Regressão Linear Ponderada na Seleção de Variáveis Independentes em Modelos de Regressão Logística

Este exemplar corresponde à redação final da tese devidamente corrigida e defendida pela Sra. Tirza Aidar, e aprovada pela Comissão Julgadora.

Campinas, 17 de Setembro de 1992

Profa. Dra. Eliana H. de Freitas Marques

Eliana H. de Freitas Marques

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Título de Mestre em Estatística

Aos meus pais, David e Nadyr e
à minha filha querida Leila.

AGRADECIMENTOS

Agradeço a oportunidade de encontrar no ambiente de trabalho tantas pessoas cujo bom humor, profissionalismo e amizade, tiveram atuação construtiva em minha formação profissional.

Agradeço à Eliana por nunca deixar que as dificuldades que surgiram interferissem no desenvolvimento deste trabalho, me incentivando sempre com sua confiança .

Aos meus colegas e amigos Aloísio, Carlos, Elaine, Gladston, etc, que participaram de perto das horas difíceis e das conquistas durante o verão de 89, nos cursos, e na elaboração da tese.

Aos amigos de sempre, Sônia, Nico, Nádia, Natália, Angela, Dominique, Marcão, Luiz, Fátima, Marlene, etc, pela sua paciência e amizade nos momentos de crise, e o jeitinho especial de cada um estar presente, torcendo, e se alegrando comigo.

E finalmente, agradeço a companhia constante da pequena Leila, que entre gracejos, chôros e solicitações, cresce com arte, arte na vontade de brincar, descobrir e questionar, me lembrando sempre do lado lúdico e construtivo do processo de vida.

Tirza.

SUMÁRIO

FIGURAS

EXEMPLOS

CAPITULO 1

INTRODUÇÃO

Temos em disponibilidade, e com relativa facilidade, farta literatura acompanhada de programas estatísticos de fácil acesso, que lidam com modelos de regressão linear de uma maneira eficiente e abrangente, contemplando os diversos aspectos de análise de dados como, estimativas de parâmetros, análise de resíduos, diagnósticos, seleção de variáveis independentes, etc.

Basicamente, os modelos de regressão buscam estudar as relações existentes entre o comportamento de uma variável resposta Y, chamada de variável dependente , e um vetor de variáveis independentes $\underset{\sim}{X}$. O modelo de regressão linear é definido por :

$$\underset{\sim}{Y} = \underset{\sim}{X}'\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad \text{com} \qquad E(\underset{\sim}{Y} \mid \underset{\sim}{X}) = \underset{\sim}{X}'\underset{\sim}{\beta}$$

onde é suposto que $\underset{\sim}{Y}$ tem distribuição Normal;
$\underset{\sim}{X}$ é a matriz do modelo correspondendo aos valores das variáveis independentes em cada observação, e X' é a sua matriz tranposta;

$\underset{\sim}{\beta}$ é o vetor de parâmetros desconhecidos,

e $\underset{\sim}{\varepsilon}$ é o vetor de erros independentes tal que $\underset{\sim}{\varepsilon} \sim N(0, I\sigma^2)$.

Quando $\underset{\sim}{Y}$ não é variável contínua, e sim discreta, indicando a ocorrência ou não de determinado evento, ou representando categorias de determinada condição, o modelo acima não se aplica, sendo a regressão logística uma ferramenta eficiente nesses casos.

McCullagh e Nelder (1989) , tratam da regressão logística como caso particular de modelos lineares generalizados, onde $\underset{\sim}{Y}$ pode assumir qualquer distribuicão pertencente à família exponencial, e o modelo é definido por uma funcão de ligação, $g(\underset{\sim}{u})$, monótona e diferenciável tal que:

$$E(\underset{\sim}{Y}|\underset{\sim}{X}) = \underset{\sim}{u} \qquad \text{e} \qquad g(\underset{\sim}{u}) = \underset{\sim}{X}'\underset{\sim}{\beta}.$$

Para o caso onde Y assume somente dois valores (Y=0 ou Y=1), a regressão logística dicotômica é definida por:

$$E(Y|\underset{\sim}{X}) = u = P(Y=1|\underset{\sim}{X}) \qquad \text{e}$$

$$g(u) = \ln(u/1-u) = \underset{\sim}{X}'\underset{\sim}{\beta} \quad ,$$

onde $g(u)$ é o logito da probabilidade de Y assumir valor igual a 1 , dado o vetor de variáveis independentes $\underset{\sim}{X}$.

Nos últimos dez anos muito se tem escrito sobre análise de dados categóricos, em particular sobre modelos baseados nos logitos das probabilidades de ocorrência das categorias da variável resposta Y, sendo o logito de uma probabilidade p definido por : $\ln(p/1-p)$.

A técnica regressão logística, já conhecida anteriormente pelos trabalhos de Cornfield (1962) e de Truett , Cornfield e Kannel (1967)

sobre doenças coronárias, passou a ser amplamente utilizada, em diversas áreas de pesquisa , após a introdução dos procedimentos de estimação por Máxima Verossimilhança por Walter e Duncan (1967), Day e Kerridge (1967) e Cox (1970). (Breslow 1980).

A aplicação clássica de regressão logística ocorre em análise de dados obtidos sob condições experimentais, como os experimentos biológicos relacionados a estudos de dose-resposta. (Bishop 1975, Kleinbaum 1982).

Grande parte da literatura existente no assunto está diretamente ligada à Bioestatística. É nas áreas de saúde, especialmente em Epidemiologia que aparecem muitas aplicações de regressão logística, com o objetivo de estudar relações existentes entre ocorrência de doenças e fatores de riscos, controlando-se fatores considerados de confundimento , como por exemplo, idade, sexo, condição social, etc. Kleimbaun (1982) apresenta uma dicussão detalhada sobre o uso de regressão logística dicotômica na área epidemiológica. Hosmer e Lemeshow (1989) apresentam em seu livro texto, uma revisão das técnicas desenvolvidas nos últimos anos para análise de regressão logística dicotômica.

O uso mais frequente de modelos de regressão logística é no estudo entre uma variável resposta binária e um conjunto de variáveis independentes, mas o modelo também se aplica à situações onde a variável resposta é politômica, apresentando k>2 categorias. Neste contexto é importante a identificação da estrutura das categorias, que pode ser nominal ou ordinal.

Se a estrutura for nominal, isto é, se as categorias representam "condições" ou "classes" sem que exista uma ordenação natural entre elas, o modelo utilizado, que será referido no trabalho como modelo logístico politômico, baseia-se nos logitos múltiplos das probabilidades de (k-1) categorias com relação à uma categoria chamada de padrão ou referência. Hosmer e Lemeshow tratam deste modelo utilizando regressões logísticas dicotômicas individuais propostas por Begg e Gray (1984).

Quando a variável resposta apresenta uma estrutura ordenada, modelos que incorporam essa informação são , em geral, mais adequados, estando entre eles o modelo de odds proporcionais (MacCullag 1980), baseado nos logitos múltiplos de probabilidades acumuladas, e o modelo de Anderson (1984) que é construído através de restrições impostas aos parâmetros do modelo logístico politômico.

O número de artigos recentemente publicados sobre regressão para análise de variáveis respostas ordenadas, demonstra o crescente interesse e utilização dos modelos existentes. (Agresti 1984, Armstrong et al 1989, Peterson et al 1990, Greenwood et al 1988, Ashby et al 1986).

Holtbügge e Shmacher (1991) comparam os modelos citados acima, verificando que o modelo de odds proporcionais produz estimativas com menor vício e erro padrão, e advertem que antes que o modelo de Anderson se torne amplamente utilizado, mais estudos devem ser desenvolvidos para avaliação da eficiência do modelo e testes de hipóteses propostos pelo autor.

Com a variedade de técnicas e o surgimento de algorítimos destinados a análise de dados categóricos, os modelos logísticos estão sendo utilizados nas diversas áreas da pesquisa científica, com possibilidades de análises cada vez mais abrangentes, inclusive no que se diz respeito à grandes conjuntos de dados.

Um importante passo da modelagem, quando se trabalha com conjuntos de dados que contenham informações sobre inúmeras variáveis, é a seleção de um subconjunto destas, de maneira que se obtenha um modelo bem ajustado aos dados, e que com um número mínimo de parâmetros, consiga de maneira parsimoniosa atingir o objetivo primeiro da análise, que , geralmente, é de descrever o comportamento da variável resposta Y com relação à um conjunto de variáveis independentes.

Em modelos de regressão linear, a seleção de modelos baseada em subconjuntos de q variáveis independentes, é feita principalmente por dois

métodos bastante conhecidos e disponíveis em programas estatísticos. O primeiro, através do ajuste de todos os $w=p^2$ possíveis modelos, para um conjunto inicial de p variáveis independentes, seleciona aqueles com maior coeficiente de determinação, R^2 , e com o coeficiente de Mallow (C(q)) nos limites aceitáveis, isto é, $C(q) \cong q$. O segundo procedimento, chamado Stepwise, faz uma escolha sistemática onde apenas alguns modelos são verificados diminuindo assim o tempo computacional. A cada passo da seleção, uma variável pode entrar ou sair do modelo conforme critérios pré-estabelecidos baseados em comparações entre os coeficientes de correlação parcial, contribuição da variável no coeficiente de determinação, e o valor da estatística F. (Draper 1981).

Apesar do procedimento Stepwise ser mais rápido e econômico, os procedimentos que ajustam todos submodelos (modelos com $q<p$ variáveis independentes), e geram medidas de comparação entre eles, possibilitam que o pesquisador tenha maior maleabilidade na escolha do modelo a ser adotado, de maneira que a decisão seja baseada não somente em uma medida estatística como também na coerência e interpretação com relação às áreas envolvidas na pesquisa.

Furnival e Wilson (1974) desenvolveram um algorítimo eficiente de ajuste de todos os possíveis modelos para regressão linear através de operações "sweep" na matriz formada pelos blocos $\underset{\sim}{X}'\underset{\sim}{X}$, $\underset{\sim}{X}'\underset{\sim}{Y}$, $\underset{\sim}{Y}'\underset{\sim}{X}$ e $\underset{\sim}{Y}'\underset{\sim}{Y}$, o que proporcionou uma alternativa aos procedimentos Stepwise sem grandes perdas em termos de tempo e gastos computacionais.

Para modelos de Regressão Logística, a extensão das metodologias mencionadas se torna mais complexa pelo fato de serem necessários métodos iterativos para obtenção de estimativas dos parâmetros de cada submodelo.

Baseados no algorítimo de Furnival e Wilson, Lawless e Singhal (1978) apresentaram uma metodologia de seleção para modelos lineares generalizados através de comparações dos submodelos com o modelo completo de acordo com estatísticas de razão de verossimilhança.

Hosmer, Jovanovic e Lemeshaw (1989) demonstraram que , para o caso particular de Regressão Logística Dicotômica, a procura do melhor modelo pode ser efetuada através de Regressão Linear com estimativas de Mínimos Quadrados Ponderados , viabilizando a utilização de algorítimos menos complexos, e por consequência mais rápidos, e accessíveis, dado hoje em dia a facilidade com que se encontra programas computacionais para regressão linear .

Como mencionado anteriormente, o uso da regressão logística tem se ampliado rapidamente. A realização desse trabalho tem por finalidade apresentar o método proposto por Hosmer et all (1989) de seleção e ajuste de submodelos para variável resposta dicotômica, e estender a técnica para modelos com variável resposta politômica nominal e ordinal. A intenção é de divulgação da técnica para que o pesquisador que não tenha acesso fácil à programas específicos de seleção de variáveis independentes em modelos logísticos, tenha, com programas de regressão linear ponderada, hoje bastante populares na área de estatística, uma alternativa para realização de análises em estudos que envolvam número grande de variáveis.

Para variáveis respostas dicotômicas, a aplicação do método é direta e seus aspectos teóricos serão apresentados no capítulo 2 após uma revisão de ajuste, teste de significância e interpretação dos parâmetros no modelo.

No terceiro capítulo é considerada a extensão do método de Hosmer e Lemeshow para seleção de variáveis independentes em modelos com variáveis respostas politômicas. Para o modelo logístico politômico, o método também se aplica ás (k-1) regressões dicotômicas estimadas individualmente. Quando Y apresenta categorias com estrutura ordenada, dois modelos são abordados, o modelo de odds proporcionais (McCullagh 1980) e o modelo unidimensional de Anderson (1984). Para o primeiro demonstrou-se que o método de Hosmer e Lemeshow pode ser também utilizado através da aplicação de função de ligação composta (Tomphson e Baker 1981). O modelo de Anderson apresentado na secção 3.4, será utilizado como uma alternativa de seleção de variáveis independentes para modelos ordenados baseada em análises descritivas dos coeficientes estimados do

modelo logístico politômico.

A metodologia apresentada nos capítulos 2 e 3 é ilustrada no quarto capítulo com análises de um conjunto de dados proveniente da CEPLAC (Comissão Executiva do Plano da Lavoura do Cacau). Os dados dizem respeito à um levantamento da situação de áreas de cacau, implantadas no período de 1976 à 1985, que teve como objetivo a avaliação qualitativa das plantações, e a identificação de fatores que de alguma maneira estejam associados à essa "qualidade".

A título de exemplificação do material apresentado, ao longo do trabalho aparecem alguns exemplos elaborados à partir dos dados citados acima. Uma descrição mais detalhada do conjunto de dados utilizado é feita no capítulo 4.

CAPITULO 2

SELEÇÃO DE VARIÁVEIS INDEPENDENTES ATRAVÉS DE AJUSTES POR MQP EM MODELOS DE REGRESSÃO LOGÍSTICA DICOTÔMICA

2.1 - INTRODUÇÃO

Neste capítulo é feita uma revisão do modelo de Regressão Logística Dicotômica, e é apresentado o método de seleção de variáveis independentes baseado em ajustes por Mínimos Quadrados Ponderados (MQP), proposto por Hosmer e Lameshow .

Suponha uma amostra de n observações independentes, proveniente de uma população de interesse, onde para cada unidade amostral, Y_i assume valor 1 se determinado evento ocorre, e 0 caso contrário. Associado a cada observação, seja o vetor $\underset{\sim}{X}_i$ de variáveis independentes, previamente escolhidas pelo pesquisador que julga, por conhecimentos anteriores, estarem associadas à distribuição da variável dependente. Ou seja :

$$Y_i = \begin{cases} 1 \text{ se o evento ocorre} \\ 0 \text{ caso contrário} \end{cases}$$

$$\underset{\sim}{X}_i' = (x_{i0}, x_{i1}, \ldots, x_{ip}) \qquad i=1,2,..n ,$$

onde x_{ij} é o valor da j-ésima variável na i-ésima observação com j=1,2,...,p para um conjunto de p variáveis independentes que podem ser categóricas ou contínuas, e $x_{i0} = 1$ para todo j .

Quando todas as variáveis independentes são categóricas , os dados podem ser estruturados numa tabela de contingência 2*N onde N é o número de grupos, ou subpopulacões, formados pelo cruzamento das categorias das variáveis explanatórias. No exemplo da Tabela 2.1 temos duas variáveis independentes, X1 assumindo dois valores, e X2 assumindo três valores, sendo N=2*3=6 o número de subpopulações e nji a frequência de casos de ocorrência do j-ésimo evento (j=0,1) na i-ésima subpopulação (i=1,2,...,N).

Tabela 2.1 : Representação de dados agrupados.

X1	X2	Y = 0	Y = 1	TOTAL
	1	n_{01}	n_{11}	$n._1$
1	2	n_{02}	n_{12}	$n._2$
	3	n_{03}	n_{13}	.
	1	n_{04}	n_{14}	.
2	2	n_{05}	n_{15}	.
	3	n_{06}	n_{16}	$n._6$
TOTAL		$n_0.$	$n_1.$	n

Em casos mais gerais, quando uma ou mais variáveis independentes assumem valores contínuos, as informações são representadas como na Tabela 2.2, onde N=n e $n._1 = n._2 = \ldots = n._n = 1$.

Tabela 2.2 : = Representação de dados não agrupados.

i	Y_i	x_{11}	x_{12}	...	x_{1p}
1	y_1	x_{11}	x_{12}	...	x_{1p}
2	y_2	x_{21}	x_{22}	...	x_{2p}
⋮	⋮	⋮	⋮		⋮
n	y_n	x_{n1}	x_{n2}	...	x_{np}

onde x_{ij} - valor da j-ésima variável na i-ésima observação,

e y_i - valor da variável resposta na i-ésima observação.
($i=1,2,\ldots,n$ e $j=1,2,\ldots,p$).

A discriminação entre as estruturas de dados é importante, pois alguns métodos de análise apropriados para dados agrupados, baseados em suposições sobre n^* tendendo à infinito, onde $n^*=\min(n.1,\ldots,n.n) \to \infty$, não se aplicam em dados não agrupados, como por exemplo os métodos para análise de modelos log-lineares.

Os dados citados no capítulo 1 se enquadram no segundo caso, contendo variáveis independentes categóricas e contínuas. Um dos interesses na escolha desse exemplo é o de se trabalhar com as diferentes estruturas de variáveis, possibilitando a abordagem dos diversos aspectos relativos à escolha de um subconjunto de variáveis independentes.

Para ilustração, nos dados utilizados a variável resposta "nota dada pelo entrevistador relativa à qualidade da área de cacau", PAR_TEC (Tabela 4.2), será tratada como variável dicotômica, onde:

$$Y = \begin{cases} 1 & \text{se PAR\_TEC = Ótimo ou Bom (sucesso), e} \\ 0 & \text{se PAR\_TEC = Regular ou Péssimo (fracasso).} \end{cases}$$

2.2 - TRANSFORMAÇÃO LOGÍSTICA.

O modelo de Regressão Logística é baseado no ajuste linear do logito da probabilidade de ocorrência da variável resposta dado o vetor de variáveis independentes $\underset{\sim}{X}$:

$$g(\Pi(\underset{\sim}{X}_i)) = \ln(\Pi(\underset{\sim}{X}_i)/1-\Pi(\underset{\sim}{X}_i)) = \underset{\sim}{X}'\underset{\sim}{\beta} \qquad (2.1)$$

onde:

$$\Pi_i = \Pi(\underset{\sim}{X}_i) = P(Y_i=1 \mid \underset{\sim}{X}_i) = \frac{\exp(\underset{\sim}{X}_i'\underset{\sim}{\beta})}{1 + \exp(\underset{\sim}{X}_i'\underset{\sim}{\beta})}$$

$i=1,2,\ldots,n$;

$\underset{\sim}{X}_i' = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{ip})$ com $\alpha_{i0}=1$ para todo i, e

$\underset{\sim}{\beta}'=(\beta_0,\beta_1,\ldots,\beta_p)$ é o vetor de parâmetros desconhecidos.

A importância da transformação logística está no fato do modelo (2.1) fornecer interpretação simples, onde os parâmetros são funções da razão de odds da probabilidade de ocorrência de Y , além de garantir estimativas das probabilidades Π_i pertencentes ao intervalo (0,1) para qualquer $\underset{\sim}{\beta} \in (-\infty, \infty)$ e $\underset{\sim}{X}_i$.

A odds de uma probabilidade (Π) é definida pelo quociente entre esta e seu complementar:

$$OP_{\pi} = \frac{\Pi}{(1-\Pi)} \qquad \text{para} \qquad 0 < \Pi < 1.$$

Ou seja, é uma medida de diferença relativa entre a probabilidade de ocorrência e de não ocorrência de um mesmo evento.

O estudo da relação entre um evento com probabilidade Π de ocorrência e um fator externo x, é feito pela avaliação da Razão de Odds:

$$\Psi(x) = \frac{P(\text{ocorrência}|x) \;/\; P(\bar{\text{n}}\ \text{ocorrência}\ |x)}{P(\text{ocorrência}|\bar{x}) \;/\; P(\bar{\text{n}}\ \text{ocorrência}|\bar{x})} = \frac{OP_{\pi}(x)}{OP_{\pi}(\bar{x})},$$

onde, x é uma variável dicotômica representando exposição (x) ou não exposição ($\bar{x}$) a determinado fator externo, às vezes chamado de fator de risco. Se a associação entre o fator de risco e o evento de interesse for fraca, a razão de odds, ou razão de riscos, será próxima de um e seu logarítimo próximo de zero. A seguir será apresentado um exemplo que ilustra a utilização de razão de odds como medida de associação.

EXEMPLO 2.1 :

Suponha que se deseja verificar se a probabilidade de sucesso da área de cacau está associada ao fator "problema de mão de obra", PMO. As frequências observadas estão descritas na Tabela 2.3:

Tabela 2.3 : Frequência observada de Parecer Técnico (Y) por Problema de Mão de Obra (PMO).

Prob . M. Obra	Parecer Técnico		
	Bom/Ótimo	Ruim/Péss.	Total
$\overline{PMO}$	285	453	738
PMO	93	271	364
Total	378	724	1102

A odds observada da probabilidade de sucesso, isto é , da área implantada no período ter parecer técnico bom ou ótimo, é dada por:

$$OP_s = \frac{378}{724} = 0.52$$

Ou seja, sem levar em consideração outros fatores, a probabilidade de uma área ser considerada boa ou ótima na amostra é aproximadamente a metade de ser ruim ou péssima.

Considere agora a possibilidade da qualidade da área estar associada à problemas como disponibilidade e/ou qualidade de mão de obra. O cálculo da razão de odds é dado por:

$$\Psi\ (PMO) = \frac{OP(PMO)}{OP(\overline{PMO})} = \frac{93/271}{285/453} = \frac{0,34}{0,63} = 0.54$$

isto é, a odds da probabilidade da área ser boa ou ótima no grupo onde ocorreu problema com mão de obra é aproximadamente metade do que nas áreas sem problemas.

O modelo (2.1) pertence à classe de Modelos Lineares Generalizados, onde $\underset{\sim}{Y}$ pode assumir qualquer distribuição da família exponencial, e não nescessariamente a média $E(\underset{\sim}{Y})= \underset{\sim}{\mu}$ deve ter uma relação linear com $\underset{\sim}{X}$, mas sim qualquer outra função $g(\underset{\sim}{\mu})$ monótona e diferenciável chamada de função de ligação . (McCullagh e Nelder 1989 pg 98).

A função de Verossimilhança para uma família exponencial é dada por :

$$f(y,\theta,\phi) = \exp\{[y\theta-b(\theta)]/a(\phi)+c(y,\phi)\} \qquad (2.2)$$

Quando ϕ é conhecido, $g(\mu)=\theta$ é chamada de função de ligação canônica, o que garante a existência de uma estatística suficiente para $\underset{\sim}{\beta}$, no preditor linear $\underset{\sim}{\eta}=\underset{\sim}{X}'\underset{\sim}{\beta}$.

Considerando o caso mais geral, isto é, quando existem variáveis independentes contínuas, como na Tabela 2.2, e supondo que para cada observação, Y_i assuma distribuição de Bernoulli com parâmetro Π_i, a função de verossimilhança da i-ésima observação é descrita por:

$$L_{Y_i}(y,\Pi_i) = \Pi_i^{y}\ (1-\Pi_i)^{1-y} \qquad y = 0 \text{ ou } 1 \quad .$$

$$L_{Y_i}(y,\Pi_i) = \exp\{y\ln(\Pi_i) + (1-y)\ln(1-\Pi_i)\} =$$

$$= \exp\{y\ln(\Pi_i/1-\Pi_i) + \ln(1-\Pi_i)\}$$

Fazendo:

$$\theta_i = \ln(\Pi_i/\ 1-\Pi_i) \qquad \text{obtêm-se} \qquad \Pi_i = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} \quad ,$$

e

$$L_{Y_i}(y;\Theta_i,\phi) = \exp\{y\Theta_i - \ln(1+\exp(\Theta_i))\} =$$

$$= \exp\{[y\Theta_i - b(\Theta_i)]/a(\phi)+c(y,\phi)\} \quad ,$$

onde : $c(y,\phi) = 0$; $b(\Theta_i) = \ln\{1+\exp(\Theta_i)\}$; $\phi = a(\phi)=1$.

Logo, a transformação logística é uma função de ligação canônica, onde $\Theta_i = \eta_i = \ln(\Pi_i/1-\Pi_i)$.

