0
4 - 2,0 - 2,5
5 - 2,5 - 3,0
6 - 3,0 - 3,5
7 - 3,5 - 4,0
8 - 4,0 - 4,5
9 - 4,5 ou (+)
& - Ignorado

23. APGAR 1'

0 - 8 9 - 9 ou 10 & - Ignorado

24 - ESTADO DO RN NA ALTA DA MÃE

1 - Vivo sadio 3 - Natimorto 5 - Morte 1-7 dias
2 - Vivo doente 4 - Morte neonatal 1ªs 24 hs. 6 - Morte 7-28 dias

25. TOXEMIA E HIPERTENSÃO

1 - Sem

2 - Hipertensão moderada pior toxemia ou toxemia moderada

3 - Hipertensão crônica essencial ou hipertensão outra origem

4 - Hipertensão severa pior toxemia ou toxemia severa

5 - Eclampsia

& - Ignorado

26. INFECÇÕES

0 - Com

& - Ignorado

1 - Sem

27. HEMORRÁGICAS

1 - Sem

2 - Metror. 1º-2º trim.

3 - PP ou PP central

4 - Rut. uterina ou Rut. cicat. cesárea

5 - DPPNI

& - Ignorado

APÊNDICE II

O programa abaixo ajusta por máxima verossimilhança um modelo de regressão logística e calcula todas as medidas de diagnósticos apresentadas no Capítulo 2, usando a aproximação a um passo no processo iterativo.