Para dados agrupados, como na Tabela 2.1, a função de verossimilhança da amostra é o produto das probabilidades Π_i, onde Π_i, $n_{.i}$ são os parâmetros da distribuição binomial, tal que:

Π_i - é a probabilidade de ocorrência de Y na i-ésima subpopulação,
$n_{.i}=y_i$ - é o número de ocorrências de Y=1 na i-ésima subpopulação, então:

$$L_Y(n_{.i},\Pi_i) = \binom{n_{.i}}{y_i} \Pi_i^{y_i} (1-\Pi_i)^{n_{.i}-y_i} =$$

$$= \exp\left\{ \ln\binom{n_{.i}}{y_i} + y_i \ln(\Pi_i/1-\Pi_i) + n_{.i}\ln(1-\Pi_i) \right\} =$$

$$= \exp\{[y\Theta_i - b(\Theta_i)]/a(\phi)+c(y,\phi)\} \quad ,$$

onde : $c(y,\phi) = \binom{n_{.i}}{y_i}$; $b(\Theta_i)=n_i \ln\{1+\exp(\Theta_i)\}$; $\phi = a(\phi)=1$.

Para a aplicação de Modelos Lineares Generalizados, o uso da função de ligação canônica não é obrigatório, mas é conveniente pois leva à propriedades estatísticas desejáveis, particularmente em amostras pequenas.

2.3 ESTIMAÇÃO E INTEPRETAÇÃO DOS PARÂMETROS, E TESTES DE QUALIDADE DE AJUSTE.

2.3.1 Estimador de Máxima Verossimilhança

Os estimadores de Máxima Verossimilhança (MV) de $\underset{\sim}{\beta}$ obtidos pela solução das equações geradas pela derivada da função de verossimilhança, podem ser calculados pelo método de Newton Raphson (NR) ou por Mínimos Quadrados Ponderados Iterativos (MQPI), apresentados à seguir.

Seja o modelo :

$$\underset{\sim}{\eta} = \ln(\underset{\sim}{\Pi} / 1 - \underset{\sim}{\Pi}) = \underset{\sim}{X}'\underset{\sim}{\beta} \qquad (2.3)$$

onde :

$\underset{\sim}{\Pi} = P(\underset{\sim}{Y}=1 \mid \underset{\sim}{X})$,

$\underset{\sim}{X} = (\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_n)$ é a matriz do modelo,

$\underset{\sim}{Y} \sim B(\underset{\sim}{\Pi})$,

$\underset{\sim}{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ é o vetor de parâmetros desconhecidos.

A função de Verossimilhança da amostra na forma canônica de uma família exponencial pode ser escrita por :

$$L(\underset{\sim}{X}_i'\underset{\sim}{\beta}, \underset{\sim}{y}) = \prod_{i=1}^{n} P(Y=\psi_i \mid \underset{\sim}{X}_i) =$$

$$= \prod_{i=1}^{n} \left(\frac{\exp(\underset{\sim}{X}_i'\underset{\sim}{\beta})}{1+\exp(\underset{\sim}{X}_i'\underset{\sim}{\beta})} \right)^{\psi_i} \left(\frac{1}{1+\exp(\underset{\sim}{X}_i'\underset{\sim}{\beta})} \right)^{1-\psi_i} \qquad \text{com } \psi_i=0 \text{ ou } 1, \text{ e}$$

$$\ln L(\underset{\sim}{X}'\underset{\sim}{\beta}; \underset{\sim}{y}) = \sum_{i=1}^{n} \ln L(\underset{\sim}{X}'\underset{\sim}{\beta}; y_i)$$

$$= \sum_{i=1}^{n} [\psi_i \Theta_i - b(\Theta_i)]/a(\phi) + c(\phi, \psi_i)$$

com $\Theta_i = \eta_i = \ln(\Pi_i/1-\Pi_i) = \underset{\sim}{X}_i'\underset{\sim}{\beta}$

e $c(\phi,\psi_i)=0 \;\; ; \; b(\Theta_i) = \ln(1+\exp(\Theta_i)) \; ; \;\; a(\phi)=1 \;\; .$

Então ,

$$\ln L(\underset{\sim}{X}'\underset{\sim}{\beta}; \underset{\sim}{y}) = \sum_{i=1}^{n} \left[\psi_i \underset{\sim}{X}_i'\underset{\sim}{\beta} - \ln(1+\exp(\underset{\sim}{X}_i'\underset{\sim}{\beta}) \right] .$$

O estimador de Máxima Verossimilhança $\hat{\underset{\sim}{\beta}}$ de $\underset{\sim}{\beta}$, é obtido pelas soluções das (p+1) equações dadas pela derivada de L com relação a $\underset{\sim}{\beta}$, isto é :

$$\frac{\partial L(\underset{\sim}{X}'\hat{\underset{\sim}{\beta}}; \underset{\sim}{y})}{\partial \hat{\beta}} = 0 \qquad \textbf{(2.4)}$$

ou ,

$$\partial \ln L(\underset{\sim}{X}'\hat{\underset{\sim}{\beta}}; \underset{\sim}{y}\ )/\partial \hat{\underset{\sim}{\beta}}_j =$$

$$= \sum_{i=1}^{n} \left[\psi_i x_{ij} - x_{ij} \frac{\partial\ b(\ \underset{\sim}{X}_i'\ \hat{\underset{\sim}{\beta}})}{\partial\ \hat{\beta}_j} \right] = \sum_{i=1}^{n} x_{ij} (\psi_i - b'(\underset{\sim}{X}_i'\hat{\underset{\sim}{\beta}}\ ) = 0 \ .$$

Onde $b'(.)$ e $b''(.)$ são as derivadas primeira e segunda da função $b(.)$ com relação à β, respectivamente.

Como $$E\left(\frac{\partial\ L_i}{\partial\ \theta_i} \right) = E(y_i - b'(\theta_i)) = 0 \ ,$$

temos $$E(y_i) = b'(\theta_i) = \frac{\exp\ (\theta_i)}{1+\exp(\theta_i)} = \Pi_i \quad \text{e} \quad \text{ainda}$$

$$E\left(\frac{\partial^2 L_i}{\partial\ \theta_i} \right) + \left(\frac{\partial\ L_i}{\partial\ \theta_i} \right)^2 = 0 =$$

$$= \ - b''(\theta_i) + E\ [y_i - b'(\theta_i)]^2 \ ,$$

Então $$V_i = V(\ y_i) = b''(\theta_i) = \Pi_i (1 - \Pi_i).$$

Logo, o sistema de equações é:

$$\underset{\sim}{X}'\underset{\sim}{s} = \underset{\sim}{X}'(\underset{\sim}{y} - \hat{\underset{\sim}{y}}) = \underset{\sim}{X}'(\underset{\sim}{y} - \hat{\underset{\sim}{\Pi}}) = 0 \qquad \textbf{(2.5)}$$

ou

$$\sum_{i=1}^{n} x_{ij} [\psi_i - 1/(1+\exp(-\underset{\sim}{X}_i'\hat{\underset{\sim}{\beta}}))] = 0$$

$$\sum_{i=1}^{n} \alpha_{ij}(y_i - \hat{\Pi}_i) = 0 \qquad \text{para} \quad j=0,1,\ldots p.$$

Como o sistema de equações é não linear em $\hat{\underset{\sim}{\beta}}$, são necessários métodos iterativos para sua solução.

Quando a segunda derivada de $\ln L(\underset{\sim}{X}'\underset{\sim}{\beta};\underset{\sim}{y})$ é de fácil determinação, como neste caso, o método de Newton-Raphson pode ser empregado, e de uma maneira simplificada pode ser descrito como :

Seja :

$$\frac{-\partial^2 \ln L\ (\ \hat{\underset{\sim}{X}}'\underset{\sim}{\beta};\ \underset{\sim}{y}\ )}{\partial\hat{\beta}_i\ \partial\hat{\beta}_j} = \frac{-\partial\ \underset{\sim}{X}'(\underset{\sim}{y}-\hat{\underset{\sim}{\Pi}})}{\partial\hat{\underset{\sim}{\beta}}} = \sum_{i=1}^{n} \alpha_{ij}\ b''(\underset{\sim}{X}_i'\hat{\underset{\sim}{\beta}})\alpha_{ij} = \underset{\sim}{X}\ \underset{\sim}{V}\ \underset{\sim}{X}' \ .$$

Para $\underset{\sim}{X}\ \underset{\sim}{V}\ \underset{\sim}{X}'$ positiva definida, as soluções do sistema **(2.5)** são pontos de máximo para $L(\underset{\sim}{X}'\underset{\sim}{\beta},\underset{\sim}{y})$. O processo de solução se inicia com $\underset{\sim}{\beta}^0$ inicial, derivando $\underset{\sim}{\beta}^1, \underset{\sim}{\beta}^2, \ldots$ sucessivamente, tal que:

$$\underset{\sim}{\beta}^{t+1} = \underset{\sim}{\beta}^t + (\ \underset{\sim}{X}\ \underset{\sim}{V}\ \underset{\sim}{X}'\ )^{-1}\ \underset{\sim}{X}\ \underset{\sim}{s} \qquad t=1,\ 2,\ldots \qquad \textbf{(2.6)}$$

onde t identifica a t-ésima iteração, e $\underset{\sim}{V}$ e $\underset{\sim}{s}=(\underset{\sim}{y}-\hat{\underset{\sim}{\Pi}})$ são recalculados em cada passo por $\underset{\sim}{\beta}^t$.

A convergência é estabelecida quando $|\underset{\sim}{\beta}^t - \underset{\sim}{\beta}^{t+1}| < \underset{\sim}{c_1}$, onde $\underset{\sim}{c_1} > \underset{\sim}{0}$ é o grau de precisão pré-estabelecido. Quando um dos critérios é satisfeito, o processo é interrompido e o estimador é dado por $\hat{\underset{\sim}{\beta}} = \underset{\sim}{\beta}^{t+1}$.

Um segundo método de solução do sistema (2.5) é o de Mínimos Quadrados Ponderados Iterativos, MQPI, implementado no pacote GLIM (Generalized Linear Interactive Model - Numerical Algorithms Group 1987) para ajustes de Modelos Lineares Generalizados. O método baseia-se no ajuste por mínimos quadrados ponderados (MQP) de Z, forma linearizada da função de ligação, com peso W, onde W^{-1} é a variância de Z.

Isto é, seja $\hat{\underset{\sim}{\eta}}_0$ estimador do preditor linear, dado um $\hat{\underset{\sim}{\beta}}_0$ inicial, e $\hat{\underset{\sim}{\Pi}}_0$ calculado pela função de ligação, então :

$$g(\hat{\Pi}_0) = \ln(\hat{\Pi}_0/1-\hat{\Pi}_0) = \hat{\underset{\sim}{\eta}}_0 = X'\hat{\underset{\sim}{\beta}}_0 \quad ,$$

$$\underset{\sim}{Z}_0 = \hat{\underset{\sim}{\eta}}_0 + (\underset{\sim}{y} - \hat{\underset{\sim}{\Pi}}_0)\left(\frac{\partial \eta}{\partial \Pi}\right)_0 \qquad \text{com} \quad \underset{\sim}{W}_0 = \hat{\underset{\sim}{\Pi}}_0(1-\hat{\underset{\sim}{\Pi}}_0)$$

pois

$$V(Z) = E(Z^2) - E^2(Z) = E[\ \eta + (y - \hat{\Pi})\frac{\partial \eta}{\partial \Pi}\ ]^2 - E^2(\eta) =$$

$$= E\left[\eta^2 + 2\eta(y - \hat{\Pi})\frac{\partial \eta}{\partial \Pi} + (y - \hat{\Pi})^2\left(\frac{\partial \eta}{\partial \Pi}\right)^2\right] - \eta^2 =$$

$$= V(y)\left(\frac{\partial \eta}{\partial \Pi}\right)^2 = W^{-1} \quad .$$

Partindo-se de $\tilde{\underset{\sim}{\beta}}_0$, ajusta-se por mínimos quadrados ponderados o modelo linear $Z_0 = \underset{\sim}{X}'\underset{\sim}{\beta}_1$ com peso W_0 , estimando-se $\tilde{\underset{\sim}{\beta}}_1$. O processo é repetido até a obtenção de $|\tilde{\underset{\sim}{\beta}}_{t+1} - \tilde{\underset{\sim}{\beta}}_t| < c$ onde c é o grau de precisão pré-determinado.

McCullagh e Nelder (1989 pg 41) sugerem como ponto inicial os valores observados de $\hat{\Pi}_0$, derivando daí $\hat{\eta}_0$, $(\partial\eta_0/\partial\hat{\Pi}_0)$ e W_0.

Os métodos iterativos de NR e de MQPI são equivalentes no se diz respeito à obtenção do estimador de MV $\hat{\underset{\sim}{\beta}}$:

$$\text{Seja} \qquad \underset{\sim}{Z}=\underset{\sim}{\eta} + (\underset{\sim}{y}-\hat{\underset{\sim}{\Pi}})\frac{\partial \eta}{\partial \underset{\sim}{\Pi}} = \underset{\sim}{X}'\hat{\underset{\sim}{\beta}} + \underset{\sim}{V}^{-1}(\underset{\sim}{y}-\hat{\underset{\sim}{\Pi}}) ,$$

escrevendo na forma matricial :

$$\underset{\sim}{Z} = \underset{\sim}{X}'\hat{\underset{\sim}{\beta}} + \underset{\sim}{V}^{-1}\underset{\sim}{s} \qquad \Rightarrow \qquad \underset{\sim}{s} = \underset{\sim}{V} (\underset{\sim}{Z} - \underset{\sim}{X}'\hat{\underset{\sim}{\beta}}) ,$$

e substituindo em (2.6) obtêm-se :

$$\underset{\sim}{\beta}^{t+1} = \underset{\sim}{\beta}^{t} + (\underset{\sim}{X} \underset{\sim}{V} \underset{\sim}{X}')^{-1} \underset{\sim}{X} \underset{\sim}{V} (\underset{\sim}{Z}^{t} - \underset{\sim}{X}'\underset{\sim}{\beta}^{t})$$

$$= \underset{\sim}{\beta}^{t} + (\underset{\sim}{X} \underset{\sim}{V} \underset{\sim}{X}')^{-1} \underset{\sim}{X} \underset{\sim}{V} \underset{\sim}{Z}^{t} - (\underset{\sim}{X} \underset{\sim}{V} \underset{\sim}{X}')^{-1} \underset{\sim}{X} \underset{\sim}{V} \underset{\sim}{X}'\underset{\sim}{\beta}^{t}$$

$$= (\underset{\sim}{X} \underset{\sim}{V} \underset{\sim}{X}')^{-1} \underset{\sim}{X} \underset{\sim}{V} \underset{\sim}{Z}^{t} = \tilde{\underset{\sim}{\beta}} .$$

Grizzle, Stamer e Koch (1969) propõe outro método de estimação em regressão logística dicotômica, o método não iterativo de MQP. Para situações onde a maioria das probabilidades Π_i são diferentes de 0 ou 1, o método de G.S.K. fornece estimadores de MV aproximados, sendo portanto apropriado para dados representados em tabelas de contingência , como a Tabela 2.1 .

2.3.2 Interpretação dos Parâmetros

Os coeficientes $\underset{\sim}{\beta}$ do modelo (2.1) são uma medida do efeito das variáveis independentes $\underset{\sim}{X}$ sobre o logito da probabilidade de ocorrência da variável resposta.

Seja , por exemplo, X_j uma variável dicotômica assumindo valor 0 ou 1 , e o logito a seguir :

$$g(\Pi) = \ln(\Pi / 1-\Pi) = \underset{\sim}{X}^{*\prime}\underset{\sim}{\beta}^{*} + X_j'\beta_j \quad ,$$

onde $X^{*\prime}$ é a matriz do modelo sem a j-ésima linha referente à variável X_j e β^{*} o vetor de parâmetros sem o parâmetro β_j. Mantendo a matriz X^{*} constante, sejam :

$$g(\Pi_1) = \ln(\Pi_1/1-\Pi_1) = \underset{\sim}{X}^{*\prime}\underset{\sim}{\beta}^{*} + X_j'\beta_j \quad ,$$

e

$$g(\Pi_0) = \ln(\Pi_0/1-\Pi_0) = \underset{\sim}{X}^{*\prime}\underset{\sim}{\beta}^{*} \quad ;$$

onde $\Pi_k = P(Y=1 \mid X^{*} \text{ e } X_j=k)$ $\qquad$ K = 0 ou 1.

A razão de odds, Ψ_j , para $X_j=1$ e $X_j=0$ escrita em função de $\underset{\sim}{\beta}$ é dada por

$$\Psi_j = \frac{\Pi_1\ (1-\Pi_0)}{(1-\Pi_1)\ \Pi_0} = \frac{\dfrac{1}{1+\exp-(\underset{\sim}{X}^{*}{}'\underset{\sim}{\beta}^{*}+\beta_j)}}{\dfrac{1}{1+\exp(\underset{\sim}{X}^{*}{}'\underset{\sim}{\beta}^{*}+\beta_j)}} \qquad \frac{\dfrac{1}{1+\exp(\underset{\sim}{X}^{*}{}'\underset{\sim}{\beta}^{*})}}{\dfrac{1}{1+\exp-(\underset{\sim}{X}^{*}{}'\underset{\sim}{\beta}^{*})}}$$

$$= \frac{\exp(\underset{\sim}{X}^{*}{}'\underset{\sim}{\beta}^{*}+\beta_j)}{\exp(\underset{\sim}{X}^{*}{}'\underset{\sim}{\beta}^{*})} = \exp(\beta_j) \qquad \text{portanto} \qquad \beta_j = \ln(\Psi_j).$$

Assim, β_j representa a mudança no logaritimo da odds da probabilidade de ocorrência de Y , quando se passa da categoria 0 para a categoria 1 da variável X_j. Esta situação é ilustrada no no exemplo 2.2.

EXEMPLO 2.2 :

Considere o modelo logístico univariado aplicado aos dados da Tabela 2.3 :

$$\text{logito}(s) = \ln\ (\Pi(s)/1-\Pi(s)) = \beta_0 + X\ \beta_1$$

$$X = \begin{cases} 0 & \text{se não houve problema com mão de obra } (\overline{PMO}) \ , \\ 1 & \text{caso contrário } (PMO) \ . \end{cases}$$

O modelo ajustado aos dados é:

$$\text{logito}(s) = \ln(0.63) + \ln(0.34/0.63)\ X =$$

$$= \ln(0.63) + [\ln(0.34) - \ln(0.63)]\ X =$$

$= -0.4634 - 0.6061X$

$\beta o = -0.4634$ (logito da probabilidade de sucesso nas áreas sem problemas de mão de obra), e em consequência,

$\exp(\beta o)=0,6291$ que é o odds da probabilidade de sucesso para $\overline{PMO}$,

$\beta = -0.6061$ (acréscimo no logito para as áreas onde ocorreu problema de mão de obra), e portanto,

$\exp(\beta o+\beta)=0,3432$ que é o odds da probabilidade de sucesso para PMO.

É importante ressaltar que a expressão $\exp(\beta_j)=\Psi_j$ foi determinada pela codificação de X_j como 0 ou 1. Existem outras parametrizações , como a chamada dos desvios médios, onde $X_j=-1$ ou $X_j=1$, levando à expressão $\exp(2\beta_j)=\Psi_j$, o que modifica a interpretação do coeficiente β_j. O programa de regressão logística BMDPLR no BMDP (Dixon 1987) oferece também esta opção. No procedimento CATMOD do SAS, (Statistical Analysis Sistem-SAS Institute Inc. 1988), a parametrização é definida pelo usuário, sendo a parametrização dos desvios médios a padrão. Hosmer e Lemeshaw (1989) fazem uma revisão detalhada sobre alguns tipos de parametrização, e suas interpretações.

Suponha agora que X_j seja uma variável politômica, com k>2 categorias não ordenadas. Neste caso, X_j será representada no modelo por (k-1) coeficientes, referentes à (k-1) variáveis independentes dicotômicas, como ilustrado no exemplo 2.3.

EXEMPLO 2.3 :

Seja, a variável PROVISOR ,que descreve através das quatro

categorias, *bom, excessivo, ruim e* $\bar{n}tem$, como foi feito o sombreamento provisório da área de cacau. O cacau, cultura perene, necessita de sombreamento feito em duas etapas, pois não é interessante para seu crescimento e produção o contato direto com a luz solar. No ato de sua plantação, e até a planta atingir aproximadamente quatro anos de idade , o sombreamento adequado, chamado de provisório, é feito por culturas de crescimento rápido e de porte mediano, como por exemplo bananeiras, que são retiradas da área logo que o cacau atinja um tamanho adequado.

Na Tabela 2.4, sob o título de parametrização estão descritas as variáveis independentes dicotômicas relaecionadas às categorias da variável PROVISOR, e são apresentadas as frequências por categoria de Parecer Técnico.

Tabela 2.4 : Frequências observadas de Parecer técnico (Y) por Sombreamento Provisório (PROVISOR).

Somb. Prov.	Parametrização			Parecer Técnico		Total
	PRV1	PRV2	PRV3	1	0	
bom	0	0	0	175	181	356
exc.	1	0	0	29	152	181
ruim	0	1	0	38	218	256
ntem	0	0	1	136	173	309
Total				378	724	1107

Desta maneira para a variável X_j são definidos três parâmetros, β_{1j}, β_{2j} e β_{3j}, representando o ln da razão de odds das categorias "*exc*" , "*ruim*" e "*ntem*" com relação à categoria de referência "*bom*", respectivamente. A diferença no logito que acompanha a mudança da categoria "*bom*" para a "*exc*" é representada pelo parâmetro β_{1j} e assim consecutivamente.

O modelo ajustado resulta :

$$\text{logito}(\Pi(s)) = \beta_0 + \beta_1 prv1 + \beta_2 prv2 + \beta_3 prv3 =$$
$$= -0.034 - 1.623 prv1 - 1.713 prv2 - 0.207 prv3 .$$

Ou seja, o valor:

$$\ln[\Psi\ (bom, exc)] = \ln\left[\frac{29/152}{175/181}\right] = -1.623 \quad ,$$

representa o efeito do sombreamento ser *excessivo* no logito da probabilidade de *sucesso* da área implantada. Nas áreas com bom sombreamento ,a odds da probabilidade de *sucesso* é de exp(-0,034)=0,97 , diminuindo para exp(-0,034-1,623)=0,19 nas áreas com sombreamento *excessivo.*

A última categoria, $\overline{ntem}$, é ambígua pois pode estar respondendo àquelas áreas onde não tem, ou não teve sombreamento provisório, ou àquelas que por já estarem em idade avançada perdeu-se informação. Esta dúvida foi levantada quando observou-se que a frequência de áreas sem sombreamento provisório é bem maior na faixa de 5 à 10 anos de idade. Apesar disso, como sua frequência é grande, decidiu-se mantê-la no modelo, mesmo que isoladamente a categoria não gere nenhuma informação.

Quando a variável X_j é contínua, β_j reflete a mudança no ln da odds da probabilidade de Y para um incremento de uma unidade na escala de X_j.

A extenção destas interpretações é direta para os termos do modelo que identificam interação ou confundimento entre as variáveis

independentes.

Suponha um estudo com duas variáveis independentes, X_1 e X_2, neste caso alguns dos possíveis modelos são :

(1) $$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$$

(2) $$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(3) $$\eta = \beta_0 + \beta_1 X_1$$

(4) $$\eta = \beta_0$$

No primeiro modelo diz-se que existe associação entre a função de ligação, η, com as duas variáveis, e além disso que X_2 atua como uma variável modificadora da associação de X_1, ou seja , a associação de η com X_1 varia a depender do valor de X_2. Neste caso a interação $X_1 X_2$ é presente e deve ser considerada. Se as duas variáveis independentes apresentarem associação com η sem que uma interfira na outra, o modelo (2) é satisfatório, onde os efeitos β_1 e β_2 são avaliados de maneira que cada variável tenha sua porção de influência em η. Os dois efeitos devem ser considerados para se evitar a obtenção de uma medida de associação β_1 irreal pelo modelo (3), caso as subpopulações não sejam homogenêas com respeito a X_2. O terceiro modelo é adotado quando somente X_1 estiver associada a η . E finalmente o modelo (4) é aquele que apresenta o logito médio de toda a amostra para o caso onde X_1 e X_2 não apresentam associação significativa com η.

2.3.3 Qualidade de Ajuste do Modelo

Após o ajuste do modelo, é necessário que se faça algumas avaliações sobre a qualidade deste ajuste.

O primeiro passo é verificar se o modelo como um todo é satisfatório, no sentido de conter informações suficientes para explicar a variação da função de ligação com relação ao vetor de variáveis independentes, e em seguida estuda-se como, e se, cada variável incluída neste modelo está contribuindo para o ajuste.

Quando os dados são agrupados (Tabela 2.1) e as frequências $n._i$ não são muito pequenas , isto é, $n._i > 5$, o teste da qualidade do ajuste pode ser feito pelas estatísticas de Pearson (X^2) , e de razão de Verossimilhança (D), também chamada de função Desvio (McCullagh-1989), definidas por :

$$X^2 = \sum_{i=1}^{N} \frac{n.^2_i (\Pi^o_i - \hat{\Pi}_i)^2}{n._i \hat{\Pi}_i (1-\hat{\Pi}_i)}$$

onde $\Pi^o_i = \frac{n_{1i}}{n._i}$ (proporção observada) ,

$$\hat{\Pi}_i = \frac{\exp\ (\underset{\sim}{X}_i{}'\underset{\sim}{\beta})}{\exp\ (1+\underset{\sim}{X}_i{}'\underset{\sim}{\beta})}$$ (proporção estimada), e

N = número de subpopulações determinado pelas combinações das categorias das p variáveis independentes.

$$D = -2\log\left\{\frac{L(\underset{\sim}{\hat{\Pi}}, \underset{\sim}{Y})}{L(\underset{\sim}{\Pi}^{\circ}, \underset{\sim}{Y})}\right\}$$

$$D = -2\left\{\sum_{i=1}^{N} n_{1i}\log(\hat{\Pi}_i/\Pi_i^{\circ}) + (n._i - n_{1i})\log\left[(1-\hat{\Pi}_i)/(1-\Pi_i^{\circ})\right]\right\}$$

As estatísticas X^2 e D são assintoticamente equivalentes, com distribuição aproximada Qui-Quadrado com (N-p-1) graus de liberdade. A aproximação é baseada nas suposições de que as observações sejam independentes com distribuição $Y_{ij} \quad B(n._j, \Pi_j)$, de maneira que :

$$D \ \text{e} \ X^2 \underset{\substack{n._j \to \infty \\ j=1,\ldots n}}{\overset{ass}{\sim}} \chi^2_{(N-p-1)}$$

À medida que N se aproxima de n , e $n._i \approx 1$ para a maior parte de i=1,2...,N , como em casos de dados não agrupados (Tabela 2.2), a

aproximação não ocorre , e as estatísticas X^2 e D não devem ser utilizadas como medidas absolutas de qualidade de ajuste. Entretanto, na comparação entre dois modelos hierárquicos, um com p, e outro com q<p termos, a diferença entre suas funções desvios equivale à estatística de razão de verossimilhança , G, e tem distribuição assintótica χ^2 com (p-q) graus de liberdade, e a aproximação baseia-se somente no limite de n tendendo para infinito.

Outra estatística bastante citada na literatura para comparação entre modelos hierárquicos é a estatística de Wald, (W).

Supondo que cada observação seja independente com distribuição $Y_i \sim B(1,\Pi_i)$, com i=1,2...n, o modelo contendo p variáveis independentes pode ser comparado com o modelo mais simples, isto é, aquele somente com o intercepto, β_0, verificando se pelo menos uma das variáveis independentes está contribuindo significativamente para o aprimoramento do ajuste.

Ou então, compara-se o modelo selecionado à priori com outro contendo além dos termos referentes aos efeitos principais, algumas interações destes, avaliando-se a necessidade ou não de inclusão de termos de ordem maior que um para um melhor ajuste.

Para a hipótese :

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0 \qquad q<p$$

$$H_1 : \beta_i \neq 0 \qquad \text{para pelo menos um} \quad i = q+1,\ldots,p ,$$

as estatísticas G, e W são definidas por :

(1) $$G = D - D_{H0} =$$

$$= 2\log\{L(\underset{\sim}{\beta}_{H0})/L(\underset{\sim}{\beta})\}$$

onde $\underset{\sim}{\beta}'_{HO} = (\beta_0, \beta_1, \ldots, \beta_q)$,

$\underset{\sim}{\beta}' = (\beta_0, \beta_1, \ldots, \beta_q, \ldots, \beta_p)$,

e

D_{HO} - função Desvio sob Ho .

(2) $$W = \hat{\underset{\sim}{\beta}}_{HO} \Sigma^{-1}_{\sim HO} \hat{\underset{\sim}{\beta}}_{HO}$$

onde $\Sigma^{-1}_{\sim HO}$ é a parte da matriz de covariância correspondente aos r=p-q parâmetros excluidos do modelo .

O estudo individual de significância de cada variável pode ser feito pela estatística de Wald (W_j) :

Seja Ho: $\beta_j = 0$,

então,

$$W_j^2 = \left[\frac{\beta_j}{\hat{S}(\beta_j)} \right]^2$$

onde $\hat{S}(\beta_j) = [\hat{\sigma}^2_j]^{1/2}$ é o erro padrão estimado de β_j e $\hat{\sigma}^2_j$ é o j-ésimo elemento da diagonal da matriz de covariância de $\underset{\sim}{\beta}$, Σ_{β}, definida por:

$$\Sigma_{\beta} = (\underset{\sim}{X}' \underset{\sim}{V} \underset{\sim}{X})^{-1} \qquad \text{e} \qquad \underset{\sim}{V} = \text{diag}\ [\underset{\sim}{\Pi}(1-\underset{\sim}{\Pi})]$$

com $$W_j^2 \sim \chi^2_1$$

Hauck e Donner (1977) concluem que este teste é conservador, não rejeitando Ho para valores altos de β_j , e sugerem que conclusões finais sejam tiradas de testes de razão de verossimilhança:

$$D = -2\log\left[\frac{\text{verossimilhança do modelo com } X_j}{\text{verossimilhança do modelo sem } X_j}\right]$$

Quando uma variável independente politômica está sendo representada por um conjunto de variáveis independentes dicotômicas, análises cuidadosas devem ser feitas de maneira a tratar esse conjunto de forma análoga. Não se deve incluir ou excluir do modelo somente parte deste conjunto; se uma destas variáveis independentes dicotômicas for significativa, todas as outras devem ser incluídas .(Hosmer e Lemeshow 1989).

Se o número p de variáveis independentes proposto inicialmente é pequeno, a busca do melhor modelo pode ser feita passo a passo através do ajuste de vários modelos e comparações entre eles, mas à medida que p cresce esta seleção se torna uma tarefa difícil.

A seguir veremos o método proposto por Hosmer e Lemeshaw para a solução deste problema.

2.4 SELEÇÃO DE VARIÁVEIS INDEPENDENTES BASEADA EM AJUSTES POR MQP

2.4.1 Introdução

No modelo de regressão, quando o número de variáveis independentes é muito extenso, faz-se necessário um estudo cuidadoso de seleção destas, de maneira que se obtenha um modelo objetivo, estável e informativo. Isto é, um modelo de regressão com um número grande de variáveis independentes, além de ser de manutenção cara, e de difícil interpretação, pode conter variáveis intercorrelacionadas, gerando estimativas com altas variâncias, adicionando pouco ao poder de predição do modelo e prejudicando sua capacidade descritiva.

O problema então é encontrar um subconjunto de variáveis independentes que seja pequeno o suficiente para facilitar sua operacionalização e interpretação, e que ao mesmo tempo contenha informações com poder de descrição dos dados e controle de predição.

As primeiras variáveis independentes que devem ser excluídas, são aquelas com altos erros de medição, que não sejam fundamentais ao estudo, e que de alguma maneira dupliquem informação de outras variáveis mais estáveis. Em seguida deve-se definir a relação linear de cada variável com a função da variável resposta , no caso o logito, utilizando-se de transformações, se necessário.

Se após uma seleção inicial, o número de variáveis independentes ainda for grande, os procedimentos de escolha baseados no ajuste de todos

os possíveis modelos, viabilizam uma busca mais rápida, com ajustes e medidas de comparação entre eles, fornecendo possibilidades ao pesquisador de encontrar um melhor modelo baseado não só em critérios estatísticos como também em julgamentos sob o ponto de vista da área de pesquisa envolvida.

2.4.2 Procedimentos usuais

O procedimento mais utilizado em regressão logística é o "Stepwise", e suas variações, "Forward" e "Backward", disponíveis nos sistemas SAS e BMDP. No primeiro o processo inicia-se com o ajuste do modelo contendo o termo constante, β_0 , apenas. Em seguida ajusta-se todos os modelos contendo uma variável, que através de testes de razão de verossimilhança, são comparados com o primeiro modelo, selecionando-se aquele com menor valor de α_{ej} (nível de significância da j-ésima variável). Na terceira etapa ajusta-se os (p-1) modelos com duas variáveis independentes, e da mesma maneira escolhe-se aquele com menor α_{ej} segundo testes de razão de verossimilhança na comparação com o modelo selecionado na etapa anterior. Após essa segunda seleção, é verificado se a primeira variável selecionada deve permanecer no modelo. Isto é, se chamarmos a primeira variável selecionada de X_{j1}, e a segunda de X_{j2} , compara-se os modelos $\eta=\beta_0+\beta_{j1}X_{j1}+\beta_{j2}X_{j2}$ e $\eta=\beta_0+\beta_{j2}X_{j2}$, se o nível de significância for maior que α_s (nível de significância pré-determinado para se retirar uma variável), então elimina-se X_{j1} do modelo, caso contrário ele permanece. Os dois últimos passos são repetidos até que entre as variáveis independentes excluídas não exista nenhuma com α_{ej} menor que α_e, onde esse último é o nível de significância escolhido para entrada de uma variável no modelo. Os valores de α_e e α_s devem ser escolhidos de maneira que não exista risco de se eliminar variáveis independentes importantes. Alguns autores recomendam $\alpha_e=0,15$ e $\alpha_s=0,20$.

No procedimento "Forward" o processo inicia-se como no

"Stepwise", sendo que as variáveis independentes já incluídas no modelo não são retiradas em passos seguintes, e o processo é finalizado quando $\alpha_{ej} > \alpha_e$ para todo X_j restante. O contrário ocorre no "Backward", quando no primeiro passo todas as variáveis independentes são consideradas, eliminando-se a cada etapa aquela com maior α_{sj}.

Para regressão linear exitem vários procedimentos de seleção de variáveis que, por não serem baseados em métodos iterativos, são mais simples e rápidos quando comparados àqueles destinados aos modelos de regressão onde não se faz suposição de normalidade. Uma das medidas utilizadas para se comparar diferentes modelos com um mesmo número de parâmetros, é o coeficiente de Mallow, mencionado no capítulo 1 e definido por

$$C(q) = \frac{SQR(q)}{SQR(p)/(n-p-1)} + 2(q+1)-n \qquad (2.7)$$

onde SQR(p) = Soma de quadrados de resíduos do modelo completo,
SQR(q) = Soma de quadrados de resíduos do submodelo ,
n = número de observações,
p = número inicial de variáveis independentes e
q = número de variáveis independentes no submodelo.

Furnival e Wilson (1974) desenvolveram, para regressão linear, um algorítimo rápido e eficiente de seleção dos melhores modelos de cada tamanho q<p, . Através de operações nas matrizes A_1 e A_2 **(2.8)**, o algorítimo fornece os ajustes de todos os 2^p possíveis modelos. Diferentes operações, chamadas de operações "sweep", podem ser utilizadas, sendo que cada uma fornece determinada informação sobre os modelos ajustados, como, por exemplo, as estimativas dos coeficientes, a matriz de covariância , ou a soma de quadrados de resíduos.

$$A_1 = \begin{bmatrix} \underset{\sim}{X}\,\underset{\sim}{X}' & \underset{\sim}{X}\,\underset{\sim}{Y} \\ \underset{\sim}{Y}'\,\underset{\sim}{X}' & \underset{\sim}{Y}'\,\underset{\sim}{Y} \end{bmatrix} \qquad \text{e} \qquad A_2 = \begin{bmatrix} [\underset{\sim}{X}\underset{\sim}{X}']^{-1} & \hat{\underset{\sim}{\beta}} \\ \hat{\underset{\sim}{\beta}} & \underset{\sim}{0} \end{bmatrix} \qquad (2.8)$$

Lawless e Singhal (1978) adaptaram o algorítimo acima para regressão não normal, utilizando a matriz de Informação , $|$, e o estimador de MV $\hat{\underset{\sim}{\beta}}$, de $\underset{\sim}{\beta}$ do modelo completo. As operações são efetuadas nas matrizes B_1 e B_2 (2.9), produzindo uma sequência de todos os submodelos com aproximações $\tilde{\underset{\sim}{\beta}}$ de $\hat{\underset{\sim}{\beta}}$.

$$B_1 = \begin{bmatrix} I & I'\hat{\underset{\sim}{\beta}} \\ \hat{\underset{\sim}{\beta}}'\,I & \hat{\underset{\sim}{\beta}}'\,I\hat{\underset{\sim}{\beta}} \end{bmatrix} \qquad \text{e} \qquad B_2 = \begin{bmatrix} I^{-1} & \hat{\underset{\sim}{\beta}} \\ \hat{\underset{\sim}{\beta}} & 0 \end{bmatrix} \qquad (2.9)$$

onde, $\hat{\underset{\sim}{\beta}}$ é o estimador de MV do modelo completo , e

$$| = -E\left[\frac{\partial^2 \log L(\underset{\sim}{y}, \underset{\sim}{X}'\underset{\sim}{\beta})}{\partial\beta_j \; \partial\beta_{j'}}\right] \qquad j, j' = 0, 1, \ldots p \; .$$

Sob hipótese $Ho: \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$, esse algorítimo fornece o estimador de MV aproximado $\tilde{\underset{\sim}{\beta}}_1$ de $\underset{\sim}{\beta}_1$ calculado por :

$$\tilde{\underset{\sim}{\beta}}_1' = (\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_q, \underset{\sim}{0}) = (\hat{\underset{\sim}{\beta}}_1 - C_{12}\, C_{22}^{-1}\, \hat{\underset{\sim}{\beta}}_2, \underset{\sim}{0})$$

$$\text{onde} \quad \hat{\underset{\sim}{\beta}}' = (\hat{\underset{\sim}{\beta}}_1', \hat{\underset{\sim}{\beta}}_2') \qquad \text{e} \qquad |^{-1} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \; .$$

Os melhores modelos são selecionados através de comparações entre os submodelos e o modelo completo realizadas com aproximações do teste de razão de verossimilhança , sem que para isso sejam necessários novos ajustes ou métodos iterativos. A estatística de razão de verossimilhança para comparação do submodelo com o modelo completo é dada por :

$$\lambda = -2\log[L(\hat{\beta}_1)/L(\hat{\beta})]$$

onde $\hat{\beta}_1$ é o estimador de MV de β sob Ho .

São propostas duas alternativas à estatística λ, a estatística de Wald $\lambda^* = \hat{\beta}_2' C_{22}^{-1} \hat{\beta}_2$ e a aproximação $\lambda^{**} = -2\log[L(\tilde{\beta}_1)/L(\hat{\beta})]$ de λ ; ambas com distribuição assintótica χ^2 com (p-q) graus de liberdade, sob Ho: $\beta_2 = 0$.

Lawless e Singhal (1987a, 1987b) implementaram um sistema de análise de regressão para modelos lineares generalizados, contendo um algorítimo do método descrito acima , de seleção de variáveis através do ajuste de todos os possíveis submodelos.

2.4.3 Procedimento com ajuste de MQP

Hosmer et all (1989) mostram que para a Regressão Logística Dicotômica, o processo de ajuste e de seleção de modelos, proposto por Lawless e Singhal, pode ser efetuado através de programas de regressão linear, com ajustes de mínimos quadrados ponderados onde as comparações são baseadas no coeficiente de Mallow , C(p).

O método utiliza o estimador de MV $\hat{\beta}$, do modelo completo, ajustando por MQP os submodelos $Z = X_1'\beta_1$, ponderados por W, **(2.10)**, através dos quais obtêm-se as estimativas aproximadas de $\hat{\beta}_1$, $\tilde{\beta}_1$, a estatística λ' para comparação entre modelos hierárquicos, e o coeficiente de Mallow.

$$Z = X'\hat{\beta} + W^{-1}(\Pi - \hat{\Pi}) \qquad \textbf{(2.10)}$$

$$W = \text{diag}(w_i) = \text{diag}(\hat{\Pi}_i(1-\hat{\Pi}_i))$$

$$\hat{\beta} = (X\,W\,X')^{-1} X\,W\,Z' \qquad \textbf{(2.11)}$$

O processo inicia-se, então, com as estimativas do modelo completo, $\hat{\underset{\sim}{\beta}}$, obtidas de um programa para regressão logística, e em seguida, com as informações geradas, utiliza um programa para ajuste e seleção de submodelos em regressão linear. Abaixo são apresentados os resultados que justificam o procedimento proposto.

Seja $\hat{\underset{\sim}{Z}}(p)=\underset{\sim}{X}'\hat{\underset{\sim}{\beta}}$ o vetor estimado pela regressão linear ponderada com p variáveis independentes. A soma de quadrados dos resíduos é obtida por:

$$SQR(p) = [\underset{\sim}{Z}-\hat{\underset{\sim}{Z}}(p)]'\underset{\sim}{W}\,[\underset{\sim}{Z}-\hat{\underset{\sim}{Z}}(p)] =$$

$$= \underset{\sim}{Z}'\underset{\sim}{W}\,\underset{\sim}{Z} + \hat{\underset{\sim}{Z}}(p)'\underset{\sim}{W}\,\hat{\underset{\sim}{Z}}(p) - 2\hat{\underset{\sim}{Z}}(p)'\underset{\sim}{W}\,\underset{\sim}{Z} =$$

$$= \underset{\sim}{Z}'\underset{\sim}{W}\,\underset{\sim}{Z} + \hat{\underset{\sim}{\beta}}'\underset{\sim}{X}\,\underset{\sim}{W}\,\underset{\sim}{X}'\hat{\underset{\sim}{\beta}} - 2[\hat{\underset{\sim}{\beta}}'\underset{\sim}{X}\,\underset{\sim}{W}\,\underset{\sim}{X}'\hat{\underset{\sim}{\beta}} + \hat{\underset{\sim}{\beta}}'\underset{\sim}{X}\,\underset{\sim}{W}\,\underset{\sim}{W}^{-1}(\underset{\sim}{\Pi}-\hat{\underset{\sim}{\Pi}})] =$$

$$= \underset{\sim}{Z}'\underset{\sim}{W}\,\underset{\sim}{Z} - \hat{\underset{\sim}{\beta}}'\underset{\sim}{X}\,\underset{\sim}{W}\,\underset{\sim}{X}'\hat{\underset{\sim}{\beta}} = \sum_{i=1}^{n} w_i[Z_i-\hat{Z}_i(p)]^2 =$$

$$= \sum_{i=1}^{n} \frac{(\Pi_i-\hat{\Pi}_i)^2}{\hat{\Pi}_i(1-\hat{\Pi}_i)} = X^2 .$$

A última expressão é a estatística X^2 de Pearson do ajuste do modelo logístico.

Para o ajuste de um submodelo de tamanho q<p particiona-se a matriz $\underset{\sim}{X}$ em $\underset{\sim}{X}'= (\underset{\sim}{X}_1,\underset{\sim}{X}_2)'$ de maneira que $\underset{\sim}{X}_1$ seja a matriz de dimensão n*(q+1) relacionada ao subconjunto das q variáveis independentes mais o termo constante e $\underset{\sim}{X}_2$ às (p-q) variáveis independentes restantes, tal que :

$$\underset{\sim}{X} = (\underset{\sim}{X}_1,\underset{\sim}{X}_2)$$

$$\underset{\sim}{\beta}' = (\underset{\sim}{\beta}_1',\underset{\sim}{\beta}_2')$$

$$| = \underset{\sim}{X}\ \underset{\sim}{W}\ \underset{\sim}{X}' = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{X}_1 \underset{\sim}{W} \underset{\sim}{X}_1' & \underset{\sim}{X}_1 \underset{\sim}{W} \underset{\sim}{X}_2' \\ \underset{\sim}{X}_2 \underset{\sim}{W} \underset{\sim}{X}_1' & \underset{\sim}{X}_2 \underset{\sim}{W} \underset{\sim}{X}_2' \end{bmatrix} .$$

O estimador $\tilde{\underset{\sim}{\beta}}_1$ de $\underset{\sim}{\beta}_1$ obtido por mínimos quadrados ponderados do modelo $\underset{\sim}{Z} = \underset{\sim}{X}_1' \underset{\sim}{\beta}_1$ é dado por :

$$\tilde{\underset{\sim}{\beta}}_1 = [\underset{\sim}{X}_1 \underset{\sim}{W}\ \underset{\sim}{X}_1']^{-1} \underset{\sim}{X}_1 \underset{\sim}{W}\ \underset{\sim}{Z} \quad . \tag{2.12}$$

Substituindo Z por (2.10) :

$$\tilde{\underset{\sim}{\beta}}_1 = (\underset{\sim}{X}_1 \underset{\sim}{W} \underset{\sim}{X}_1')^{-1} \underset{\sim}{X}_1 \underset{\sim}{W} [(\underset{\sim}{X}_1'\ \underset{\sim}{X}_2') \begin{bmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix} + \underset{\sim}{W}^{-1} (\underset{\sim}{\Pi} - \hat{\underset{\sim}{\Pi}}) =$$

$$= (\underset{\sim}{X}_1 \underset{\sim}{W} \underset{\sim}{X}_1')^{-1} \underset{\sim}{X}_1 \underset{\sim}{W}\ [\underset{\sim}{X}_1' \hat{\underset{\sim}{\beta}}_1 + \underset{\sim}{X}_2' \hat{\underset{\sim}{\beta}}_2] + \underset{\sim}{W}^{-1} (\underset{\sim}{\Pi} - \hat{\underset{\sim}{\Pi}})$$

$$= \hat{\underset{\sim}{\beta}}_1 + I_{11}^{-1} I_{12} \hat{\underset{\sim}{\beta}}_2 + I_{11}^{-1}\ \underset{\sim}{X}_1'\ (\underset{\sim}{\Pi} - \hat{\underset{\sim}{\Pi}})$$

$$= \hat{\underset{\sim}{\beta}}_1 + I_{11}^{-1} I_{12} \hat{\underset{\sim}{\beta}}_2 \quad . \tag{2.13}$$

Sendo que $\hat{\underset{\sim}{\beta}}_1, \hat{\underset{\sim}{\beta}}_2, I_{11}$, e I_{12} são quantidades obtidas inicialmente pelo ajuste do modelo completo, tornando assim desnecessários novos cálculos que dependam de métodos iterativos.

Vê-se que o valor do estimador de Máxima Verossimilhança $\hat{\underset{\sim}{\beta}}_1$ de $\underset{\sim}{\beta}_1$, é obtido por métodos iterativos até a convergência de (2.12), e nada mais é do que o modelo $\underset{\sim}{\eta} = \underset{\sim}{X}' \underset{\sim}{\beta}$ sob hipótese Ho: $\underset{\sim}{\beta}_2 = 0$.