O leitor deverá ter cuidados ao usar o programa listado abaixo, pois ele não foi generalizado. Por exemplo,a dimensão dos vetores e das matrizes, assim como as repetições do comando DO estão feitas de acordo com o problema tratado no texto. Tais comandos devem então ser modificados de forma a adaptar-se aos dados de cada aplicação em particular.

```
C                    PROGRAMA PRINCIPAL
      DIMENSION X(306,6),Y(306),P(306),S(6),XINF(6,6),A(222),CO(222)
      DIMENSION XQUI(222),T(222),DES(222),CONTA(222),NSR(306),DIF(306)
      DIMENSION TESTE(6),B(6),BT(6),D(222),V(222),R(222),Q(222)
      DIMENSION CHA(222,6),XHAT(222),XPA(222,6),CN(222),CO(222)
      DIMENSION NC(306),NS(306),PN(306),XN(222,6),NCR(306)
      DO 40 I=1,306
      READ(8,41) X(I,2),X(I,3),X(I,4),X(I,5),X(I,6),Y(I)
 41   FORMAT(1(,6I1)
      IF(Y(I).LT.7)Y(I)=0
      IF(Y(I).GT.6)Y(I)=1
 40   CONTINUE
      DO 45 I=1,306
 45   X(I,1)=1.0
      DO 50 I=1,6
 50   B(I)=0
       M=0
      N=306
      K=6
 60   CALL PROB(K,N,B,X,P)
      CALL SCORE(K,N,Y,X,P,S)
      CALL INFO(N,K,X,P,XINF)
      L1=6
      L=6
      LO=0
      CALL SINV(XINF,K,L,LO,PVT,IFAULT)
      CALL PARAM(K,N,B,XINF,S,BT)
      M=M+1
      DO 70 I=1,6
 70   TESTE(I)=ABS(BT(I)-B(I))/BT(I)
      IF(TESTE(1).GE.0.0001) GO TO 75
      IF(TESTE(2).GE.0.0001) GO TO 75
      IF(TESTE(3).GE.0.0001) GO TO 75
```

```
      IF(TEST(E(4).GE.0.0001) GO TO 75
      IF(TEST(E(5).GE.0.0001) GO TO 75
      IF(TEST(E(6).GE.0.0001) GO TO 75
      GO TO 78
75    DO 80 I=1,6
80    B(I)=BI(I)
      GO TO 50
78    CALL PROB(K,N,B,X,P)
      CALL INFO(N,K,X,P,XINF)
      CALL SINV(XINF,K,L,LO,PVT,IFAULT)
      DO 22 I=1,306
      NC(I)=0
22    NS(I)=0
      DO 24 I=1,306
      NS(I)=0
      N=0
      DO 26 J=1,306
      IF(P(I).EQ.P(J)) GO TO 28
      GO TO 26
28    N=N+1
      NS(I)=NS(I)+Y(J)
26    CONTINUE
      NC(I)=N
24    CONTINUE
      YN=0
      DO 30 J=1,6
      DO 30 I=1,222
30    XN(I,J)=0
      DO 32 I=1,222
      NCR(I)=0
      NSR(I)=0
32    PN(I)=0
      DO 34 I=1,306
      NCONT=0
      IF(NC(I).NE.1) GO TO 36
37    Z=I-YN
      PN(Z)=P(I)
      XN(Z,1)=X(I,1)
      XN(Z,2)=X(I,2)
      XN(Z,3)=X(I,3)
      XN(Z,4)=X(I,4)
      XN(Z,5)=X(I,5)
      XN(Z,6)=X(I,6)
      NCR(Z)=NC(I)
      NSR(Z)=NS(I)
      GO TO 34
36    DO 38 J=1,306
      IF(PN(J).EQ.P(I)) NCONT=NCONT+1
38    CONTINUE
      IF(NCONT.EQ.0) GO TO 37
      YN=YN+NCONT
34    CONTINUE
      O=0
      DO 14 I=1,222
      D(I)=NSR(I)-NCR(I)*PV(I)
      V(I)=NCR(I)*PN(I)*(1.0-PN(I))
      R(I)=SQRT(V(I))
      Q(I)=D(I)**2
      O=O+Q(I)/V(I)
      T(I)=D(I)/R(I)
14    CONTINUE
      CONT=0
      DO 11 I=1,222
      BOGI=0
      DIF(I)=NCR(I)-NSR(I)
      IF(DIF(I).EQ.0) GO TO 12
      IF(DIF(I).EQ.NCR(I))GO TO 13
      RECC=ALOG(NSR(I)/DIF(I))
```

```
      DO 17 J=1,6
   17 BOGI=B(J)*XN(I,J)+BOGI
      PARCIA=NSR(I)*(RECO-BOGI)
      ATETA=NCR(I)*ALOG(1.0 +NSR(I)/DIF(I))
      ABETA=NCR(I)*ALOG(1.0 +EXP(BOGI))
      A(I)=2*(PARCIA+ATETA+ABETA)
      IF(RECO.GT.BOGI) GO TO 16
      CD(I)=-SQRT(A(I))
      GO TO 15
   16 CD(I)=SQRT(A(I))
      GO TO 15
   12 A(I)=-2*NCR(I)*ALOG(PN(I))
      CD(I)=SQRT(A(I))
      GO TO 15
   13 A(I)=-2*NCR(I)*ALOG(1.0-PN(I))
      CD(I)=-SQRT(A(I))
   15 CONT=CONT+A(I)


   11 CONTINUE
      SOMA=0
      DO 20 I=1,222
      SOMA=SOMA+1
      CONTA(I)=SOMA
   20 CONTINUE

C                     SAIDA PADRAO

      WRITE(6,91) B(1),B(2),B(3),B(4),B(5),B(6)