A soma de quadrados de resíduos do modelo ajustado $\tilde{\underset{\sim}{Z}}(q) = \underset{\sim}{X}_1' \tilde{\underset{\sim}{\beta}}_1$ é calculado por :

$$SQR(q) = [\underset{\sim}{Z} - \tilde{\underset{\sim}{Z}}(q)]' \underset{\sim}{W} [\underset{\sim}{Z} - \tilde{\underset{\sim}{Z}}(q)] =$$

$$= \underset{\sim}{Z}'\underset{\sim}{W}\underset{\sim}{Z} + \tilde{\underset{\sim}{Z}}(q)'\underset{\sim}{W}\,\tilde{\underset{\sim}{Z}}(q) - 2\tilde{\underset{\sim}{Z}}(q)'\underset{\sim}{W}\,\underset{\sim}{Z} =$$

$$= \underset{\sim}{Z}'\underset{\sim}{W}\,\underset{\sim}{Z} - \tilde{\underset{\sim}{\beta}}_1'(\underset{\sim 1}{X}\underset{\sim}{W}\,\underset{\sim 1}{X}')\tilde{\underset{\sim}{\beta}}_1$$

$$= X^2 + \hat{\underset{\sim}{\beta}}'(\underset{\sim}{X}\,\underset{\sim}{W}\,\underset{\sim}{X}')\hat{\underset{\sim}{\beta}} - \tilde{\underset{\sim}{\beta}}_1'(\underset{\sim 1}{X}\underset{\sim}{W}\,\underset{\sim 1}{X}')\tilde{\underset{\sim}{\beta}}_1$$

$$= X^2 + \hat{\underset{\sim}{\beta}}_2'(I_{22} - I_{21}I_{11}^{-1}I_{12})\hat{\underset{\sim}{\beta}}_2 \qquad \textbf{(2.14)}$$

$$= X^2 + \lambda^* \ .$$

Então : $\lambda^* = SQR(q) - SQR(P)$.

A matriz $\hat{\sum}_{(\hat{\beta}2)} = (\ I_{22} - I_{21}I_{11}^{-1}I_{12}\ )^{-1} = C_{22}$ é a matriz de covariância estimada de $\hat{\underset{\sim}{\beta}}_2$. Segue então que λ^* é a estatística de Wald para o teste de $Ho: \underset{\sim}{\beta}_2 = 0$.

As expressões **(2.13)** e **(2.14)** fornecem a base para o uso do estimador de mínimos quadrados ponderados da variável Z, no subconjunto de q variáveis independentes, como uma aproximação do estimador de máxima verossimilhança , e para a interpretação do ajuste.

Substituindo **(2.14)** em **(2.7)** :

$$C(q) = \frac{X^2 + \lambda^*}{X^2/(n-p-1)} + 2(q+1) - n \quad . \qquad \textbf{(2.15)}$$

Sob $Ho: \underset{\sim}{\beta}_2 = \underset{\sim}{0}$, λ^* tem distribuição assintótica χ^2 com (p-q) graus de liberdade , e assumindo que X^2 tem distribuição aproximada χ^2 com (n-p-1) graus de liberdade, o valor esperado de C(q) é aproximadamente (q+1). Ou seja, se a exclusão das (p-q) variáveis independentes do modelo não contribuir para um aumento significativo na soma de quadrados residual, o valor de C(q) deve se aproximar do número de parâmetros do modelo em questão, caso contrário terá distribuição

Qui-Quadrado não central o que leva a um valor de C(q) maior que (q+1). Portanto, o melhor modelo entre aqueles do mesmo tamanho q, é aquele com menor coeficiente de Mallow.

Para a comparação entre modelos com número diferentes de parâmetros, pode-se, como proposto por Lawless e Singhal (1978), utilizar a estatística $\lambda^* = SQR(q)-SQR(c)$, onde $p \geq c > q$.

Hanck e Donner (1977) e Jennings (1986) mostram que , especialmente para n pequeno, λ^* não tem comportamento monótono como a estatística de razão de verossimilhança, decrescendo para grandes valores de $\underset{\sim}{\beta}$, diminuindo o poder do teste. Ou seja, a estatística de Wald é conservadora, podendo, em alguns casos, aceitar Ho: $\underset{\sim}{\beta}_2=0$, quando a estatística de razão de verossimilhança rejeita ao mesmo nível de significância. Apesar destas restrições, a utilização do coeficiente de Mallow como definido em (2.15) é viável, pois, mesmo que seu valor seja superestimado, na comparação entre modelos de mesmo tamanho, aquele com menor valor de λ', e consequentemente menor valor de C(q) é aquele que melhor se ajusta aos dados observados.

O método proposto viabilisa, de maneira rápida e com facilidades computacionais, uma seleção preliminar de alguns modelos segundo critérios citados. A escolha do modelo não deve ser baseada somente em uma medida estatística, decisões finais devem ser elaboradas através de avaliações de acordo com estatística de razão de verossimilhança e considerações na área de interesse.

Hosmer e Lameshaw advertem que a utilização de métodos como Stepwise, de regressão linear, não são adequados para seleção de variáveis independentes em modelo Logístico Dicotômico através da pseudo variável dependente Z, pois pouco se sabe da interpretação da estatística F nesse caso.

A extensão do método para modelos logísticos com variáveis respostas politômicas será apresentada no próximo capítulo.

CAPITULO 3

MÉTODO DE SELEÇÃO DE VARIÁVEIS INDEPENDENTES BASEADO EM AJUSTE POR MQP EM MODELOS LOGÍSTICOS PARA VARIÁVEL RESPOSTA POLITÔMICA

3.1- INTRODUÇÃO

Neste capítulo serão apresentados modelos logísticos para variáveis respostas politômicas, com extensão do método de seleção de variáveis independentes proposto por Hosmer e Lemeshow .

Dependendo da estrutura apresentada pelas categorias de uma variável resposta politômica alguns modelos são mais adequados que outros, gerando interpretações que exploram esta estrutura no sentido de fornecer mais informações sobre o comportamento dos dados e associações de interesse.

As categorias de uma variável politômica podem apresentar uma estrutura nominal ou ordenada definidas por:

(1) Escala Nominal :

A variável Y assume k (k>2) valores , ou classes, distintos

que representam características da unidade amostral sem nenhuma estrutura de ordem , isto é, não faz sentido a ordenação das categorias já que nenhuma pode ser considerada maior , melhor ou pior que as outras. Exemplo:

$$Y = \text{cor de olhos} \begin{cases} \text{verdes} \\ \text{azuis} \\ \text{castanhos} \\ \text{pretos} \end{cases}$$

(2) Escala Ordenada :

Neste caso as k categorias apresentam estrutura ordenada, que é intrínseca à variável, podendo estar representando graus de uma segunda variável resposta, Z, contínua não observada. Exemplos:

	Y	Z (nota)
Y = Aproveitamento dos dos alunos de certo curso	ótimo	85-100
	bom	70- 84
	regular	51- 69
	ruim	00- 50

$$Y = \text{Avaliação visual de uma plantação de cacau} \begin{cases} \text{ótima} \\ \text{boa} \\ \text{regular} \\ \text{ruim} \end{cases}$$

O primeiro modelo a ser apresentado, denominado de regressão logística politômica, baseia-se nos logitos múltiplos das probabilidades das (k-1) categorias com relação à uma única categoria selecionada como

categoria padrão, e não leva em consideração nenhuma estrutura de ordem, sendo portanto adequado para variáveis respostas nominais.

Quando a estrutura das categorias da variável resposta Y é ordenada, McCullagh (1980) propõe um modelo em função dos logitos múltiplos das probabilidades acumuladas das categorias . O modelo de McCullagh, também chamado de modelo de odds proporcionais, parte da suposição de que a ordem das categorias de Y é fixa, e não depende do vetor $\underset{\sim}{X}$ de variáveis independentes, mesmo que esteja associado à ele.

Ao contrário do modelo anterior, Anderson (1984) apresenta um modelo, ao qual denominou "Steriotype Model", que através de restrições impostas aos parâmetros do modelo de Regressão Politômica, busca analisar como, e se, o vetor $\underset{\sim}{X}$ identifica a ordem assumida à priori pela variável resposta Y .

3.2 - MODELO DE REGRESSÃO LOGÍSTICA POLITÔMICA

3.2.1 Introdução

Seja uma amostra aleatória de tamanho n , onde as unidades amostrais são classificadas segundo k categorias de uma variável resposta Y, e, associadas à cada observação, p medidas formando o vetor $\underset{\sim}{X}_i'=(x_{i0},x_{i1},\ldots,x_{ip})$ de variáveis independentes, com i=1,2,...,n , onde $x_{i0}=1$ para obtenção do termo constante do modelo.

Seja Z uma variável indicadora, tal que :

$$\eta_{di} = \begin{cases} 1 & \text{se } Y_i = d \\ 0 & \text{caso contrário} \end{cases} \qquad d=1,2,\ldots,k.$$

Ou seja, $\eta_{di}=1$ se a i-ésima observação pertencer à d-ésima categoria de Y, e $\Pi_{di}=P[\eta_{di}=1|\underset{\sim}{X}_i]$, (d=1,2,...,k-1), é a probabilidade de Y_i assumir valor d.

O modelo logístico politômico baseia-se nos logitos múltiplos das probabilidades Π_{di} com relação à probabilidade de ocorrência da "categoria de referência", que será, por convenção, a k-ésima categoria, isto é :

$$\eta_{di}=\ln\left(\frac{\Pi_{di}}{\Pi_{ki}}\right)=\underset{\sim}{X}_i'\underset{\sim}{\beta}$$

e

$$\Pi_{di} = \frac{1}{1+\exp(-\underset{\sim}{X}_i'\underset{\sim}{\beta}_d)} \qquad (3.1)$$

onde $\underset{\sim}{\beta}_d = (\beta_{d0}, \beta_{d1}, \ldots, \beta_{dp})$ é o vetor de parâmetros desconhecidos, com $\underset{\sim}{\beta}_k = \underset{\sim}{0}$, e d=1,2,...,k-1.

3.2.2 Estimação dos parâmetros

Supondo independência das n observações, o ln da função de verossimilhança da amostra é definido como :

$$L(\underset{\sim}{\beta}, \underset{\sim}{X}) = \prod_{i=1}^{n} \prod_{d=1}^{k} \left(\Pi_{di}\right)^{\gamma_{di}} = \prod_{i=1}^{n} \prod_{d=1}^{k} \frac{\exp(\underset{\sim}{X}_i'\underset{\sim}{\beta}_d)}{\sum_{j=1,\ldots k} \exp(\underset{\sim}{X}_i'\underset{\sim}{\beta}_j)}$$

$$\ln L(\underset{\sim}{\beta}, \underset{\sim}{X}) = \sum_{i=1}^{n} \sum_{d=1}^{k} \gamma_{di} \left[\underset{\sim}{X}_i'\underset{\sim}{\beta}_d - \ln\left(\sum_{j=1}^{k} \exp(\underset{\sim}{X}_i'\underset{\sim}{\beta}_j)\right)\right] \qquad (3.2).$$

Os estimadores de Máxima Verossimilhança de $\underset{\sim}{\beta}$, $\hat{\underset{\sim}{\beta}}$, são obtidos pela solução das (p+1)×(k-1) equações geradas pela primeira derivada de (3.2) com relação aos parâmetros:

$$\frac{\partial \ln L(\underset{\sim}{\beta}, \underset{\sim}{X})}{\partial \beta_{dj}} = \sum_{i=1}^{n} \gamma_{di} x_{ij} - \left\{ \sum_{d'=1}^{k} \frac{\gamma_{d'i} x_{ij} \exp(\underset{\sim}{X}_i' \underset{\sim}{\beta}_d)}{\sum_{l=1}^{k} \exp(\underset{\sim}{X}_i' \underset{\sim}{\beta}_l)} \right\} =$$

$$= \sum_{i=1}^{n} \gamma_{di} x_{ij} - x_{ij} \Pi_{di} \sum_{d'=1}^{k} \gamma_{d'i} = \sum_{i=1}^{n} x_{ij} (\gamma_{di} - \hat{\Pi}_{di}) = 0 .$$

O método iterativo de Newton Raphson para a solução das equações utiliza a segunda derivada de $L(\hat{\underset{\sim}{\beta}}, \underset{\sim}{X})$ com relação à $\hat{\underset{\sim}{\beta}}$ dada por :

$$\frac{\partial^2 \ln L(\hat{\underset{\sim}{\beta}}, \underset{\sim}{X})}{\partial \beta_{dj} \partial \beta_{dj'}} = - \sum_{i=1}^{n} x_{ij} x_{ij'} (\hat{\Pi}_{di})(1-\hat{\Pi}_{di}) \quad ,$$

e

$$\frac{\partial^2 \ln L(\hat{\underset{\sim}{\beta}}, \underset{\sim}{X})}{\partial \beta_{dj} \partial \beta_{d'j'}} = \sum_{i=1}^{n} x_{ij} x_{ij'} (\hat{\Pi}_{di})(\hat{\Pi}_{d'i}) \quad ,$$

A matriz de informação $\mathcal{I}$ é a matriz quadrada de dimensão $(p+1)(k-1)$ composta por $(k-1)\times(k-1)$ submatrizes $V_{dd'}$, de dimensão $(p+1)\times(p+1)$, cujos (j,l)-ésimos elementos são :

$$(v_{dd'})_{jl} = \sum_{i=1}^{n} x_{ij} x_{il} \Pi_{di} [\, \delta(d,d') - \Pi_{d'i} \,] \qquad j,l = 0,1 \ldots p \quad ,$$

$$\text{com} \quad \delta(d,d') = \begin{cases} 1 & \text{se } d=d' \\ 0 & \text{caso contrário.} \end{cases}$$

3.2.3 Ajuste utilizando regressões individuais

Begg e Gray (1984) sugerem outro método de ajuste do Modelo Politômico, utilizando de Regressões Individuais. Os autores propõem o ajuste de cada uma das (k-1) categorias numa regressão logística dicotômica, e demonstram que os estimadores obtidos são consistentes com eficiência aproximada à do método politômico. Advertem entretanto, que essa eficiência decresce com o aumento do número de parâmetros e com baixas frequências da categoria padrão, ou de referência. Portanto, é importante que na escolha da categoria padrão, sua frequência, ou probabilidade de ocorrência, seja considerada.

O modelo baseado nas Regressões Logísticas Individuais com relação à categoria padrão pode ser escrito da forma:

$$\ln(\theta_d/\theta_k) = \underset{\sim}{X}'\underset{\sim}{\beta}_d^* \qquad (3.3)$$

onde :

$$\theta_d = P(y_d = 1 \mid y_d + y_k = 1, \underset{\sim}{X}) ,$$

$$\theta_k = P(y_d = 0 \mid y_d + y_k = 1, \underset{\sim}{X}) ,$$

e $\underset{\sim}{\beta}^*$ são vetores de parâmetros desconhecidos das (k-1) regressões logísticas dicotômicas.

Verifica-se que (3.1) e (3.3) são equivalentes no que diz respeito aos parâmetros, isto é, $\underset{\sim}{\beta} = \underset{\sim}{\beta}^*$. Pelo teorema de Bayes, temos que:

$$\theta_d = \frac{P([\mathcal{Y}_k = 1] \cup [\mathcal{Y}_d = 1] \mid \underset{\sim}{X}, \mathcal{Y}_d = 1) P(\mathcal{Y}_d = 1 \mid \underset{\sim}{X})}{P(\mathcal{Y}_k + \mathcal{Y}_d = 1 \mid \underset{\sim}{X})} =$$

$$= \frac{P(\mathcal{Y}_d = 1 \mid \underset{\sim}{X})}{P(\mathcal{Y}_k = 1 \mid \underset{\sim}{X}) + P(\mathcal{Y}_d = 1 \mid \underset{\sim}{X})} = \frac{\Pi_d}{\Pi_d + \Pi_k} \Rightarrow$$

$$\Rightarrow \quad \underset{\sim}{X}' \underset{\sim}{\beta}_d^* = \ln(\theta_d / \theta_k) = \ln\left(\frac{\Pi_d}{\Pi_d + \Pi_k} \times \frac{\Pi_d + \Pi_k}{\Pi_k} \right) = \ln(\Pi_d / \Pi_k) = \underset{\sim}{X}' \underset{\sim}{\beta}_d \quad .$$

Também é demonstrado que o estimador $\hat{\underset{\sim}{\beta}}^{*\prime} = (\hat{\underset{\sim}{\beta}}_1^*, \ldots, \hat{\underset{\sim}{\beta}}_{k-1}^*)$ tem média $\underset{\sim}{\beta}$ e matriz de covariância $H^{-1} W H^{-1}$, onde H é a matriz bloco diagonal com os blocos $H_1, H_2, \ldots, H_{k-1}$ tal que :

$$H_d = W_{dd'} \qquad d, d' = 1, 2, \ldots, k-1 ,$$

onde :

$$(W_{dd'})_{jl} = \sum_{i=1}^{n} x_{ij} x_{il} \Pi_{ki} \theta_{di} \theta_{d'i}^{[1-\delta(d,d')]}$$

$$j, l = 0, 1, \ldots, p.$$

O método Individualizado fornece uma alternativa de ajuste do modelo Logístico Politômico possibilitando análises de grandes conjuntos de dados, principalmente na fase de análise exploratória , onde a seleção de variáveis independentes pode ser feita por algoritimos menos complexos como de regressão logística dicotômica e de regressão linear ponderada.

A avaliação da qualidade do ajuste de um modelo Logístico

Politômico com variáveis independentes categóricas e contínuas, é feita , como discutido em **(2.3)**, pelo estudo de comparações entre modelos hierárquicos através da estatística de razão de verossimilhança. Hosmer et all (1989 pg 216), propõem como avaliação da qualidade de ajuste um estudo conjunto e descritivo das regressões individuais.

. A seguir será apresentado um exemplo de ajuste de modelo politômico pelos dois métodos citados.

3.2.4 Ilustração

EXEMPLO 3.1 :

Seja a variável resposta politômica definida por :

$$Y=\begin{cases} 1 & \text{se PAR\_TEC = Ótimo} \\ 2 & \text{se PAR\_TEC = Bom} \\ 3 & \text{se PAR\_TEC = Ruim} \\ 4 & \text{se PAR\_TEC = Péssimo} \end{cases},$$

e as variáveis referentes à identificação das áreas onde se faz controle de pragas, CPE, e uso de adubos, ADUB, com frequências observadas apresentadas na Tabela 3.1.

Tabela 3.1 : Frequências observadas de Parecer Técnico (Y) por Controle de pragas (CPE) e Utilização de adubo (ADUB).

ADUB	CPE	Y				TOTAL
		1	2	3	4	
0	0	1	4	57	62	114
	1	4	23	40	34	101
1	0	1	17	46	55	119
	1	48	280	312	128	768
TOTAL		54	324	445	279	1102

As estimativas dos parâmetros do modelo politômico e das regressões individuais (Tabela 3.2) apresentam valores aproximados, e de acordo com o teste de Wald rejeita-se a hipótese Ho:β_{ij}= 0, para todo J=0,1,2, e i=1,2 , ao nível α = 5% , com exceção do parâmetro β_{21}, referente ao fator ADUB para o primeiro logito.

Tabela 3.2 : Estimativa e erro padrão dos parâmetros do modelo politômico e das regressões individuais, estatística de Wald e nível de significância para Ho: β_{ij}=0.

COVAR	Polit $\hat{\beta}$	INDIVID. $\hat{\beta}^*$	E.PADRAO $(\hat{\beta})$	E.PADRAO $(\hat{\beta}^*)$	WALD $(\hat{\beta})$	Nivel de (%) Signif.
Const	-4,588	-4,682	0,786	0,799	34,08	0,00
	-2,377	-2,507	0,292	0,307	66,09	0,00
	-0,452	-0,446	0,169	0,168	7,19	0,73
CPE	2,638	2,655	0,742	0,739	12,63	0,04
	2,001	2,045	0,266	0,264	56,68	0,00
	0,850	0,848	0,181	0,181	22,16	0,00
ADUB	0,935	1,034	0,501	0,505	3,48	6.21
	1,139	1.257	0,254	0,258	20,03	0.00
	0,449	0,439	0,189	0,190	5,67	1,17

Os valores da estatística de razão de verossimilhança G, para comparações entre os modelos apresentados e os modelos contendo interações são :

Modelo Politômico G = 4,45 (α=22% , 3 g.l.)

Modelo Individual (η_1) G_1= 0,45 (α=50% , 1 g.l.) .
(η_2) G_2= 0,38 (α=54% , 1 g.l.)
(η_3) G_3= 2,77 (α=10% , 1 g.l.) .

Os resultados mostram que não existe evidências de interações entre os efeitos das variáveis CPE e ADUB.

3.2.5 Seleção de Variáveis Independentes baseado em ajuste por MQP

O método de seleção de variáveis independentes com ajuste de Mínimos Quadrados Ponderados apresentado no capítulo 2, pode ser aplicado às regressões dicotômicas individuais propostas por Begg e Gray (1984), ou diretamente ao modelo logístico politômico onde :

$$\underset{\sim}{Z}_i = \underset{\sim}{X}_i^{*\prime}\,\hat{\underset{\sim}{\xi}} + \underset{\sim}{W}_i^{-1}(\underset{\sim}{\Pi}-\hat{\underset{\sim}{\Pi}}) \qquad , \qquad \hat{\underset{\sim}{\xi}}' = (\hat{\underset{\sim}{\beta}}_1', \ldots, \hat{\underset{\sim}{\beta}}_{k-1}') \qquad ,$$

$$\underset{\sim}{W}_i = \operatorname{diag}\frac{\partial \Pi_i}{\partial \eta_i} = \operatorname{diag}\,(\Pi_d(1-\Pi_d)) \qquad \begin{matrix} d=1,2\ldots,k-1 \\ i=1,2\ldots,n \end{matrix} \quad ,$$

e $$X_i^{*\prime} = \underset{(k-1)\times(k-1)(p+1)}{\left[\begin{array}{c|ccc|ccc|ccc|ccc} 1\ 0\ 0\ldots 0 & x_{i1} & \ldots & x_{ip} & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0\ 1\ 0\ldots 0 & 0 & \ldots & 0 & x_{i1} & \ldots & x_{ip} & \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & & & & & & \vdots & & \vdots & \vdots & & \vdots \\ 0\ 0\ 0\ldots 1 & 0 & \ldots & 0 & 0 & \ldots & 0. & 0 & \ldots & 0 & x_{i1} & \ldots & x_{ip} \end{array}\right]} \quad .$$

Os resultados da secção 2.4, no capítulo anterior, são diretamente extendidos para o modelo politômico onde W é a matriz quadrada bloco diagonal de dimensão n(k-1), formada pelos blocos $\underset{\sim}{W}_i$ (i=1,2...,n).

A aplicação do procedimento acima pode se tornar inviável, dado que a dimensão de X^* cresce com p e k ; por exemplo, para p=10 e k=4 temos

que $\underset{\sim}{X}^{*}$, terá 3n linhas por 3x11=33 colunas, referentes aos parâmetros a serem estimados, gerando 2^{30} possíveis modelos para serem comparados. A utilização das regressões individuais é, nestes casos, aconselhável. Alguns cuidados devem ser considerados, pois ao ajustarmos (k-1) modelos, a seleção pode ocorrer de forma diferente para cada um deles. Sugere-se que, se uma covariável X_j for selecionada para apenas uma das funções logito η_d, por exemplo, $\beta_{d',j}$ seja restrito ao valor 0 para todo d' diferente de d. (Hosmer e Lemeshow 1989 pg 232).

3.3 - MODELO DE ODDS PROPORCIONAIS

3.3.1 Introdução

Seja Y uma variável resposta assumindo k categorias com estrutura ordenada, e seja:

$$\Pi_d = P(Y=d|\underset{\sim}{X}) \qquad e \qquad \gamma_j = \sum_{l=1}^{j} \Pi_l \qquad j,d = 1,2,\ldots,k$$

com $\gamma_k = 1$, a probabilidade pontual e acumulada das categorias de Y .

O modelo de McCullagh (1983), definido por:

$$\ln\left(\frac{P(Y<j|\underset{\sim}{X})}{1-P(Y<j|\underset{\sim}{X})}\right) = \ln\left(\frac{\gamma_j}{1-\gamma_j}\right) = \theta_j + \underset{\sim}{X}'\underset{\sim}{\beta} \qquad (3.4)$$

onde $\underset{\sim}{\theta}$ e $\underset{\sim}{\beta}$ são vetores de parâmetros desconhecidos satisfazendo a desigualdade $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{k-1}$, tem um número reduzido de parâmetros

quando comparado aos modelos (3.1) e (3.3) por encorporar a estrutura de ordem fornecida pelos dados. O nome "odds proporcionais" provem da propriedade da razão de odds do evento $Y \leq j$, para $\underset{\sim}{X}_1$ e $\underset{\sim}{X}_{1'}$, ser independente da categoria j. Isto é :

$$\frac{\gamma_{j_1}/(1-\gamma_{j_1})}{\gamma_{j_1'}/(1-\gamma_{j_1'})} = \frac{\exp(\theta_j + \underset{\sim}{X}_1'\underset{\sim}{\beta})}{\exp(\theta_j + \underset{\sim}{X}_{1'}'\underset{\sim}{\beta})} = \exp \underset{\sim}{\beta}'(\underset{\sim}{X}_1 - \underset{\sim}{X}_{1'}) \quad .$$

Segundo McCullagh (1980), o modelo pode ser derivado de uma variável aleatória contínua não observada Z, tal que se a variável Z estiver no intervalo $\theta_{j-1} \leq Z \leq \theta_j$ então Y=j ocorre, isto é, $\theta_1, \ldots, \theta_{k-1}$ são os pontos de divisão na escala de Z, e :

$$P(Y \leq j) = P(Z \leq \theta_j) = P(Z + \underset{\sim}{X}'\underset{\sim}{\beta} \leq \theta_j + \underset{\sim}{X}'\underset{\sim}{\beta}) =$$

$$= \exp(\theta_j + \underset{\sim}{X}'\underset{\sim}{\beta})/(1+\exp(\theta_j + \underset{\sim}{X}'\underset{\sim}{\beta})) \ .$$

Na prática, em geral, a existência de Z é de difícil verificação, e na realidade não é necessária para a interpretação do modelo, que pode ser feita baseada simplesmente nas categorias da variável resposta Y .

3.3.2 Estimação dos Parâmetros

O logarítimo natural da função de Verossimilhança da amostra com n observações independentes é escrito da forma:

$$\ln L(\underset{\sim}{\theta},\underset{\sim}{\beta},\underset{\sim}{X}) = \sum_{i=1}^{n} \sum_{j=1}^{k} P(Y=j \mid \underset{\sim}{X}_i) =$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} P(Y \leq j \mid \underset{\sim}{X}_i) - P(Y \leq j-1 \mid \underset{\sim}{X}_i)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \left\{ \frac{\exp(\theta_j + \underset{\sim}{X}_i'\underset{\sim}{\beta})}{1+\exp(\theta_j + \underset{\sim}{X}_i'\underset{\sim}{\beta})} - \frac{\exp(\theta_{j-1} + \underset{\sim}{X}_i'\underset{\sim}{\beta})}{1+\exp(\theta_{j-1} + \underset{\sim}{X}_i'\underset{\sim}{\beta})} \right\} . \qquad (3.5)$$

O estimador de máxima Verossimilhança de $\underset{\sim}{\xi}=(\underset{\sim}{\theta},\underset{\sim}{\beta})$ é calculado pela maximização de (3.5), que não fornece solução explícita. Portanto métodos interativos são necessários. Anderson (1981) sugere o método de Quasi-Newton de Gill e Murray (1972) implementado na biblioteca do programa computacional NAG (1978).

Thompson e Baker (1981) mostram que para modelos lineares generalizados com mais de um valor predito para cada observação, como é o

caso do modelo de odds proporcionais, o estimador de MV de $\underset{\sim}{\xi}$ pode ser obtido por MQPI através de função de ligação composta. Green (1984) demonstra que o método, nestes casos, equivale à solução das equações geradas por $\partial L/\partial \hat{\underset{\sim}{\xi}} = 0$ pelo método de Newton Raphson com os Escores de Fisher.

Hastie e Tibshirani (1987) descrevem um algoritimo para ajuste do modelo de Odds Proporcionais utilizando a função de ligação composta na determinação da pseudo variável dependente Z e seu peso W .

Para facilitar a apresentação do algoritimo proposto por Hastie e Tbshirani (1987), a notação introduzida à seguir se refere a uma observação apenas, para o caso de dados não agrupados, e à uma subpopulação em casos de dados agrupados.

Sejam as observações y_i , i=1,2,...N, independentes com distribuição Multinomial $M_k(n_i \underset{\sim}{\Pi}_i, n_i \underset{\sim}{\Sigma}_i)$, onde :

K = número de categorias da variável resposta,

$\underset{\sim}{n}_i = (n_{1i}, n_{2i}, \ldots, n_{ki})$ = número de observações na k-ésima categoria da i-ésima observação,

$\underset{\sim}{\Pi}_i = (\Pi_{1i}, \Pi_{2i}, \ldots, \Pi_{ki})$ tal que $\underset{\sim}{\Pi}_i = E(\underset{\sim}{p}_i)$ com $p_{di} = n_{di}/n._i$ $(d=1, 2, \ldots, k)$

vetor de probabilidades na i-ésima observação, e

$n_i \underset{\sim}{\Sigma}_i = n_i \text{ diag } (\underset{\sim}{\Pi}_i) - \underset{\sim}{\Pi}_i \underset{\sim}{\Pi}_i'$ = matriz de covariância de y_i .

O índice i será omitido para efeito de simplificação.

Seja L o ln da função de verossimilhança da distribuição Multinomial :

$$L = \sum_{j=1}^{k} n_j \ln \Pi_j$$

e

$$\underset{\sim}{\gamma} = \underset{\sim}{H}\ \underset{\sim}{\Pi} = \underset{(k-1)\times(k)}{\begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & 0 \dots & 0 \\ \vdots & & & \vdots \\ 1 & 1 \dots & 1 & 0 \end{bmatrix}} \underset{(k)\times 1}{\begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_3 \end{bmatrix}}$$

Então

$$\frac{\partial \underset{\sim}{\Pi}}{\partial \underset{\sim}{\gamma}} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & & -1 & 1 \\ 0 & 0 & 0 & \dots & 0 & -1 \end{bmatrix} = \underset{\sim}{H}^{-} \quad .$$

Derivando L com respeito à Π_d sob a restrição $\sum_{d=1}^{k} \Pi_d = 1$, obtêm-se , (McCullagh e Nelder 1989 pg 171) :

$$\frac{\partial L}{\partial \Pi_d} = \frac{n_d - \Pi_d \sum_{i=1}^{k} n_i}{\Pi_d}$$

ou:

$$\frac{\partial L}{\partial \Pi_d} = \underset{\sim}{n}\ \Sigma^{-}\ (\underset{\sim}{p} - \underset{\sim}{\Pi})$$

com $\Sigma^{-} = \text{diag}(\Pi_i^{-1})$, matriz inversa generalizada de Σ.

Logo :

$$\frac{\partial L}{\partial \underset{\sim}{\gamma}} = \left(\frac{\partial \underset{\sim}{\Pi}}{\partial \underset{\sim}{\gamma}}\right)' \left(\frac{\partial L}{\partial \underset{\sim}{\Pi}}\right) = (H^{-})' \underset{\sim}{n}\, \Sigma^{-}(\underset{\sim}{p}-\underset{\sim}{\Pi}) = u$$

e

$$\frac{\partial L}{\partial \underset{\sim}{\xi}} = \left(\frac{\partial \underset{\sim}{\eta}}{\partial \underset{\sim}{\xi}}\right)\left(\frac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\eta}}\right)\left(\frac{\partial L}{\partial \underset{\sim}{\gamma}}\right) = \underset{\sim}{D}'\underset{\sim}{u} \qquad \text{onde :}$$

(1) $\xi' = (\theta_1, \ldots, \theta_{k-1}, \beta_1, \ldots, \beta_p) = (\underset{\sim}{\theta}, \underset{\sim}{\beta})$

(2) $\underset{\sim}{\eta} = \underset{\sim}{\theta} + 1\ \underset{\sim}{X}'\underset{\sim}{\beta} \qquad\qquad \underset{\sim}{X}' = (\alpha_1, \alpha_2, \ldots, \alpha_p)$

(3) $\underset{\sim}{D} = C\underset{\sim}{X}^{*}$

onde $\underset{\sim}{C} = \dfrac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\eta}} = \text{diag}(\gamma_d(1-\gamma_d))$ e $\underset{\sim}{X}^{*} = [I \,|\, 1\ \underset{\sim}{X}']$

para d=1,2,...,k-1.

e

(4) $| = \underset{\sim}{D}'\underset{\sim}{A}\ \underset{\sim}{D}$ onde $\underset{\sim}{A} = E\left(\dfrac{\partial L}{\partial \underset{\sim}{\gamma}}\right)\left(\dfrac{\partial L}{\partial \underset{\sim}{\gamma}}\right)' = \underset{\sim}{H}^{-}{}'\underset{\sim}{n}\,\Sigma^{-}\,\underset{\sim}{H}^{-}$.

A solução do sistema $\underset{\sim}{D}'\underset{\sim}{u} = 0$ pelo método dos Escores de Fisher equivale à solução por MQPI do modelo (Green 1984):

$$\underset{\sim}{Z} = \underset{\sim}{C}^{-1}(\underset{\sim}{\gamma}-\hat{\underset{\sim}{\gamma}}) + \underset{\sim}{X}^{*}{}'\hat{\underset{\sim}{\xi}} \quad \text{em} \quad \underset{\sim}{X}^{*} \quad \text{com peso} \quad \underset{\sim}{W}=\underset{\sim}{C}'\underset{\sim}{A}\ \underset{\sim}{C} \ . \qquad (3.6)$$

Portanto, fazendo $\hat{\underset{\sim}{\eta}} = \underset{\sim}{X}^{*}\hat{\underset{\sim}{\xi}}$, e $\hat{\underset{\sim}{\gamma}}=1/1+\exp(-\hat{\underset{\sim}{\eta}})$ onde $\hat{\underset{\sim}{\xi}}$ é o estimador de M.V. de $\underset{\sim}{\xi}$, o estimador $\tilde{\underset{\sim}{\xi}}$ de M.Q.P. de $\underset{\sim}{\xi}$, da Regressão Linear Ponderada de $\underset{\sim}{Z}=\underset{\sim}{X}^{*}\underset{\sim}{\xi}$, com peso W é idêntico ao estimador de M.V. $\hat{\underset{\sim}{\xi}}$.

O procedimento Logistic do pacote estatístico SAS, e o GLIM, utilizam o método de Mínimos Quadrados Ponderados Iterativos para o cálculo das estimativas de MV de $\underset{\sim}{\xi}$.

A avaliação da qualidade do ajuste do modelo, assim como da contribuição de cada covariável é feita através das estatísticas de Razão de Verossimilhança e de Wald respectivamente como descrito no modelo Logístico Dicotômico, secção (2.3).

O modelo de odds proporcionais foi ajustado no exemplo 3.2, onde são consideradas as variáveis referentes ao controle de pragas (CPE) e a utilização de adubos (ADUB), com frequências descritas na Tabela 3.1.

EXEMPLO 3.2:

Seja o modelo de odds proporcionais ajustado aos dados da Tabela 3.1, onde:

$$\eta_j = \ln\left(\frac{\gamma_j}{1-\gamma_j}\right) = \theta_j + \underset{\sim}{X}'\underset{\sim}{\beta} = \theta_1 + \theta_2 + \theta_3 + \beta_1(adub) + \beta_2(cpe)$$

j=1,2,3.

O vetor de parâmetros $(\theta_1, \theta_2, \theta_3, \beta_1, \beta_2)$ representa :

θ_1 - ln da odds da probabilidade da área estar ótima versus estar boa, ruim ou péssima, quando não se faz adubação nem controle de pragas, (*cpe=frad*= 0).

θ_2 - idem para o ln da odds da probabilidade da área estar ótima ou boa, versus ruim ou péssima.

θ_3 - idem para o ln da odds da probabilidade da área estar ótima, boa ou ruim , versus estar péssima.

β_1 - incremento para os logitos devido ao controle de pragas.

β_2 - incremento para os logitos devido ao uso de adubo.

Os resultados do ajuste estão dispostos na tabela 3.3 :

Tabela 3.3: Ajuste do modelo de odds proporcionais dos dados da tabela 3.1. =

Parametro	g.l	Parametro estimado	Erro Padrao	WALD	Nivel Sign.%
θ_1	1	-4,731	0,215	484,93	0,01
θ_2	1	-2,334	0,169	191,14	0,01
θ_3	1	-0,406	0,152	7,10	0,77
β_1 (*adub*)	1	1,334	0,155	73,95	0,01
β_2 (*cpe*)	1	0,681	0,156	19,02	0,01

$G = -2\log \{ L(\underset{\sim}{\beta}_{HO})/L(\underset{\sim}{\beta})\} = 144,833$ (2 graus de liberdade).

Para Ho : $\beta_1=\beta_2= 0$, o teste $X^2_{(2\ g.l)}$ resulta na rejeição de Ho ao nível $\alpha=1\%$, sugerindo que pelo menos uma das variáveis está contribuindo de modo significativo para o modelo. O teste de Wald (colunas 5 e 6 da tabela 3.3) confirma este resultado, rejeitando as hipóteses que β_1 ou β_2 sejam iguais à zero.

3.3.3 Seleção de Variáveis Independentes baseada em ajustes por MQP

Baseado nos resultados de Green (1984) para estimação dos parâmetros do modelo, e no algorítimo descrito por Hastie e Tibshirani (1987), é apresentada nesta secção a extensão do método de seleção de variáveis independentes de Hosmer e Lemeshow para o modelo de odds proporcionais. Os resultados são demonstrados para o caso particular de K=3 categorias.

Para variável resposta tricotômica, onde k=3, a pseudo variável dependente Z , e seu peso W **(3.6)** são definidos por :

$$\underset{\sim}{Z} = \underset{\sim}{C}^{-1}(\underset{\sim}{\gamma}-\hat{\underset{\sim}{\gamma}}) + \underset{\sim}{X}^{*}\,\underset{\sim}{\xi}$$

$$\underset{\sim}{Z}=\begin{bmatrix} 1/\gamma_1(1-\gamma_1) & 0 \\ 0 & 1/\hat{\gamma}_2(1-\hat{\gamma}_2) \end{bmatrix}\begin{bmatrix} \gamma_1-\gamma_1 \\ \gamma_2-\hat{\gamma}_2 \end{bmatrix}+\left[\begin{array}{cc:ccc} 1 & 0 & \alpha & \alpha & \cdots \alpha \\ 0 & 1 & \alpha_1^1 & \alpha_2^2 & \cdots \alpha_p^p \end{array}\right]\begin{bmatrix} \theta \\ \theta^1 \\ \theta_2 \\ \hdashline \underset{\sim}{\beta} \end{bmatrix} \qquad (3.7)$$

$$\underset{\sim}{W} = \begin{bmatrix} (1-\hat{\gamma}_1)^2\,\hat{\gamma}_2\,\hat{\gamma}_1 & -\hat{\gamma}_1(1-\hat{\gamma}_1)\hat{\gamma}_2(1-\hat{\gamma}_2) \\ -\hat{\gamma}_1(1-\hat{\gamma}_1)\hat{\gamma}_2(1-\hat{\gamma}_2) & \hat{\gamma}_2{}^2(1-\hat{\gamma}_1)(1-\hat{\gamma}_2) \end{bmatrix} \frac{1}{(\hat{\gamma}_2-\hat{\gamma}_1)},$$

onde :

$$\underset{\sim}{H} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \qquad \underset{\sim}{H}^{-} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} ,$$

$$\underset{\sim}{\Pi} = (\Pi_1, \Pi_2, \Pi_3) \qquad \underset{\sim}{\gamma} = (\gamma_1, \gamma_2) = (\Pi_1, \Pi_1+\Pi_2) ,$$

$$\Pi_1+\Pi_2+\Pi_3 = 1 ,$$

$$\Sigma = \begin{bmatrix} \Pi_1(1-\Pi_1) & -\Pi_2\Pi_1 & -\Pi_3\Pi_1 \\ -\Pi_1\Pi_2 & \Pi_2(1-\Pi_2) & -\Pi_3\Pi_2 \\ -\Pi_1\Pi_3 & -\Pi_2\Pi_3 & \Pi_3(1-\Pi_3) \end{bmatrix} \qquad \Sigma^{-} = \begin{bmatrix} 1/\Pi_1 & 0 & 0 \\ 0 & 1/\Pi_2 & 0 \\ 0 & 0 & 1/\Pi_3 \end{bmatrix},$$

$$\underset{\sim}{C} = \begin{bmatrix} \gamma_1(1-\gamma_1) & 0 \\ 0 & \gamma_2(1-\gamma_2) \end{bmatrix} .$$

Se de um conjunto incial de p variáveis independentes, deseja-se selecionar um subconjunto de tamanho q<p de maneira que se obtenha um modelo reduzido com bom ajuste , a seleção pode ser feita por um programa de regressão linear com ajuste de Mínimos Quadrados Ponderados à partir do ajuste do modelo completo e do estimador de M.V. $\hat{\xi}$ de ξ.

A Regressão Linear Ponderada de $\underset{\sim}{Z}$ em $\underset{\sim}{X}^{*}$, com peso $\underset{\sim}{W}$, onde $\underset{\sim}{Z}$ e $\underset{\sim}{W}$ são obtidos por $\hat{\underset{\sim}{\xi}}$, fornece os valores necessários para a comparação de submodelos através do coeficiente de Mallow como descrito na secção (2.4) .

A soma de quadrados de resíduos gerada pela Regressão Linear nas p variáveis independentes equivale ao X^2 de Pearson, possibilitando assim a utilização de algorítimos de seleção de submodelos baseados nas diferenças das somas de quadrados residuais entre os submodelos e o modelo completo.

Seja SQR(p) a soma de quadrados residuais obtidos do ajuste por MQP do modelo $Z=[I\,|\,1X]'\xi$, onde $\underset{\sim}{Z}=\underset{\sim}{C}^{-1}(\underset{\sim}{\gamma}-\hat{\underset{\sim}{\gamma}})+\underset{\sim}{X}^{*}{}'\hat{\underset{\sim}{\xi}}$ e $\hat{\gamma}=1/(1+\exp-(\underset{\sim}{X}^{*}{}'\hat{\underset{\sim}{\xi}}))$:

$$SQR(p)= [\underset{\sim}{Z} - \hat{\underset{\sim}{Z}}(p)]'\underset{\sim}{W}\,[\underset{\sim}{Z} - \hat{\underset{\sim}{Z}}(p)]$$

$$=\sum_{i=1}^{n} (\underset{\sim}{\gamma}_i - \hat{\underset{\sim}{\gamma}}_i)\, \underset{\sim}{C}_i^{-1}\underset{\sim}{W}_i\underset{\sim}{C}_i^{-1}(\underset{\sim}{\gamma}_i - \underset{\sim}{\gamma}_1)$$

$$=\sum_{i=1}^{n} \frac{(\gamma_{1i}-\hat{\gamma}_{1i})^2\,\hat{\gamma}_{2i}}{\hat{\gamma}_{1i}(\hat{\gamma}_{2i}-\hat{\gamma}_{1i})}+\frac{(\gamma_{2i}-\hat{\gamma}_{2i})^2(1-\hat{\gamma}_{1i})}{(1-\hat{\gamma}_{2i})(\hat{\gamma}_{2i}-\hat{\gamma}_{1i})} + \frac{-2(\gamma_{1i}-\hat{\gamma}_{1i})(\gamma_{2i}-\hat{\gamma}_{2i})}{(\hat{\gamma}_{2i}-\hat{\gamma}_{1i})} .$$

Agora, desenvolvendo o X^2 de Pearson em termos de $\underset{\sim}{\gamma}$, para uma observação obtem-se :

$$X^2 = \frac{(\Pi_1-\hat{\Pi}_1)^2}{\hat{\Pi}_1} + \frac{(\Pi_2-\hat{\Pi}_2)^2}{\hat{\Pi}_2} + \frac{(\Pi_3-\hat{\Pi}_3)^2}{\hat{\Pi}_3}$$

$$= \frac{(\gamma_1-\hat{\gamma}_1)^2}{\hat{\Pi}_1} + \frac{(\gamma_2-\gamma_1-\hat{\gamma}_2+\hat{\gamma}_1)^2}{\hat{\Pi}_2} + \frac{(1-\gamma_2-1+\hat{\gamma}_2)^2}{(1-\hat{\gamma}_2)}$$

$$= \frac{(\gamma_1-\hat{\gamma}_1)^2}{\hat{\Pi}_1} + \frac{(\gamma_2-\hat{\gamma}_2)^2}{\hat{\Pi}_3} + \frac{(\gamma_2-\hat{\gamma}_2)^2}{\hat{\Pi}_2} + \frac{(\gamma_1-\hat{\gamma}_1)^2}{\hat{\Pi}_2} - \frac{2(\gamma_1-\hat{\gamma}_1)(\gamma_2-\hat{\gamma}_2)}{\hat{\Pi}_2}$$

$$= \frac{(\gamma_1-\hat{\gamma}_1)^2(\hat{\Pi}_1+\hat{\Pi}_2)}{\hat{\Pi}_1\hat{\Pi}_2} + \frac{(\gamma_2-\hat{\gamma}_2)^2(\hat{\Pi}_2+\hat{\Pi}_3)}{\hat{\Pi}_2\hat{\Pi}_3} - \frac{2(\gamma_1-\hat{\gamma}_1)(\gamma_2-\hat{\gamma}_2)}{\hat{\Pi}_2}$$

$$\frac{(\hat{\gamma}_1-\hat{\gamma}_1)^2\,\hat{\gamma}_2^2}{\hat{\gamma}_1(\hat{\gamma}_2-\hat{\gamma}_1)} + \frac{(\gamma_2-\hat{\gamma}_2)^2(1-\hat{\gamma}_1)}{(\hat{\gamma}_2-\hat{\gamma}_1)(1-\hat{\gamma}_1)} - \frac{2(\gamma_1-\hat{\gamma}_1)(\gamma-\hat{\gamma}_2)}{(\hat{\gamma}_2-\hat{\gamma}_1)} .$$

O ajuste de um submodelo com q<p variáveis independentes é calculado particionando-se a matrizes X^*, ξ^* e | tal que :

$$\tilde{Z}(q)=\underset{\sim}{X}_1^{*}\,\hat{\underset{\sim}{\xi}}_1 \qquad \text{onde} \qquad \hat{\xi}=(\hat{\theta}_1,\hat{\theta}_2,\hat{\underset{\sim}{\beta}}_1 \vdots \hat{\underset{\sim}{\beta}}_2)=$$

$$=(\hat{\underset{\sim}{\theta}},\hat{\beta}_1,\ldots\hat{\beta}_q \vdots \hat{\beta}_{q+1},\ldots\hat{\beta}_p)=(\hat{\xi}_1,\hat{\xi}_2) ,$$

$$\underset{\sim}{X}^{*\prime}=(\underset{\sim}{X}_1^{*},\underset{\sim}{X}_2^{*})' = \left[\begin{array}{cccc|ccc} 1 & 0 & \alpha_{11} & \ldots\alpha_{1q} & \alpha_{1(q+1)} & \ldots & \alpha_{1p} \\ 0 & 1 & \alpha_{11} & \ldots\alpha_{1q} & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & \alpha_{n1} & \ldots\alpha_{nq} & \alpha_{n(q+1)} & \ldots & \alpha_{np} \end{array}\right],$$

$$| = \left[\begin{array}{cc} \underset{\sim}{X}_1^{*}\underset{\sim}{W}\,\underset{\sim}{X}_1^{*\prime} & \underset{\sim}{X}_1\underset{\sim}{W}\,\underset{\sim}{X}_2^{*\prime} \\ \underset{\sim}{X}_2^{*}\underset{\sim}{W}\,\underset{\sim}{X}_1^{*\prime} & \underset{\sim}{X}_2^{*}\underset{\sim}{W}\,\underset{\sim}{X}_2^{*\prime} \end{array}\right] = \left[\begin{array}{cc} I_{11} & I_{12} \\ I_{21} & I_{22} \end{array}\right] ,$$

$$\tilde{\underset{\sim}{\xi}}_1=(I_{11})^{-1}\underset{\sim}{X}^{*}\underset{\sim}{W}\,\underset{\sim}{Z} \quad .$$

Substituindo (3.7) obtem-se :

$$\tilde{\underset{\sim}{\xi}}_1 = (I_{11})^{-1}\underset{\sim}{X}_1^{*}\underset{\sim}{W}\,[\underset{\sim}{C}^{-1}(\underset{\sim}{\gamma}-\hat{\underset{\sim}{\gamma}}) + \underset{\sim}{X}^{*}\,'\hat{\underset{\sim}{\xi}}\,]$$

$$=(I_{11})^{-1}\underset{\sim}{X}_1^{*}\underset{\sim}{W}\,\underset{\sim}{C}^{-1}(\underset{\sim}{\gamma}-\hat{\underset{\sim}{\gamma}})+(I_{11})^{-1}\underset{\sim}{X}_1^{*}\underset{\sim}{W}(\underset{\sim}{X}_1^{*}\,'\hat{\underset{\sim}{\xi}}_1+\underset{\sim}{X}_2^{*}\,'\hat{\underset{\sim}{\xi}}_2)$$

$$=\hat{\underset{\sim}{\xi}}_1+ I_{11}^{-1}I_{12}\hat{\underset{\sim}{\xi}}_2 \quad .$$

E finalmente, a soma de quadrados residuais do submodelo ajustado é :

$$SQR(q)=[\underset{\sim}{Z}-\tilde{\underset{\sim}{Z}}(q)]'\underset{\sim}{W}\,[\underset{\sim}{Z}-\tilde{\underset{\sim}{Z}}(q)]$$

$$=\underset{\sim}{Z}'\underset{\sim}{W}\,\underset{\sim}{Z} - 2[\tilde{\underset{\sim}{\xi}}_1'\underset{\sim}{X}_1^{*}]\underset{\sim}{W}[\underset{\sim}{C}^{-1}(\underset{\sim}{\gamma}-\hat{\underset{\sim}{\gamma}})+\underset{\sim}{X}^{*}\,'\hat{\underset{\sim}{\xi}}\,) + \tilde{\underset{\sim}{\xi}}_1'\underset{\sim}{X}_1^{*}\underset{\sim}{W}\,\underset{\sim}{X}_1^{*}\,'\tilde{\underset{\sim}{\xi}}_1$$

$$=X^2 - \hat{\underset{\sim}{\xi}}'\,|\,\hat{\underset{\sim}{\xi}} + \tilde{\underset{\sim}{\xi}}_1'I_{11}\tilde{\underset{\sim}{\xi}}_1= \underset{\sim}{X}^2+ \hat{\underset{\sim}{\xi}}_2'[I_{22}-I_{21}I_{11}^{-1}I_{12}]\hat{\underset{\sim}{\xi}}_2$$

$$= X^2+ \lambda^{*} \quad .$$

Os resultados acima garantem a utilização de ajustes por MQP para seleção de variáveis independentes no modelo de odds proporcionais, onde a comparação entre modelos pode ser efetuada através do coeficiente C(q), ou da estatística de Wald, λ^{*}, como aproximação da estatística de razão de verossimilhança, λ. (Lawless e Singhal 1978).