   91 FORMAT('1',1X,131('='),//,60X,'SAIDA PADRAO',//,1X,131('='),
     *////,55X,'PARAMETROS ESTIMADOS',//,50X,'CONSTANTE =',F9.6,//,
     *50X,'AMEMORREIA =',F9.6,//,50X,'ESTADO =',F9.6,//,50X,'CURCIL ='
     *,F9.6,//,50X,'CORDAO =',F9.6,//,50X,'PESO =',F9.6,////,54X,
     *'PROBABILIDADES ESTIMADAS',//)
      WRITE(6,92)(CONTA(I),PI(I),NCR(I),NSR(I),I=1,222)
   92 FORMAT(4(10X,I3,2X,F7.5,2X,I2,2X,I2),/)
      WRITE(6,93)
   93 FORMAT(///,50X,'MATRIZ DE VARIANCIA-COVARIANCIA',//)
      WRITE(6,94)(XINF(I,1),XINF(I,2),XINF(I,3),XINF(I,4),XINF(I,5)
     *,XINF(I,6),I=1,6)
   94 FORMAT(30X,F9.6,4X,F9.6,4X,F9.6,4X,F9.6,4X,F9.6,4X,F9.6,/)
      WRITE(6,95) Q
   95 FORMAT(///,54X,'ESTATISTICA QUI-QUADRADO =',F9.5,////,52X,
     *'COMPONENTES DO QUI-QUADRADO',//)
      WRITE (6,96)(CONTA(I),E(I),I=1,222)
   96 FORMAT(5(12X,I3,2X,F9.5),/)
      WRITE(6,97) CONT
   97 FORMAT(///,52X,'DESVIO =',F15.8)
      WRITE(6,104)
  104 FORMAT('1',1X,131('='),//,60X,
     *'DIAGNOSTICOS',//,1X,131('='),////,62X,'COMPONENTES DO DESVIO'
     *,///)
      WRITE(6,98)(CONTA(I),CQ(I),I=1,222)
   98 FORMAT(5(8X,I3,2X,F9.6),/)
      N=222
      CALL CHAP(N,K,XN,PN,XINF,CHA,XHAT)
      WRITE(6,99)
   99 FORMAT(///,50X,'ELEMENTOS DA DIAGONAL DA MATRIZ DE PROJECAO',//)
      WRITE(6,90) (CONTA(I),XHAT(I),I=1,222)
   90 FORMAT(5(8X,I3,2X,F9.5),/)
      CALL IMP(Y,S,XN,D,XHAT,XINF,XPA)
      WRITE(6,100)
  100 FORMAT(///,50X,'IMPACTO SOBRE OS PARAMETROS',//,20X,
     *'B0',12X,'B1',12X,'B2',12X,'B3',12X,'B4',12X,'B5',/)
      WRITE(6,101)(CONTA(I),(XPA(I,J),J=1,6),I=1,222)
  101 FORMAT(15X,I3,5X,F8.5,5X,F8.5,5X,F8.5,5X,F8.5,5X,F8.5,5X,F8.5,/)
      DO 110 I=1,222
      CN(I)=(Q(I)/V(I))*XHAT(I)/(1.0-XHAT(I))
      CQ(I)=CN(I)/(1.0-XHAT(I))
```

```
      DES(I)=A(I)+CN(I)
      XQUI(I)=CN(I)/XHAT(I)
  110 CONTINUE
      WRITE(6,112)
  112 FORMAT(///,30X,'OCOOK',13X,'OCOOK-1',11X,'SQDESVIO',10X,
     *'SQQUI-QUADRADO',34X,//)
      WRITE(6,115)(CONTA(I),CC(I),CN(I),DES(I),XQUI(I),I=1,222)
  115 FORMAT(15X,I3,5X,F8.5,12X,F8.5,12X,F8.5,12X,F8.5,/)
      STOP
      END
```

```
C
C             SUBROTINA PARA CALCULAR OS VALORES DAS
C             PROBABILIDADES DE DIMENSAO N
C
      SUBROUTINE PROB(K,N,B,X,P)
      REAL NUM
      DIMENSION X(N,K),P(N),B(K)
      DO 15 I=1,N
      SOMA=0
      DO 10 J=1,K
   10 SOMA=X(I,J)*B(J)+SOMA
      NUM=EXP(SOMA)
   15 P(I)=NUM/(NUM+1.0)
      RETURN
      END
```

```
C
C             SUBROTINA PARA CALCULAR OS VALORES DOS
C             ELEMENTOS DO VETOR SCORE DE DIMENSAO K
C
      SUBROUTINE SCORE(K,N,Y,X,P,S)
      DIMENSION X(N,K),Y(N),P(N),S(K)
      T=0
      R=0
      DO 15 J=1,K
      DO 10 I=1,N
      T=X(I,J)*Y(I)+T
   10 R=X(I,J)*P(I)+R
      S(J)=T-R
      T=0
   15 R=0
      RETURN
      END
```

```
C
C             SUBROTINA PARA CALCULAR OS ELEMENTOS DA
C             MATRIZ DE INFORMACAO DE DIMENSAO K*K
C
      SUBROUTINE INFO(N,K,X,P,XINF)
      DIMENSION X(N,K),P(N),XINF(K,K)
      DO 10 J=1,K
      DO 10 L=1,K
   10 XINF(L,J)=0
      DO 20 L=1,K
      DO 20 J=1,L
      DO 20 I=1,N
   20 XINF(L,J)=XINF(L,J)+X(I,J)*X(I,L)*P(I)*(1.0-P(I))
      DO 30 L=1,K
      DO 30 J=1,L
   30 XINF(J,L)=XINF(L,J)
      RETURN
      END
```

```
C
C                    SUBROTINA PARA INVERSAO DE MATRIZ SIMETRICA
      SUBROUTINE SINV(A,K,L,LO,PVT,IFAULT)
      DIMENSION A(L,L)
      REAL A,AA,AIP,BIG,EP,PVT,SMALL,T
      INTEGER P,PM,PP
      DATA BIG,SMALL/1.0E68,1.0E-7/