A extensão dos resultados para variáveis resposta com k=4 categorias é direta onde W é uma matriz simétrica definida por:

$$\begin{bmatrix} \dfrac{\hat{\gamma}_1^2(1-\hat{\gamma}_1)^2\hat{\gamma}_2}{\hat{\gamma}_1(\hat{\gamma}_2-\hat{\gamma}_1)} & -\dfrac{\hat{\gamma}_1(1-\hat{\gamma}_1)\hat{\gamma}_2(1-\hat{\gamma}_2)}{(\hat{\gamma}_2-\hat{\gamma}_1)} & 0 \\ & \dfrac{\hat{\gamma}_2^2(1-\hat{\gamma}_2)^2(\hat{\gamma}_3-\hat{\gamma}_1)}{(\hat{\gamma}_2-\hat{\gamma}_1)(\hat{\gamma}_3-\hat{\gamma}_2)} & \dfrac{\hat{\gamma}_2(1-\hat{\gamma}_2)\hat{\gamma}_3(1-\hat{\gamma}_3)}{(\hat{\gamma}_3-\hat{\gamma}_2)} \\ & & \dfrac{\hat{\gamma}_3^2(1-\hat{\gamma}_3)^2(1-\hat{\gamma}_2)}{(\hat{\gamma}_3-\hat{\gamma}_2)(1-\hat{\gamma}_3)} \end{bmatrix}$$

3.4 STEREOTYPE MODEL

Anderson (1984) em seu trabalho observa que quando as categorias de Y não estão diretamente relacionadas à intervalos de uma variável contínua, e o interesse do estudo é verificar se o vetor $\underset{\sim}{X}$ apresenta uma relação ordenada com as categorias da variável resposta, modelos baseados em restrições impostas aos parâmetros do modelo logístico Politômico podem ser mais adequados.

O autor propõe novos modelos, com dimensão variando de 1 à min(k-1,p), formando uma hierarquia de modelos que, conforme informações de $\underset{\sim}{X}$, podem conter estrutura ordenada, intermediária ou nominal. A dimensão do modelo é determinada pelo número de funções lineares necessárias para se descrever a relação observada entre $\underset{\sim}{X}$ e Y.

O modelo unidimensional supõe que os efeitos das variáveis independentes no j-ésimo logito do modelo (3.1), diminuem de maneira uniforme à medida que a categoria j se aproxima da categoria padrão k, segundo parâmetro Φ_j. Essa relação ordenada, chamada por Anderson de paralelismo entre os parâmetros $\underset{\sim}{\beta}_d$ do modelo politômico, é considerada no modelo unidimensional fazendo $\underset{\sim}{\beta}_d = \Phi_d \underset{\sim}{\beta}$, onde $\underset{\sim}{\beta}$ e Φ_d, d=1,2...,k, são novos parâmetros desconhecidos, e o modelo é definido por :

$$\Pi_d(\underset{\sim}{X}_1) = \frac{\exp(\beta_{0d} - \Phi_d \underset{\sim}{X}_1' \underset{\sim}{\beta})}{\sum_{j=1}^{k} \exp(\beta_{0j} - \Phi_j \underset{\sim}{X}_1' \underset{\sim}{\beta})} \quad . \qquad (3.8)$$

A relação ordenada das categorias de Y com respeito ao vetor $\underset{\sim}{X}$, é confirmada com a verificação de que os parâmetros $\underset{\sim}{\Phi}$ sejam decrescentes, isto é :

$$\Phi_1 \equiv 1 \geq \Phi_2 \geq \ldots \geq \Phi_k \equiv 0 \qquad (3.9)$$

Se a desigualdade (3.9) é satisfeita, modelos ordenados como os modelos (3.3) e (3.8), são adequados para se estudar as associações entre as variáveis independentes e a variável resposta. Caso contrário, outros modelos devem ser considerados.

O próximo passo na construção do modelo unidimensional é a verificação de que o vetor $\underset{\sim}{X}$, além de confirmar a ordem assumida à priori, identifica ou distingue duas categorias adjacentes, isto é, se Y=j e Y=j+1 são distinguiveis com respeito à $\underset{\sim}{X}$. Isto pode ser feito por testes de hipóteses como : $H_0 : \beta_j = \beta_{j+1} \equiv H_0 : \Phi_j = \Phi_{j+1}$. Caso não exista evidência para rejeição de H_0, outros modelos devem ser avaliados, considerando a combinação das categorias j e j+1, ou reavaliando a contribuição de cada covariável para o ajuste do modelo e estimativa de Φ_j.

O modelo (3.8) tem dimensão um, e pode ser extendido para modelos com dimensão de até q=min(k-1,p). O modelo de Anderson com dimensão igual à dois é definido por :

$$\Pi_d(\underset{\sim}{X}_i) = \frac{\exp(\beta_{0d} - \Phi_{d\sim 1}\underset{\sim}{X}'\underset{\sim}{\beta} - \psi_{d\sim 1}\underset{\sim}{X}'\underset{\sim}{\alpha})}{\sum_k \exp(\beta_{0j} - \Phi_{d\sim 1}\underset{\sim}{X}'\underset{\sim}{\beta} - \psi_{d\sim i}\underset{\sim}{X}'\underset{\sim}{\alpha})}$$

onde $\underset{\sim d}{\beta} = -\Phi_{d\sim}\underset{\sim}{\beta} - \psi_{d\sim}\underset{\sim}{\alpha}$

Métodos de estimação por MV e construção de testes de hipóteses para os

parâmetros dos modelos apresentados são discutidos no trabalho de Anderson (1984).

Resultados obtidos por Holtbrügge e Shumacher (1991), mostram que as estimativas dos parâmetros do modelo de odds proporcionais são menos viciadas do que no modelo de Anderson, e que, para esse último, o vício diminui quando categorias adjacentes são combinadas. O mesmo comportamento é observado com respeito aos testes de significância para os coeficientes da regressão.

A proposta desse trabalho, no que se refere ao modelo de Anderson, é utilizar a estrutura sugerida do modelo unidimensional para, através de análises exploratórias, apresentar uma alternativa que venha auxiliar no processo de escolha de modelos e subconjuntos de variáveis independentes para estudos com variável resposta ordenada.

Greenwood e Farewell (1988) ajustam o modelo (3.8) à partir de estimativas do modelo logístico politômico. Os dados utilizados no trabalho são de um estudo sobre fatores que influenciam na recuperação de pacientes submetidos à tranplante de rins, sendo a variável resposta composta de k=3 categorias e o vetor de variáveis independentes de tamanho p=7. Inicialmente ajustou-se o modelo Logístico Politômico através do procedimento CATMOD do S.A.S. obtendo os estimadores:

$$\hat{\beta}_{10}, \hat{\beta}_{11}, \hat{\beta}_{12}, \ldots, \hat{\beta}_{17}, \hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{22}, \ldots, \hat{\beta}_{27} \ .$$

Os autores obtiveram estimativa do escalar Φ_2 por média aritimética das razões dos parâmetros que pelo teste de Wald foram significativos para ambos os logitos. De maneira que, considerando $\hat{\underset{\sim}{\beta}} = \hat{\underset{\sim}{\beta}}_1$, $\hat{\Phi}_1 = 1$ e $\hat{\Phi}_3 = 0$, calculou-se :

$$\hat{\Phi}_2 = \left[(\hat{\beta}_{11}/\hat{\beta}_{21})+(\hat{\beta}_{12}/\hat{\beta}_{22})+(\hat{\beta}_{13}/\hat{\beta}_{23})+(\hat{\beta}_{16}/\hat{\beta}_{26})+(\hat{\beta}_{17}/\hat{\beta}_{27})\right]\frac{1}{5}$$

O estudo descritivo das razões $r_1=\hat{\beta}_{11}/\hat{\beta}_{21},\ldots,r_p=\hat{\beta}_{1p}/\hat{\beta}_{2p}$, auxilia o processo de seleção de submodelos . Para o ajuste do modelo unidimensional, as variáveis que apresentarem r_j $(j=1,2\ldots,p)$ aproximadamente iguais, respeitando a relação (3.9), evidenciam a suposição de "paralelismo" assumida pelo modelo.

No exemplo **(3.1)** $r_{21}=0,76$, $r_{22}= 0,32$, $r_{31}= 1,22$ e $r_{32}=0,48$. Fazendo :

$$\hat{\Phi}_2 = (0,76+1,22)/2 = 0,99 \quad ,$$

e

$$\hat{\Phi}_3 = (0,32+0,48)/2 = 0,40 \quad ,$$

temos que $\hat{\Phi}_1$ e $\hat{\Phi}_2$ são bem próximos , evidenciando que , com respeito à CPE e ADUB os logitos das categorias *otimo e bom* não são diferenciados, ao contrário das categorias *otimo e ruim*. Dois fatores devem ser considerados aqui : (1) a baixa frequência observada na primeira categoria de Y , (2) a não existência de associação da covariável ADUB com relação à primeira função logito (Wald=3,48 α=6,2%) .

Combinando as categorias 1 e 2 , de Y, em uma única categoria, o ajuste do modelo politômico, gera os seguintes resultados:

$$\hat{\underset{\sim}{\beta}} = (-2,27;\ -0,45;\ 2,08;\ 0,85;\ 1,11;\ 0,45) \quad , \quad \hat{\Phi}_2 = 0,41 \ .$$

Resumindo, através do estudo dos valores de $R_d=(r_{d1},\ldots,r_{dp})$, é possível a seleção de variáveis independentes que tenham uma relação ordenada com Y, isto é, verifica-se quais dos fatores são importantes para a descrição da relação existente entre X e Y através de modelos ordenados.

Os dados sobre plantação de cacau foram analisados, no capítulo 4, secção 4.3, onde foi aplicado o método de seleção de variáveis independentes de Hosmer e Lemeshow ao modelo politômico e às regressões individuais. Através de análises descritivas aplicadas às razões dos coeficientes estimados, verificou-se quais os fatores dão suporte à suposição de paralelismo no modelo unidimensional de Anderson, e que diferenciam as categorias da variável resposta Parecer Técnico (PAR_TEC).

Também é apresentado no próximo capítulo, aplicações do método de seleção para os modelos dicotômico e de odds proporcionais, nas secções **(4.2)** e **(4.4)**, respectivamente. Na secção **(4.6)** é feito um resumo dos pacotes estatísticos mais conhecidos que podem ser utilizados para aplicação das técnicas mencionadas.

CAPITULO 4

APLICAÇÃO

4.1 INTRODUÇÃO :

Neste capítulo são feitas aplicações das técnicas de seleção de variáveis independentes para modelos logísticos, apresentadas nos capítulos anteriores.

O conjunto de dados utilizados é proveniente de um estudo realizado pelo Centro de Pesquiza e Departamento de Extensão da CEPLAC (Comissão Executiva do Plano da Lavoura Cacaueira), cujo objetivo foi de fazer uma avaliação qualitativa de áreas de cacau implantadas no período de 1976 à 1985, duração do programa nacional de incentivo à expansão de novas áreas de cacau nos estados da Bahia, Espírito Santo e Amazônia. Aqui são considerados somente os dados da Bahia .

A necessidade da avaliação qualitativa das novas áreas surgiu da preocupação de que, apesar das metas quantitativas, em hectares, terem

sido atingidas, o estímulo à expansão tenha levado à plantação de cacau em áreas impróprias, trazendo como consequência uma possível queda na produtividade e a necessidade de maiores gastos com insumos e tratamento do solo.

Partindo-se de informações levantadas pelos Escritórios Locais do Departamento de Extensão , foi possível o registro de todas as áreas plantadas no período, (1976-1985), para cada ano, fazenda e município. Com estas informações elaborou-se um plano amostral estratificado por tamanho de área plantada em 13 regiões da Bahia, vizando uma representatividade de todo o estado .

Num total aproximado de 26.247 áreas implantadas no período, foram avaliadas 1174 áreas selecionadas aleatoriamente dentro das regiões, que diferenciavam-se basicamente por localização e características climáticas.

As informações foram levantadas através de questionários aplicados por equipes de agrônomos do Centro de Pesquisa e técnicos agrícolas do dep. de Extensão, que foram até às áreas amostradas onde além de entrevistarem o proprietário e/ou administrador da fazenda, avaliaram visualmente a área dando uma nota pela sua qualidade técnica.

Será feita a análise dos dados com utilização de modelagem através da transformação logística, com o objetivo de se estudar como os diversos fatores levantados estão relacionados com o Parecer Técnico dado pelos entrevistadores, buscando além da descrição dos dados o conhecimento dos fatores mais importantes que levaram essas implantações ao sucesso ou ao fracasso.

A Tabela 4.1 faz uma descrição resumida das variáveis envolvidas na pesquisa que foram consideradas no trabalho. As dezesseis variáveis descritas na tabela, foram selecionadas de um conjunto inicial de 46 contidas no questionário. Os critérios utilizados neste primeiro passo da seleção foram baseados na frequência de respostas obtidas e na confiabilidade da informação.

Tabela 4.1: Descrição dos Fatores Incluídos no Estudo

FATORES	NOTAÇÕES	DESCRIÇÃO
TIPO DE SOLO	T_SOLO	1= Latossolo; 2= nÃo Latossolo; 3= Ambos
FERTILIDADE DO SOLO	FER	1= Baixa; 2= Média; 3= Alta
PROFUNDIDADE DO SOLO	PROF	1= Raso; 2= Médio; 3= Profundo
SOMBREAMENTO PROVISÓRIO	PROVISOR	1= Bom; 2= Excessivo; 3= Deficiente; 4= Inexistente (NTEM)
SOMBREAMENTO DEFINITIVO	DEFINIT	1= Bom; 2= Excesso; 3= Deficiente
ADUBAÇÃO	ADUB	0= Não Aduba; 1= Aduba
FREQUÊNCIA DE ADUBAÇÃO	FRAD	1= Esporádica; 2= À cada 3 anos; 3= À cada 2 anos; 4= Todo ano
CONTROLE DE PODRIDÃO PARDA	CPP	0= Não faz Controle; 1= Faz Controle
CONTROLE DE PRAGAS	CPE	0= Não faz Controle; 1= Faz Controle
NÚMERO DE LIMPEZAS NA ÁREA (POR ANO)	LIM	0= Nunca; 1= 1 vez; 2= 2 vezes; 3= 3 vezes 4= Mais de 3 vezes
ESPASSAMENTO ENTRE AS PLANTAS(DENSIDADE)	DEN	0= Com Falhas; 1= Sem Falhas
IDADE DA PLANTAÇÃO EM 1986	IDADE	1, 2, ... 9 anos
TAMANHO DA ÁREA IMPLANTADA (HECTARES)	ESTRATO	1= (0-4); 2= (5-9); 3= (10-14); 4= (15-19); 5= (20-29); 6= (30-39); 7= (40-69); 8= (70-100); 9= (100-)
SE TEVE PROBLEMA COM MÃO DE OBRA NA ÉPOCA DO PLANTIO	PMO	0= Teve Problema; 1= NÃo Teve Problema
Nº DE VISITAS DO PROPRIETARIO À ÁREA	FRVIS	1= Esporádica; 2= Mensal; 3= Semanal; 4= Diária
FREQUÊNCIA DE ASSISTÊNCIA TÉCNICA	FRASS	0= Não tem Assistência Técnica; 1= Esporádica; 2= 1 vez p/ano; 3= 2vezes p/ano; 4= 3 vezes p/ano; 5= 4vezes p/ano; 6= + de 4 vezes p/ano

Após esta seleção preliminar, foi feito um estudo descritivo através de gráficos e tabelas de frequências onde se decidiu por eliminar as variáveis FRAD e CPP. A primeira devido ao número relativamente alto de não respostas (≅20%), apesar de apresentar relação linear com os logitos das probabilidades das categorias *ótimo, bom e regular* , com relação à categoria *péssimo* . A variável CPP foi também excluída por estar fortemente associada à variável CPE. O valor de $\psi_{(CPP=0,CPE=0)}=7,58$ mostra que onde não ocorre controle de pragas, a odds da probabilidade de que também não ocorra controle de podridão parda é aproximadamente 8 vezes maior do que nas áreas com controle de pragas. Na amostra, 94% das áreas onde não se fez controle de pragas também não foi feito controle de podridão parda.

À variável FRASS aplicou-se a transformação raiz quadrada com o objetivo de diminuir sua variabilidade e de obter uma relação mais próxima da linear com as funções logito.

Dada a estrutura ordenada das categorias dos fatores FER e PROF, decidiu-se por incorporá-los ao modelo como variáveis contínuas, evitando assim um aumento no número de parâmetros do modelo pela criação de novas variáveis dicotômicas.

A Tabela 4.2 descreve as variáveis independentes e parametrizações incluídas no processo de seleção de melhores submodelos. Dos 1173 questionários aplicados, 73 apresentam informações incompletas, restando n=1100 observações utilizadas no trabalho.

Nas seções 4.2 , 4.3 e 4.4 serão apresentadas aplicações do método de seleção de variáveis de Hosmer e Lemeshow para o modelo logístico dicotômico, politômico e de odds proporcionais, respectivamente. O modelo de Anderson é usado como apoio ao processo de seleção de variáveis independentes no caso de variáveis respostas politômicas ordinais.

Tabela 4.2 : Descrição das variáveis incluídas no processo de seleção .

Variável	Descrição	Parametrização
T_SOLO	Categórica	Slo1 Sol2 1 = Latossolo 0 0 2 = N-Latossolo 1 0 3 = Ambos 0 1
FER	Contínua	1 = Baixa 2 = Media 3 = Alta
PROF	Contínua	1 = Raso 2 = Medio 3 = Profundo
PROVISOR	Categórica	PRV1 PRV2 PRV3 1 = Bom 0 0 0 2 = Excessivo 1 0 0 3 = Ruim 0 1 0 4 = Ntem 0 0 1
DEFINIT	Categórica	DEF1 DEF2 1 = Bom 0 0 2 = Ruim 1 0 3 = Excessivo 0 1
ADUB	Categórica	0 = Nao faz adubacao 1 = Faz adubacao
CPE	Categórica	0 = Nao 1 = Sim
DEN	Categórica	0 = Com falhas 1 = Sem falhas
PMO	Categórica	0 = Sim 1 = Nao
IDADE	Contínua	1, 2,..., 10 (anos).
LIM	Contínua	0, 1, 2, 3, 4 .
FRVIS	Contínua	1, 2, 3, 4.
FRASS2	Contínua	$(FRASS)^{1/2}$
ESTRATO	Contínua	1, 2,...,9

4.2 APLICAÇÃO DO MÉTODO DE HOSMER E LEMESHOW PARA SELEÇÃO DE VARIÁVEIS INDEPENDENTES NO MODELO LOGÍSTICO DICOTÔMICO .

Como no segundo capítulo, nesta secção a variável resposta será tratada como dicotômica, de maneira que Y=1 se PAR_TEC=ótimo ou bom , e Y=0 se PAR_TEC=regular ou péssimo.

Inicialmente ajustou-se o modelo completo , contendo 18 variáveis, pelo procedimento CATMOD do S.A.S., e em seguida , através dos parâmetros estimados $\hat{\underset{\sim}{\beta}}$ por M.V. , determinou-se o valor da pseudo variável dependente $\underset{\sim}{Z}$ e seu peso $\underset{\sim}{W}$, segundo **(2.10)**.

Os parâmetros estimados por M.V., $\hat{\beta}_j$, para o modelo completo, e seus respectivos erros padrão, estão apresentados na Tabela 4.3 juntamente à estatística de Wald e nível de significância para o teste da hipótese de que β_j seja igual à zero. A Tabela também mostra os parâmetros estimados por M.Q.P., $\tilde{\underset{\sim}{\beta}}$, do modelo $\underset{\sim}{Z}=\underset{\sim}{X}'\underset{\sim}{\beta}$.

Através do ajuste por M.Q.P. de todos os possíveis submodelos $\underset{\sim}{Z}=\underset{\sim}{X}'\underset{\sim}{\beta}_q$ (q=1,2,...,p), onde q é o número de variáveis independentes considerado, selecionou-se dois de cada tamanho q, que apresentaram menores coeficientes de Mallow , C(q), obtidos pelo procedimento REG do S.A.S . Ao todo, então, foram avaliados 34 modelos, dos quais dez estão dispostos na Tabela 4.4, com seus respectivos valores de C(q), λ, λ^* e nível de significância para o teste da hipótese de que os (p-q) termos excluídos do ajuste sejam simultaneamente iguais à zero.

Os valores λ^*, λ e C(q) foram obtidos por :

$$\lambda^* = SQE(q) - SQE(p) = SQE(q) - 1041,47 \quad ,$$

$$\lambda = -2\ln\{L(\hat{\underset{\sim}{\beta}})_q - L(\hat{\underset{\sim}{\beta}})_p\} = -2\ln\{L(\hat{\underset{\sim}{\beta}})_q\} - 952,45 \quad ,$$

$$C(q) = \frac{SQE(q)}{SQE(p)}\,(n-q-1) + 2(q+1) - n \quad ,$$

onde n=1100 e p=18.

Tabela 4.3 : Estimativas dos parâmetros para o modelo de regressão dicotômica ajustado por MV e MQP; estatística de Wald e nível de significância para $Ho: \beta_j = 0$.

VARIAV.	Estim. M.V.	E.Padrao M.V.	Estim. M.Q.P	E.Padrao M.Q.P	Estat. WALD	Nivel (%) Signif.
Const.	-9,547	0,982	-9,547	0,964	94,48	0,00
Sol1	0,186	0,239	0,186	0,235	0,60	43,70
Sol2	0,134	0,351	0,134	0,345	0,14	70,40
FER	0,544	0,176	0,544	0,173	9,54	0,20
PROF	1,059	0,232	1,059	0,227	20,89	0,00
Prv1	-1,492	0,287	-1,492	0,282	26,97	0,00
Prv2	-0,841	0,253	-0,841	0,248	11,04	0,01
Prv3	0,256	0,207	0,256	0,203	1,53	21,60
Def1	-0,786	0,195	-0,786	0,191	16,32	0,00
Def2	-0,494	0,285	-0,498	0,280	3,04	8,10
ADUB	0,668	0,261	0,668	0,256	6,58	1,00
CPE	0,930	0,277	0,930	0,271	11,54	0,10
DEN	1,637	0,171	1,637	0,168	91,49	0,00
PMO	0,166	0,181	0,161	0,178	0,80	37,30
IDADE	0,099	0,040	0,099	0,040	5,99	1,40
LIM	0,490	0,118	0,490	0,116	17,23	0,00
FRVIS	0,229	0,083	0,229	0,082	7,59	0,60
FRASS2	0,707	0,203	0,707	0,199	12,11	0,10
ESTRATO	0,036	0,004	0,036	0,004	0,97	32,40

Tabela 4.4 : Modelos selecionados por M.Q.P.; soma de quadrado de resíduos; estatística de Wald (λ^*) e de razão de verossimilhança (λ); nível de significância e coeficiente de Mallow (C(q)).

Variáveis incluídas no submodelo	SQE(q) λ^*	$-2\ln L(\hat{\beta})$ λ .	Nivel de sign. % (gl)	C(q)
Prv1-Prv3,CPE,DEN	1143,46 101,99	1079,52 127,07	0,00 (13)	98,78
Prv1-Prv3,CPE,DEN,LIM	1123,40 81,93	1048,65 96,20	0,00 (12)	80,09
Prv1-Prv3,CPE,DEN,LIM, FRASS2	1108,73 67,26	1030,06 77,61	0,00 (11)	66,86
Prv1-Prv3,DEF1,DEF2,CPE, DEN,LIM,FRASS2	1091,39 49,92	1008,86 56,41	0,00 (9)	52,86
FER,PROF,Prv1-Prv3, CPE,DEN,LIM,FRASS2	1083,48 42,01	997,36 44,91	0,00 (9)	44,65
FER,PROF,Prv1-Prv3,Def1 Def2,CPE,DEN,LIM,FRASS2	1065,84 24,37	978,71 26,26	0,00 (7)	30,34
FER,PROF,Prv1-Prv3,Def1, Def2,CPE,DEN,LIM,FRASS2, FRVIS	1058,04 16,57	969,99 17,54	1,00 (6)	24,14
FER,PROF,Prv1-Prv3,Def1, Def2,CPE,DEN,IDADE,LIM, FRASS2	1058,03 16,56	970,06 17,61	1,00 (6)	24,12
FER,PROF,Prv1-Prv3,Def1, Def2,ADUB,CPE,DEN,LIM, FRVIS,FRASS2	1050,69 9,22	961,95 9,50	5,00 (5)	18,62
FER,PROF,Prv1-Prv3,Def1, Def2,ADUB,CPE,DEN,IDADE, LIM,FRVIS,FRASS2	1044,76 3,29	954,98 2,53	25,00 (4)	14,46

Os resultados mostram que as estatísticas de Wald , λ^*, e de razão de verossimilhança , λ, apresentam valores aproximados, como o esperado, sendo os valores da primeira sempre menores, demonstrando ser mais conservadora. O valor de C(q) decresce com λ^*, se aproximando de (q+1) à medida que SQE(q) se aproxima de SQE(p).

Note que em todos os modelos da Tabela 4.4 os grupos de variáveis (PRV1 PRV2 PRV3) e (DEF1 DEF2) aparecem completos. Isso não ocorreu necessariamente no procedimento de seleção, mas se uma das variáveis dicotômicas foi incluída no ajuste, as restantes do mesmo grupo foram da mesma forma consideradas.

Neste exemplo verifica-se que não existe um número pequeno de fatores determinantes na resposta das implantações do período, mas sim um conjunto de procedimentos como qualidade de sombreamento, controle de pragas, etc, associados à qualidade do solo. O tamanho da área implantada (ESTRATO), não se apresenta como fator relevante ao sucesso ou fracasso das novas áreas. As variáveis independentes PROVISOR, CPE, e DEN, foram incluidas em todos os melhores modelos de tamanho q>5, sugerindo serem os fatores que apresentam maior associação com o logito da probabilidade de sucesso.

Na Tabela 4.5 estão apresentadas estimativas de M.V. para os sete últimos modelos selecionados da tabela anterior.

Em geral os valores estimados dos parâmetros não sofreram grandes modificações com a inclusão dos fatores IDADE, ADUB e FRVIS no modelo. A maior variação relativa ocorre na variável Prv3, que representa a categoria *ntem* para o fator PROVISOR (sombreamento provisório) , o que vem reforçar a hipótese levantada na secção 2.3 de que o efeito desta categoria se confunde com a variável IDADE.

Para comparações entre modelos hierárquicos, por exemplo o quarto e o quinto modelos da Tabela 4.4, utiliza-se o teste de razão de verossimilhança G definido por:

$$G = L_5 - L_6$$

onde L_j é o valor de $-2\ln L(\hat{\beta})_j$ do j-ésimo modelo .

No exemplo, G=18,75 tem distribuição assintótica χ^2 com 2 graus

de liberdade, rejeitando a hipótese de que os parâmetros referentes às variáveis Def1 e Def2 sejam iguais à zero, ao nível α=0,001.

Tabela 4.5 : Ajuste dos modelos selecionados.

VARIAVEL	Modelo (4)	Modelo (5)	Modelo (6)	Modelo (7)	Modelo (8)	Modelo (9)	Modelo (10)
FER		0,607	0,536	0,547	0,599	0,647	0,651
PROF		0,933	1,010	1,005	1,023	1,016	1,008
Prv1	-1,352	-1,420	-1,490	-1,547	-1,377	-1,390	-1,440
Prv2	-0,895	-1,071	-0,860	-0,886	-0,840	-0,834	-0,857
Prv3	0,409	0,313	0,422	0,292	0,392	0,379	0,267
Def1	-0,841		-0,822	-0,832	-0,832	-0,819	-0,826
Def2	-0,557		-0,630	-0,534	-0,587	-0,597	-0,508
ADUB						0,714	0,688
CPE	1,091	1,171	1,108	1,053	1,174	0,980	0,932
DEN	1,761	1,744	1,679	1,674	1,653	1,638	1,635
IDADE				0,113			0,104
LIM	0,505	0,518	0,515	0,523	0,501	0,483	0,491
FRVIS					0,232	0,231	0,220
FRASS2	0,847	0,799	0,845	0,807	0,791	0,776	0,742

O modelo (10) coincide com o modelo selecionado pelo procedimento Stepwise, efetuado no SAS, onde os níveis de significância para entrada e saída das variáveis foram α_e=0,15 e α_s=0,20, respectivamente.

O processo de seleção, dada a estrutura dos dados, não atingiu o objetivo de se encontrar um modelo com número reduzido de parâmetros, entretanto auxilia na identificação dos fatores que estão de alguma maneira associados ao parecer técnico dado pelos entrevistadores. Outras análises mais detalhadas devem ser efetuadas, considerando o comportamento das variáveis independentes por época de plantio e região.

4.3 APLICAÇÃO DO MÉTODO DE HOSMER E LEMESHOW PARA O MODELO LOGÍSTICO POLITÔMICO, E UTILIZAÇÃO DO MODELO UNIDIMENSIONAL DE ANDERSON PARA SELEÇÃO DE VARIÁVEIS INDEPENDENTES EM MODELOS ORDENADOS.

Para exemplificação do modelo politômico serão consideradas as quatro categorias da variável resposta PAR_TEC como descrita na Tabela 4.2, sendo Y=4 (PAR_TEC = Péssimo) a categoria de referência.

Como primeira etapa do processo de seleção de variáveis independentes, ajustou-se o modelo logístico politômico **(3.1)** , e os modelos logísticos dicotômicos individuais **(3.3)**, considerando as 18 variáveis propostas inicialmente. Os resultados dispostos na Tabela 4.6 mostram que, principalmente para os parâmetros que pela estatística de Wald não foram significantes para o ajuste, as estimativas das regressões individuais não se aproximam daquelas obtidas pelo ajuste do modelo politômico, sendo maior a diferença quanto maior for a distância da categoria em relação à categoria de referência. O mesmo não foi observado no exemplo 3.1, sugerindo que esta instabilidade seja consequência da quantidade de parâmetros no modelo, associada à baixa frequência de Y=1, n_1=54 (Tabela 3.1).

Em seguida através da aplicação do método de seleção de variáveis independentes de Hosmer e Lemeshow, ao modelo politômico e às regressões individuais, selecionou-se o menor submodelo cujo nível de significância, para o teste de Wald λ^*, foi maior ou igual à 10%. (Tabela 4.7).

Tabela 4.6 : Estimativa do modelo politômico , dos modelos dicotômicos individuais e razões r_2 e r_3 ,onde $r_d=\beta_d/\beta_1$.
(Obs : Para as estimativas com sinal (') o teste de Wald rejeita Ho:$\beta_{ij}=0$ ao nível α=5%).

VARIAV.	REG.POLITOMICA $\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	REG.INDIVIDUAIS $\hat{\beta}_1^*$	$\hat{\beta}_2^*$	$\hat{\beta}_3^*$	r_2	r_3
Const.	-20,293'	-11,120'	-3,398'	-29,826'	-12,244'	-3,286'		
Sol1	0,849	0,103	-0,015	-0,093	-0,370	0,031	0,121	-0,018
Sol2	-1,022	0,162	-0,040	-1,925	0,186	-0,104	-0,158	0,039
FER	1,435'	0,931'	0,563'	2,714'	1,090'	0,547'	0,648	0,392
PROF	2,074'	1,408'	0,527'	4,253'	1,684'	0,532'	0,679	0,254
Prv1	-2,658'	-1,962'	-0,660'	-4,797'	-1,770'	-0,698'	0,738	0,248
Prv2	-3,008'	-1,727'	-1,338'	-6,337'	-1,520'	-1,338'	0,574	0,445
Prv3	0,369	0,193	-0,057	-1,533	0,431	-0,047	0,524	-0,154
Def1	-0,599	-0,891'	-0,102	-1,324	-0,692'	-0,166	1,488	0,171
Def2	-1,438	-0,692	-0,306	-3,971'	-0,525	-0,379	0,481	0,212
ADUB	0,929	0,989'	0,423'	1,136	1,101'	0,451'	1,064	0,455
CPE	1,748	1,395'	0,678'	0,926	1,558'	0,633'	0,798	0,388
DEN	3,805'	2,272'	0,961'	6,371'	2,606'	0,956'	0,597	0,253
PMO	0,332	-0,228	-0,450'	-0,632	-0,395	-0,453'	-0,687	-1,353
IDADE	0,243'	0,155'	0,081	0,580'	0,202'	0,079	0,636	0,334
LIM	1,432'	0,876'	0,548'	2,775'	0,941'	0,531'	0,611	0,382
FRVIS	0,327	0,329'	0,131	0,194	0,420'	0,104	1,006	0,401
FRASS2	0,738	0,620'	-0,108	0,123	-0,048	-0,057	0,840	-0,147
ESTRATO	0,059	-0,003	-0,042	-0,113	0,047	-0,043	-0,057	-0,707

Tabela 4.7 : Seleção de variáveis independentes para as regressões individuais, e o modelo politômico.

Método	Função	Variáveis (gl)	λ^* (α %)	λ (α %)	C(q)
Individual	$\eta_1=\ln(\Pi_1/\Pi_4)$	FER, PROF, Prv1-Prv3, DEN, LIM (12)	11,00 (25,0)	18,26 (5,0)	35,40
	$\eta_2=\ln(\Pi_2/\Pi_4)$	FER, PROF, Prv1-Prv3, ADUB, CPE, DEN, IDADE, LIM, FRVIS (7)	7,99 (25,0)	8,27 (25,0)	12,18
	$\eta_3=\ln(\Pi_3/\Pi_4)$	FER, PROF, Prv1-Prv3, CPE, DEN, PMO, LIM (9)	12,57 (10,0)	14,26 (10,0)	11,60
Politômico	η_1	FER, PROF, Prv1-Prv3, CPE, DEN, IDADE, LIM	λ^* = 8,2	λ^{**} =34,7	44,20
	η_2	FER, PROF, Prv1-Prv3, Def1-Def2, ADUB, CPE DEN, IDADE, LIM, FRVIS FRASS2	9 gl α=25,0	20 gl 2,5<α<5	
	η_3	FER, PROF, Prv1-Prv3, ADUB, CPE, DEN, PMO, IDADE, LIM (20)			

Observa-se na tabela acima, que quando a seleção é feita nas regressões individuais, a diferença entre λ e λ^* no primeiro logito (η_1) é bem maior que nos logitos η_2 e η_3. Dado o número pequeno de observações para Y=(1), este resultado é esperado, pois a aproximação da distribuição das duas estatísticas para a distribuição χ^2 com (p-q) graus de liberdade, é assintótica. Para o modelo politômico, λ^* foi calculado considerando o mesmo grupo de variáveis para os três logitos, e λ^{**} foi obtida restringindo-se às variáveis selecionadas, que são basicamente as mesmas dos modelos individuais, diferenciando-se apenas nas inclusões de DEFINIT, FRASS2 e PMO.

O ajuste e seleção de variáveis independentes para o modelo politômico foi efetuado a título de exemplificação do material apresentado na secção 3.2 . A estrutura ordenada da variável resposta indica que modelos que exploram essa ordenação, como o modelo de odds proporcionais e unidimensional de Anderson, são mais adequados além de conterem um menor número de parâmetros a ser estimados.

À seguir, baseado na estrutura do modelo unidimensional de Anderson, e através das estimativas obtidas na tabela 4.6, foi feito um estudo descritivo das razões r_2 e r_3, buscando identificar os fatores que possam contribuir para o ajuste de modelos ordenados.

Considerando as estimativas $\hat{\Phi}_1=1$, $\hat{\Phi}_4=0$, $\hat{\beta} = \hat{\beta}_1$, $\hat{\Phi}_2= 0,62$ e $\hat{\Phi}_3=0,25$, onde 0,62 e 0,25 são as medianas de r_2 e r_3, respectivamente, verifica-se que as variáveis Sol1-Sol2, PMO, Prv3, Def1, FRASS2 e ESTRATO não evidenciam uma relação ordenada, nem a identificação entre as funções logito, pois seus respectivos valores de r_2 e r_3 não se aproximam de $\hat{\Phi}_2$ e $\hat{\Phi}_3$ (Figura 1). A Tabela 4.8 decreve novas estimativas do modelo politômico onde, com exceção de Def1 e Prv3, essas variáveis são excluídas.

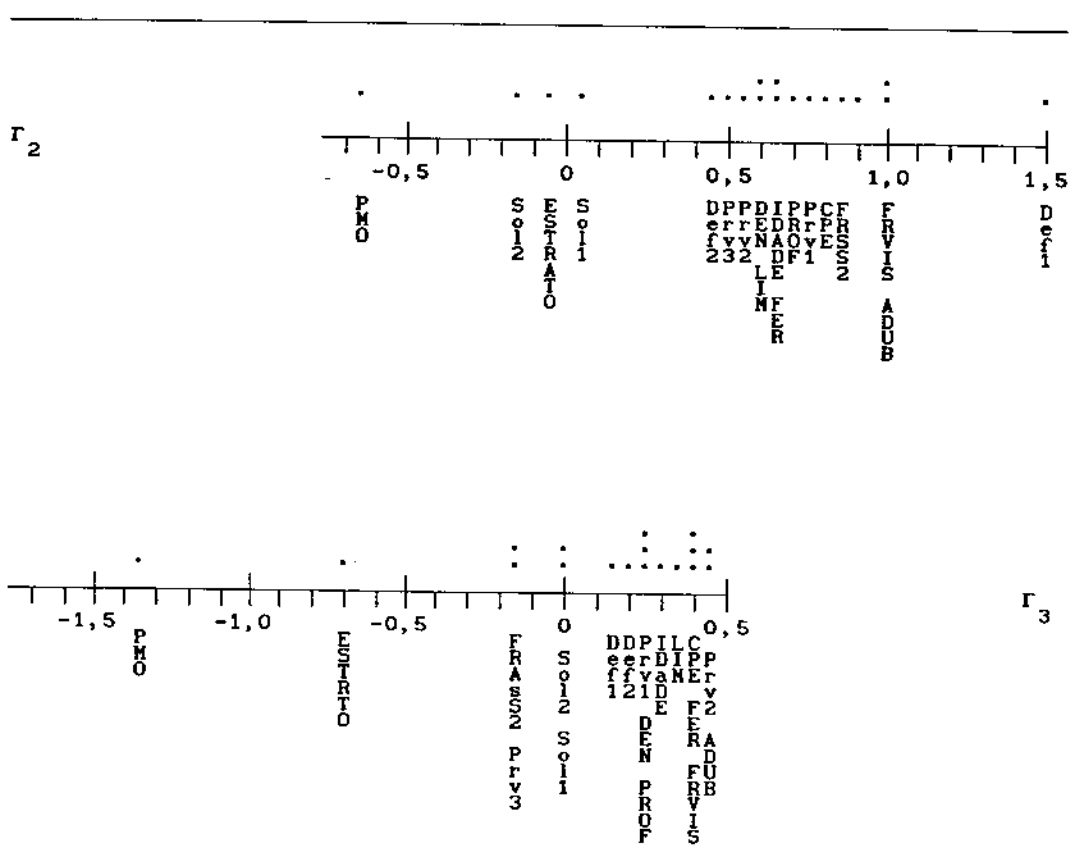

Figura 1: Representação ordenada das variáveis segundo r_2 e r_3.

Tabela 4.8 : Parâmetros estimados do submodelo e análise descritiva de r_2 e r_3 .

VAR	$\hat{\beta}_1$	$\hat{\beta}_2$	β_3	r_2	r_3
Const.	-18,244	-10,698	-3,894	0,576	0,213
FER	1,822	0,991	0,516	0,544	0,283
PROF	1,835	1,416	0,563	0,771	0,307
Prv1	-2,484	-2,025	-0,694	0,815	0,279
Prv2	-2,945	-1,680	-1,303	0,570	0,442
Prv3	0,314	0,128	-0,091	0,407	-0,291
Def1	-0,703	-0,840	-0,044	1,195	0,063
Def2	-1,444	-0,576	-0,245	0,399	0,170
ADUB	1,064	0,981	0,367	0,977	0,366
CPE	1,759	1,546	0,627	0,879	0,356
DEN	3,727	2,228	0,953	0,598	0,256
IDADE	0,246	0,170	0,080	0,686	0,323
LIM	1,457	0,911	0,512	0,625	0,351
FRVIS	0,307	0,356	0,131	1,158	0,428
		3^{o} Quartil		0,879	0,356
		Mediana		0,686	0,307
		1^{o} Quartil		0,570	0,256

Usando as medianas observadas, estimou-se $\hat{\Phi}_2$=0,69 e $\hat{\Phi}_3$=0,31, obtendo uma aproximação do ajuste do modelo unidimensional de Anderson onde a suposição de paralelismo é confirmada, pois $\hat{\Phi}_1 \geq \hat{\Phi}_2 \geq \hat{\Phi}_3 \geq \hat{\Phi}_4$. Dos fatores incluídos, Def1, Prv1, ADUB, CPE, e FRVIS ($\bar{r}_2$=1,001) sugerem não haver diferenças entre a primeira e a segunda função estimada, isto é, as cinco variáveis não distinguem as categorias *ótimo e bom*. Considerando esses resultados e a baixa frequência da primeira categoria de Y, os mesmos procedimentos foram repetidos para o modelo tricotômico, onde as categorias (1) e (2) de Y foram combinadas, tal que:

$$Y = \begin{cases} 1 & \text{se} \quad \text{PAR\_TEC} = \text{Ótimo ou } \textit{bom} \ , \\ 2 & \text{se} \quad \text{PAR\_TEC} = \textit{Regular} \ , \\ 3 & \text{se} \quad \text{PAR\_TEC} = \textit{Péssimo} \ . \end{cases}$$

Os resultados estão dispostos nas tabelas 4.9, 4.10, e 4.11 .

Tabela 4.9 : Ajuste do modelo logístico tricotômico; dos modelos dicotômicos individuais; estatística de Wald para Ho: $\beta_{1j}=0$ (i=1,2; j=0,...,p) ; razões dos parâmetros β_{2j} e β_{1j} (r_2).

VAR	REG. POLIT. $\hat{\beta}_1$	REG. POLIT. $\hat{\beta}_2$	REG. INDIV. $\hat{\beta}^*_1$	REG. INDIV. $\hat{\beta}^*_2$	Estat. de WALD (') α=5%	Estat. de WALD ('') α=1%	r_2	r^*_2
Const.	-11,547	- 3,360	-12,423	- 3,286	86,08''	13,01''		
Sol1	0,167	- 0,021	- 0,368	0,031	0,28	0,01	-0,126	-0,084
Sol2	0,107	- 0,034	0,136	- 0,104	0,05	0,01	-0,313	-0,764
FER	0,981	0,557	1,126	0,547	16,07''	7,11''	0,568	0,486
PROF	1,467	0,522	1,726	0,532	27,72''	6,98''	0,356	0,308
Prv1	- 2,021	- 0,657	- 1,881	- 0,698	29,98''	5,15''	0,325	0,371
Prv2	- 1,798	- 1,333	- 1,600	- 1,338	31,61''	28,24''	0,742	0,836
Prv3	0,198	- 0,056	0,330	- 0,047	0,42	0,04	-0,284	-0,143
Def1	- 0,863	- 0,107	- 0,757	- 0,166	9,08''	0,17	0,124	0,219
Def2	0,738	- 0,303	- 0,591	- 0,379	3,95''	0,99	0,411	0,642
ADUB	0,981	0,426	1,091	0,451	10,23''	3,98'	0,434	0,414
CPE	1,417	0,678	1,529	0,633	19,79''	10,11''	0,478	0,414
DEN	2,392	0,950	2,700	0,956	75,02''	13,57''	0,397	0,354
PMO	- 0,197	- 0,453	- 0,359	- 0,453	0,72	6,08'	2,296	1,262
IDADE	0,162	0,081	0,218	0,080	9,34''	3,45	0,499	0,362
LIM	0,924	0,543	0,966	0,531	31,60''	15,35'	0,587	0,550
FRVIS	0,332	0,131	0,424	0,104	10,24''	2,70	0,395	0,245
FRASS2	0,621	- 0,107	- 0,045	- 0,057	6,03''	0,31	-0,172	1,271
ESTRATO	0,002	- 0,042	0,047	- 0,043	0,00	0,98	-22,15	-0,907

Tabela 4.10: Seleção de variáveis independentes para as regressões individuais, e o modelo tricotômico.

Método	Função	Variáveis (gl)	λ^* (α %)	λ (α %)	C(q)
Indiv.	$\eta_1=\ln(\Pi_1/\Pi_3)$	FER, PROF, Prv1-Prv3, ADUB, CPE, DEN, IDADE, LIM, FRVIS (7)	8,10 (25.0)	8,65 (25,0)	12,64
	$\eta_2=\ln(\Pi_2/\Pi_3)$	FER, PROF, Prv1-Prv3, PMO , CPE, DEN, LIM (9)	12,50 (10,0)	14,25 (10,0)	11,55
Tricot.	η_1	FER, PROF, Prv1-Prv3, Def1, Def2, ADUB, CPE DEN, IDADE, LIM, FRVIS	λ^*=2,37 6 gl	λ^{**}=20,8 11 gl	24,76
	η_2	FER, PROF, Prv1-Prv3, ADUB, CPE, DEN, PMO, IDADE, LIM, FRASS2 (11)	α=25,0	α=5,0	

O método de Hosmer e Lemeshow aplicado ao modelo tricotômico e às regressões individuais, apresenta resultados semelhantes no que diz respeito aos subconjuntos de variáveis independentes selecionadas. Todas as variáveis selecionadas pelas regressões individuais foram também incluídas no modelo tricotômico.