      IFAULT=3
      IF(IABS(LO).GT.K.OR.K.LT.1.OR.K.GT.L)RETURN
      IFAULT=0
      IF(LO.GE.0) GO TO 1
      EP=1.0
      P=-LO
      PVT=ABS(A(P,P))
      T=PVT
      GO TO 3
    1 EP=-1.0
      P=1
      PVT=BIG
    2 IF(P.EQ.LO) GO TO 12
      T=ABS(A(P,P))
      IF(T.LT.PVT) PVT=T
    3 IF(T.LT.SMALL) IFAULT=1
      IF(T.EQ.0.0) GO TO 15
      PM=P-1
      PP=P+1
      AA=1.0/A(P,P)
      A(P,P)=-AA
      IF(P.EQ.1) GO TO 8
      DO 7 I=1,PM
      AIP=A(I,P)*AA
      DO 4 J=I,PM
    4 A(I,J)=A(I,J)-AIP*A(J,P)
      IF(P.EQ.K) GO TO 6
      DO 5 J=PP,K
    5 A(I,J)=A(I,J)-AIP*A(P,J)
    6 A(I,P)=AIP*EP
    7 CONTINUE
    8 IF(P.EQ.K) GO TO 11
      DO 10 I=PP,K
      AIP=A(PP,I)*AA
      DO 9 J=I,K
    9 A(I,J)=A(I,J)-AIP*A(P,J)
      A(P,I)=AIP*EP
   10 CONTINUE
   11 IF(EP.GT.0.0) RETURN
   12 P=P+1
      IF(P.LE.K) GO TO 2
C
C                    CORRECAO DO SINAL
      DO 14 I=1,K
      DO 13 J=I,K
   13 A(I,J)=-A(I,J)
   14 CONTINUE
      DO 16 I=1,K
      DO 17 J=I,K
   17 A(J,I)=A(I,J)
   16 CONTINUE
      RETURN
   15 IFAULT=2
      RETURN
      END
```

```
C
C                    SUBROTINA PARA CALCULAR O VETOR DE PARAMETROS
C                    DO MODELO DE DIMENSAO K
C
      SUBROUTINE PARAM(K,N,B,XINF,S,BT)
      DIMENSION B(K),XINF(K,K),S(K),BT(K)
      SOMA=0
      DO 40 I=1,K
      DO 30 J=1,K
 30   SOMA=XINF(I,J)*S(J)+SOMA
      BT(I)=B(I)+SOMA
 40   SOMA=0
      RETURN
      END


C
C     SUBROTINA PARA ACHAR OS ELEMENTOS DA DIAGONAL
C     DA MATRIZ DE PROJECAO
C
      SUBROUTINE CHAP(N,K,X,P,XINF,CHA,XHAT)
      DIMENSION X(N,K),P(N),XINF(K,K),CHA(N,K),XHAT(N)
      DO 10 J=1,N
      DO 10 I=1,K
 10   CHA(J,I)=0
      DO 20 I=1,N
      DO 30 L=1,K
      SOMA=0
      DO 40 J=1,K
 40   SOMA=X(I,J)*XINF(J,L)+SOMA
 30   CHA(I,L)=SOMA
 20   CONTINUE
C     A MATRIZ CHA E EXATAMENTE X(X'VX)-1 DE DIMENSAO N*K
      DO 50 I=1,N
 50   XHAT(I)=0
      DO 60 I=1,N
      SOMA=0
      DO 70 J=1,K
 70   SOMA=CHA(I,J)*X(I,J)+SOMA
      XHAT(I)=SOMA*P(I)*(1.0-P(I))
 60   CONTINUE
      RETURN
      END


C
C     SUBROTINA PARA CALCULAR O IMPACTO SOBRE OS PARAMETROS
C     QUANDO DA RETIRADA DE UM CASO
C
      SUBROUTINE IMP(N,K,X,D,XHAT,XINF,XPA)
      DIMENSION X(N,K),D(N),XHAT(N),XINF(K,K),XPA(N,K)
      DO 10 I=1,N
      DO 10 J=1,K
 10   XPA(I,J)=0
      DO 20 I=1,N
      DO 30 J=1,K
      SOMA=0
      DO 40 L=1,K
 40   SOMA=SOMA+X(I,L)*XINF(J,L)
      XPA(I,J)=SOMA
 30   XPA(I,J)=XPA(I,J)*D(I)/(1.0-XHAT(I))/SQRT(XINF(J,J))
 20   CONTINUE
      RETURN
      END
```

BIBLIOGRAFIA

. BICKEL, P.J. & DOKSUM, K.A. Mathematical Statistics; basic ideas and selected topics. San Francisco, Holden-Day, 1977.

. CHAMBERS, J.M. Fitting nonlinear models: numerical techniques. Biometrika, 60 (1): 1-13, apr. 1973.

. COOK, R.D. & WEISBERG, S. Residuals and influence in regression. New York, Chapman Hall, 1982.

. COX, D.R. The analysis of binary data. London, Methuen, 1970.

. DRAPER, N.R. & SMITH, H. Applied regression analysis. 2ªed. New York, John Wiley & Sons, 1981.

. EVERITT, B.S. The analysis of contingency tables. London, Chapman Hall; New York, John Wiley & Sons, 1977.

. FIENBERG, S.E. The analysis of cross - classified categorical data. 2ªed. Cambridge, The mit, 1981.

. FINNEY, D.J. Probit analysis. 2ªed. Cambridge, University, 1952.

. HOAGLIN, D.C. & WELSCH, R.E. The hat matrix in regression and ANOVA. The American Statistician, 32 (1): 17-22, feb. 1978.