Na Tabela 4.11 estão dispostos os resultados do ajuste do modelo tricotômico contendo as variáveis selecionadas, para o modelo de Anderson. Pela Figura 2 observa-se que as variáveis que apresentam valores de r_2 que mais se distanciam da mediana $\hat{\Phi}_2$=0,40 são : Estrato, Sol1, Sol2, Prv3, FRASS2, e PMO.

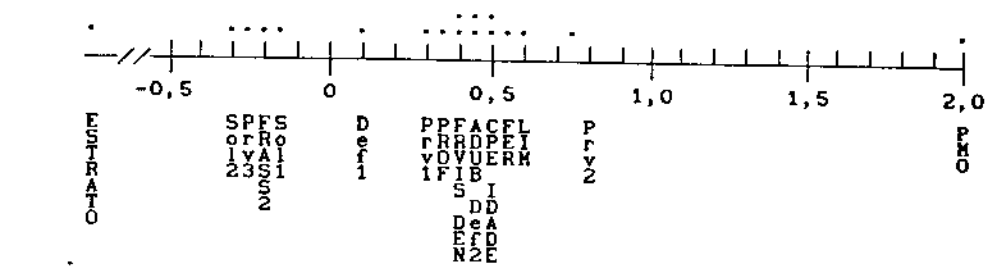

Figura 2: Representação ordenada das variáveis segundo valores de r_2.

Tabela 4.11: Parâmetros estimados do submodelo e análise descritiva de r_2.

VAR.	$\hat{\beta}_1$	$\hat{\beta}_2$	r_2
Const.	-11.036	-3,862	0,350
FER	1,066	0,509	0,477
PROF	1,455	0,561	0,385
Prv1	-2,062	-0,692	0,335
Prv2	-1,748	-1,298	0,743
Prv3	0,137	-0,091	-0,670
Def1	-0,826	-0,047	0,057
Def2	-0,630	-0,242	0,384
ADUB	0,923	0,369	0,375
CPE	1,559	0,628	0,403
DEN	2,349	0,941	0,401
IDADE	0,177	0,079	0,449
LIM	0,958	0,507	0,529
FRVIS	0,354	0,132	0,372
		3º Quartil	0,448
		Mediana	0,385
		1º Quartil	0,372

Ao contrário do modelo politômico (Tabela 4.8), os valores de r_2 obtidos na tabela acima, com exceção de r_2=-0,670 e r_2=0,057 referentes às variáveis Prv3 e Def1, são próximos do mediana (0,385), e pertencem ao intervalo aberto (0,1), o que vem reforçar a suposição de relação ordenada entre $\underset{\sim}{X}$ e Y, e de identificação das funções logito estimadas.

4.4 APLICAÇÃO DO MÉTODO DE HOSMER E LEMESHOW PARA SELEÇÃO DE VARIÁVEIS INDEPENDENTES NO MODELO DE ODDS PROPORCIONAIS .

A seleção de variáveis independentes para o modelo de odds proporcionais foi efetuada para k=4, e k=3 categorias da variável resposta, respectivamente.

Como na secção 4.2, partiu-se do ajuste do modelo completo, efetuado pelo procedimento LOGISTIC, do S.A.S., com resultados apresentados na tabela 4.12.

As estimativas obtidas são próximas dos valores estimados para os parâmetros do modelo dicotômico, o que já era esperado, dado que os coeficientes do modelo de odds proporcionais são invariantes com relação à combinação de categorias adjacentes.

Através do ajuste por MQP de todos os possíveis submodelos aplicados à pseudo variável dependente $\underset{\sim}{Z}$ com peso $\underset{\sim}{W}$, dado pela expressão **(3.6)** , selecionou-se aqueles com menores valores de C(q), dispostos na Tabela 4.13 .

Tabela 4.12 : Parâmetros estimados do modelo de odds proporcionais, estatística de Wald e nível de significância para Ho: β_j=0.

VAR	M.V.		M.Q.P.		Estat.	Nivel
	Estim.	Erro Padrao	Estim.	Erro Padrao	WALD	signif. (%)
const1	-11,363	0,734	-11,228	0,610	239,449	0,01
Const2	- 8,273	0,692	- 8,218	0,547	142,850	0,01
Const3	- 5,615	0,664	- 5,564	0,507	71,569	0,01
Sol1	0,226	0,176	0,225	0,132	1,644	19,20
Sol2	- 0,007	0,274	- 0,007	0,205	0,007	97,88
FER	0,615	0,136	0,609	0,102	20,609	0,01
PROF	0,881	0,156	0,874	0,118	31,910	0,01
Prv1	- 1,190	0,207	- 1,180	0,157	33,055	0,01
Prv2	- 1,176	0,182	- 1,170	0,139	41,523	0,01
Prv3	0,190	0,164	0,189	0,122	1,352	24,49
Def1	- 0,522	0,155	- 0,515	0,116	11,343	0,08
Def2	- 0,446	0,212	- 0,440	0,158	4,437	3,52
ADUB	0,512	0,172	0,510	0,129	8,842	0,29
CPE	0,803	0,177	0,801	0,133	20,643	0,01
DEN	1,570	0,146	1,554	0,115	115,393	0,01
PMO	- 0,061	0,132	- 0,061	0,099	0,213	64,46
IDADE	0,094	0,030	0,093	0,023	9,860	0,17
LIM	0,562	0,090	0,556	0,068	38,897	0,01
FRVIS	0,197	0,060	0,195	0,045	10,907	0,10
FRASS2	0,389	0,143	0,386	0,107	7,454	0,63
ESTRATO	0,016	0,028	0,016	0,021	0,310	57,78

Tabela 4.13 : Modelos selecionados por M.Q.P.; soma de quadrado de resíduos; estatística de Wald e de razão de verossimilhança; nível de significância , e coeficiente Mallow.

Variáveis incluídas no submodelo	SQE(Q) λ^*	$-2\ln L(\hat{\beta})$ λ	Nível de sign. (%) (gl)	C(q)
FER,PROF,Prv1-Prv3,CPE DEN,LIM,	1885,06 53,07	2131,88 60,74	0,00 (10)	95,99
FER,PROF,Prv1-Prv3,CPE, DEN,LIM,FRVIS	1874,24 42,25	2118,74 47,60	0,00 (9)	78,89
FER,PROF,Prv1-Prv3, CPE,DEN,IDADE,LIM,FRVIS	1862,60 31,61	2105,39 34,25	0,00 (8)	60,05
FER,PROF,Prv1-Prv3,ADUB CPE,DEN,IDADE,LIM,FRVIS	1853,28 21,29	2095,24 24,10	1,00 (7)	45,37
FER,PROF,Prv1-Prv3,Def1, Def2,ADUB,CPE,DEN,IDADE, LIM,FRVIS	1841,71 9,72	2082,84 11,70	5,00 (5)	28,75
FER,PROF,Prv1-Prv3,Def1, Def2,ADUB,CPE,DEN,IDADE, LIM,FRVIS,FRASS2	1834,12 2,13	2074,71 3,54	25,00 (4)	17,07

Os fatores PROVISOR, CPE e DEN aparecem, como na Tabela 4.4, em todos os modelos selecionados de tamanho maior ou igual à cinco, já os fatores DEFINIT e FRASS2 só foram incluídos naqueles com mais de q=11 e q=13 variáveis, respectivamente.

Na Tabela 4.14 são apresentadas estimativas para os parâmetros dos modelos selecionados. Através de comparações entre os modelos hierárquicos , pela estatística de razão de verossimilhança, conclui-se

que todos os parâmetros do modelo (6) são diferentes de zero (G_6-G_5=8,13 , 1 gl).

Tabela 4.14 : Modelos ajustados por MV .

Variavel	Modelo (1)	Modelo (2)	Modelo (3)	Modelo (4)	Modelo (5)	Modelo (6)
FER	0,643	0,710	0,729	0,768	0,712	0,707
PROF	0,831	0,833	0,830	0,826	0,889	0,837
Prv1	-1,203	-1,093	-1,120	-1,129	-1,171	-1,134
Prv2	-1,247	-1,237	-1,269	-1,275	-1,534	-1,169
Prv3	0,244	0,220	0,115	0,105	0,178	0,193
Def1					-0,527	-0,546
Def2					-0,441	-0,461
ADUB				0,541	0,522	0,506
CPE	1,126	1,159	1,094	0,918	0,885	0,788
DEN	1,658	1,637	1,619	1,603	1,543	1,558
IDADE			0,107	0,103	0,103	0,100
LIM	0,630	0,610	0,622	0,589	0,592	0,560
FRVIS		0,207	0,202	0,200	0,203	0,187
FRASS2						0,398

O modelo selecionado no procedimento Stepwise, se equivale ao modelo (6) acima.

Como para o caso dicotômico, as estimativas são estáveis com relação à inclusão ou exclusão de novos fatores no modelo.

Os fatores selecionados na secção anterior , Tabela 4.8, são os mesmos incluídos no modelo (5) acima, mostrando coerência nos dois métodos de seleção. Por outro lado, a variável FRASS2, presente no sexto

modelo, apresenta valores de r_2 e r_3, (0,84 e -0,15), próximos de 1 e 0, respectivamente (Tabela 4.6), sugerindo paralelismo entre as funções logitos, mas não identifica as categorias (1) e (2), e (3) e (4) de Y. Para o modelo dicotômico, FRASS2 aparece em todos os melhores modelos, segundo valor d C(q), de tamanho q≥7.

Abaixo estão dispostos os resultados quando considerou-se a combinação das duas primeiras categorias de Y.

Tabela 4.15 : Ajuste do modelo de odds proporcionais para k=3 categorias.

VAR.	Estim. M.V.	E.Padrao M.V.	Estim. M.Q.P.	E.Padrao M.Q.P.	Estat. WALD	Nivel Signif.
Const1	-8,023	0,712	-8,001	0,625	126,791	0,000
Const2	-5,337	0,684	-5,362	0,579	61,665	0,000
Sol1	0,150	0,186	0,150	0,155	0,654	0,419
Sol2	0,102	0,287	0,102	0,239	0,127	0,722
FER	0,587	0,143	0,585	0,120	16,899	0,000
PROF	0,853	0,159	0,851	0,134	28,704	0,000
Prv1	-1,214	0,212	-1,210	0,179	32,914	0,000
Prv2	-1,179	0,186	-1,176	0,158	40,096	0,000
Prv3	0,147	0,174	0,147	0,145	0,718	0,397
Def1	-0,646	0,165	-0,644	0,139	15,287	0,000
Def2	-0,462	0,218	-0,461	0,183	4,484	0,034
ADUB	0,536	0,145	0,535	0,147	9,392	0,002
CPE	0,783	0,178	0,781	0,149	19,406	0,000
DEN	1,509	0,150	1,505	0,130	101,138	0,000
PMO	-0,089	0,137	-0,089	0,114	0,422	0,516
IDADE	0,097	0,031	0,096	0,026	9,437	0,002
LIM	0,539	0,095	0,538	0,080	32,256	0,000
FRVIS	0,201	0,061	0,201	0,051	10,796	0,001
FRASS2	0,439	0,148	0,438	0,124	8,845	0,003
ESTRATO	0,003	0,030	0,003	0,025	0,008	0,927

Tabela 4.16 : Seleção de submodelos segundo método de Hosmer e Lemeshow para o modelo de odds proporcionais com k=3 categorias.

Variáveis incluídas no submodelo	SQE(q) λ^*	$-2\ln L(\hat{\beta}_q)$ λ	Nivel de sign. (%) (gl)	C(q)
FER,PROF,Prv1-Prv3,CPE DEN,LIM,	1577,30 55,98	1855,01 63,00	0,00 (10)	76,27
FER,PROF,Prv1-Prv3,CPE, DEN,LIM,FRVIS	1565,77 44,45	1841,06 49,05	0,00 (9)	61,75
FER,PROF,Prv1-Prv3,ADUB, CPE,DEN,LIM,FRVIS	1555,87 34,55	1830,24 38,23	0,00 (8)	49,56
FER,PROF,Prv1-Prv3,CPE, DEN,IDADE,LIM,FRVIS	1555,31 33,99	1829,07 37,06	0,00 (8)	48,76
FER,PROF,Prv1-Prv3,Def1, Def2,CPE,DEN,IDADE,LIM, FRVIS	1540,52 19,20	1813,14 21,13	1,00 (6)	31,56
FER,PROF,Prv1-Prv3,Def1, Def2,ADUB,CPE,DEN,LIM, FRVIS,FRASS2	1532,53 11,21	1804,51 12,50	5,00 (5)	21,12
FER,PROF,Prv1-Prv3,Def1, Def2,ADUB,CPE,DEN,IDADE, LIM,FRVIS	1531,37 10,05	1803,54 11,53	10,00 (5)	20,40
FER,PROF,Prv1-Prv3,Def, Def2,ADUB,CPE,DEN,IDADE, LIM,FRVIS,FRASS2	1522,34 1,20	1793,97 1,96	25,00 (4)	9,51

Tabela 4.17: Ajuste dos modelos selecionados.

Variavel	Modelo (4)	Modelo (5)	Modelo (7)	Modelo (8)
FER	0,704	0,611	0,652	0,645
PROF	0,816	0,838	0,830	0,827
Prv1	-1,113	-1,205	-1,214	-1,183
Prv2	-1,267	-1,146	-1,154	-1,175
Prv3	0,165	0,139	0,127	0,145
Def1		-0,640	-0,624	-0,654
Def2		-0,468	-0,453	-0,471
ADUB	0,580		0,537	0,523
CPE	0,972	1,060	0,885	0,776
DEN	1,561	1,499	1,486	1,498
IDADE		0,107	0,103	0,100
LIM	0,547	0,601	0,567	0,532
FRVIS	0,257	0,217	0,214	0,198
FRASS2				0,446

As variáveis independentes incluídas no sétimo modelo da tabela acima coincidem com aquelas selecionadas segundo os valores de r (Tabela 4.12), e o modelo (8) foi também selecionado pelo procedimento Stepwise.

Observa-se então que, para o caso dos modelos dicotômico e de odds proporcionais, os resultados alcançados pela aplicação do método de Hosmer e Lemeshow, referentes à seleção de variáveis independentes segundo o coeficiente de Mallow e os testes de Wald, são os mesmos obtidos pelo procedimento Stepwise .

Com as análises elaboradas, têm-se um conjunto de informações à respeito do comportamento dos modelos avaliados, e de como os fatores

considerados contribuem para o ajuste em cada um deles.

Não é objetivo deste capítulo escolher um modelo para os dados analisados, mas exemplificar as metodologias apresentadas nos capítulos anteriores. A escolha de um determinado modelo, ou subconjunto de variáveis independentes, deve ser feita na presença de um profissional da área de interesse, considerando critérios como interpretação e coerência nas áreas envolvidas na pesquisa.

4.5 CONCLUSÕES

A estrutura do modelo unidimensional de Anderson avaliada pelo ajuste do modelo logístico politômico, apresentada na secção 4.3, mostra ser uma boa alternativa para seleção de variáveis independentes em modelos com variável resposta ordenada. As análises descritivas das razões dos parâmetros estimados, é neste exemplo, uma ferramenta bastante "útil e explicativa" no estudo das relações existentes entre a variável resposta o vetor de variáveis independentes.

A utilização das regressões individuais no processo de seleção de variáveis independentes do modelo politômico, forneceu resultados coerentes, apesar das diferenças observadas entre as estimativas $\hat{\underset{\sim}{\beta}}^{*}$ e $\hat{\underset{\sim}{\beta}}$.

A contribuição do método de Hosmer e Lemeshow é fornecer uma alternativa aos procedimentos Stepwise disponíveis atualmente somente nos grandes pacotes estatísticos, criando uma ferramenta adicional com facilidades computacionais. Com programas de regressão linear ponderada, o pesquisador pode, através de análises preliminares, obter informações que auxiliam no processo de seleção de subconjuntos de variáveis independentes para modelos logísticos.

Com o objetivo de extensão das técnicas apresentadas, outros modelos pertencentes à classe dos Modelos Lineares Generalizados podem ser pesquisados. Estudos relativos ao comportamento da estatística F no ajuste da pseudo variável dependente Z em X por MQP, pode possibilitar a utilização dos procedimentos STEPWISE de regressão linear, para seleção de variáveis independentes em modelos logísticos.

4.6 PACOTES ESTATÍSTICOS E O PROCESSO DE SELEÇÃO DE VARIÁVEIS INDEPENDENTES EM MODELOS LOGÍSTICOS .

Tabela 4.18: Programas disponíveis para seleção de variáveis independentes em regressão linear ponderada e modelos de regressão logística.

OBS : (*)- Procedimento incluído no pacote,

(o)- Procedimento programável através de subrotinas.

MODELO	PACOTE	MÉTODO ESTIMAÇÃO	SELEÇÃO DE V. INDEPEND.	
			STEPWISE	COMPARAÇAO ENTRE TODOS SUBMODELOS
REGRESSÃO LINEAR PONDERADA	BMDP	MQP	*	
	GLIM	MQPI	*	o
	SAS (REG-GLM)	MQP	*	*
	SPSS (2)	MQP	*	*
	SOC (1)	MQP	*	o
REGRESSÃO LOGISTICA DICOTÔMICA	BMDP (BMDP3R)		*	
	GLIM	MQPI	*	o
	SAS (CATMOD)	MV		
	SAS (LOGISTIC)	MQPI	*	
	SPSS			
	SOC		o	o
REGRESSÃO LOGISTICA POLITÔMICA	BMDP (BMDPAR) (BMDP3R)			
	GLIM	MQPI	o	o
	SAS (CATMOD)	MV		
	SPSS			
	SOC		o	o
MODELO DE ODDS PROPORC.	BMDP (BMDPAR) (BMDP3R)		*	
	GLIM	MQPI	*	o
	SAS (LOGISTIC)	MQPI	*	
	SPSS			
	SOC		o	o

(1) SOC - Software para PC, desenvolvido pelo DMQ e NTIA da EMBRAPA.
(2) SPSS- Statistical Package for the Social Science.

REFERÊNCIA BIBLIOGRÁFICA :

AGRESTI,A. (1984). "Analysis of Ordinal Categorical Data". J.Wiley & Sons, New York.

ANDERSON,J.A. (1984) . Regression and ordered categorical variables. **Journal of the Royal Statistical Society ,** série B, 46 : 1-30.

ARMSTRONG,B.G. & SLOAN,M. (1989). Ordinal regression models for epidemiologic data. **American Journal of Epidemilogi,** 129:191-204.

ASHBY,D. ; STUART,J.P. & SHAPER,G. (1986). Ordered polytomous regression: An exemple relating serum biochemistry and haematology to alcohol consumption. **APPLIED STATISTICS,** 35 : 289-301.

BEGG,C.B. & GRAY,R. (1984). Calculation of polychotomous logistic regression parameters using individualized regression. **Biometrika,** 71:11-18.

BISHOP,Y.M.M.; FIENBERG,S.E. & HOLLAND,P.W. (1975). "Discrete Multivariate Analysis: Theory and Practice ". Cambridge, The MIT Press.

BRESLOW,N.E. & DAY,N.E. (1980). "Statistical Methods in Cancer Research". Vol. 1 . Lyon, International Agency for Research on Cancer.

CORNFIELD,J. (1962). Join dependent of the risk of coronary heart diseases on serum cholesterol and systolic blood pressure: A discriminant function analysis.

COX,D.R (1970). "The Analysis of Binary Data". Chapman and Hall . Methuen, London.

DAY,N.E. & KERRIDGE,D.E. (1967). A general maximum lekehood discriminant. **Biometrics**, 23 : 313-323.

DIXON,W.J. (1987). "BMDP Statistical Software". University of California Press, Berkeley.

DRAPER,N.R. & SMITH,H. (1981). "Applied Regression Analysis" . 2ª edição. J.Willey & Sons, New York.

FURNIVAL,G.M. & WILSON,R.W.Jr. (1974). Regressions by leaps and bounds. **Technometrics** 16 : 499-511.

GILL,P.E. & MURRAY,W. (1972). Quasi-Newton methods for unconstrained optimization. **J.Inst.Math. e Applic.**, 9:91-108.

GREEN,J.P. (1984) . Iteratively Reweighted Least Square for maximum likelihood estimation, and some robust and resistant alternatives. **Journal of Royal Statistics Society** , série B 46 : 149-192.

GREENWOOD, C. & FAREWELL, V. (1988). A comparison of regression models for ordinal data in analysis of transplanted - kidney Function. **The Canadian Journal of Statistics**, 16 : 325-335.

GRIZZLE,J.; STAMER,F. & KOCH,G. (1969). Analysis of categorical data by linear models. **Biometrika**, 69 : 309-314.

HASTIE,T. & TIBSHIRANI,R. (1987). Nonparametric logistic and proportional odds regression. **Applied Statistics**, 36 : 260-276.

HAUCK,W.W.JR. & DONNER, A. (1977). Wald's Test as applied to hypotheses in logit analysis. **Journal of the American Statistical Association**, 72 : 851-853.

HOSMER,W.D.; JOVANOVIC,B. & LEMESHOW,S. (1989). Best subsets logistic regression. **Biometrics**, 45 : 1265-1270.

HOSMER,W.D. & LEMESHOW,S. (1989). "Applied Logistic Regression". J.Willy e Saus, New York.

HOTBRÜGGE,W. & SCHUMACHER,M. (1991). A comparison of regression models for the analysis of ordered categorical data. **Applied Statistics**, 40:249-259.

JENNINGS, D.E. (1986). Judging inference adequeacy in logistic regression. **Journal of the American Statistical Association**, 81:471-476.

KLEIMBAUM, D.G.; Kuper,L.L. & Morgenstern, H. (1982). "Epidemiologic Research Principles and Quaantitative Methods". Van Nostrand Reinhold, New York.

LAWLESS,J.F. & SINGHAL,K. (1978). Efficient screening of non normal regression models. **Biometrics**. 34:318-327.

LAWLESS,J.F. & SINGAL,K. (1987a). ISMOD: An all subsets regression program for generalized linear models I. Statistical and computational background. **Computer Methods and Programs in Biomedicine**, 24 : 117-124.

LAWLESS,J.F. & SINGAL,K. (1987b). ISMOD: An all subsets regression program for generalized linear models II. Program guide and examples. **Computer Methods and Programs in Biomedicine**, 24:125-134.

McCULLAGH,P. (1980). Regression models for ordinal data. **Journal of Royal Statistical Society**, série B, 42 : 109-142.

McCULLAGH,P. e NELDER (1989). "Generalized Linear Moldels". 2ª edição, London, Academic, Chapman and Hall.

Numerical Algorithms Group (1987). "The Generalized Linear Interactive Modelling System: The GLIM System". Versão 3.77. Royal Statistical Society, London.

PETERSON,B. & HARRELL,F.E.Jr. (1990). Partial proportional odds models for ordinal response variables. **Applied Statistics**, 39:205-217

PREGIBON, D. (1981). Logistic regression diagnostics. **The Annals of Statistics**, 9 : 705-724.

SAS Institute Inc. (1988). SAS User's Guide, versão 6.03. SAS Institute Inc., Cary, NC.

THOMPSON,R. & BAKER,R.J. (1981). Composite linear models. **Applied Statistics**, 30 : 125-131.

TRUETT,J. ; CORNFIEL,J. & KANEL,W. (1967). A multivariate analysis of the risk of coronarey heart disease in Framingham. **Journal of Chronic Diseases**, 20 : 511-524.

WALKER,S.H. & DUNCAN,D.B. (1967). Estimation of the probability of an event as a function of several independent variables. **Biometrika**, 54:167-179.