ANÁLISE DE EXPERIMENTOS INDUSTRIAIS COM RESPOSTAS CATEGÓRICAS ORDENADAS: MÉTODO DE TAGUCHI E MODELO DE McCULLAGH

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pelo Sr. Álvaro Vigo e aprovada pela Comissão Julgadora.

Prof. Dr. Armando Mario Infante

Orientador

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para a obtenção do Título de MESTRE EM ESTATÍSTICA.

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E CIÊNCIA DA COMPUTAÇÃO - IMECC

ANÁLISE DE EXPERIMENTOS INDUSTRIAIS COM RESPOSTAS CATEGÓRICAS ORDENADAS: MÉTODO DE TAGUCHI E MODELO DE McCULLAGH

ÁLVARO VIGO

PROF. DR. ARMANDO MARIO INFANTE

ORIENTADOR

CAMPINAS - SÃO PAULO

1994

Aos meus pais.

"Statistics should be introduced to engineers as a means of catalyzing engineering and scientific reasoning by way of design and data analysis. . . . if taught on a wide enough scale, could markedly improve quality and productivity and our overall competitive position".

G. E. P. BOX (1988)

AGRADECIMENTOS

Ao Professor Armando Mario Infante, pelo incentivo, dedicação e grandes ensinamentos.

À querida Nádia, pelo companheirismo e carinho.

Aos amigos de Porto Alegre, pelo incentivo e pela torcida. Aos grandes amigos de Campinas, pelo apoio nas horas difíceis e pelas alegrias nas horas de lazer. Também à turma do futebol, pelos bons momentos de descontração.

Aos professores da UFRGS e da UNICAMP, pelo estímulo e grandes contribuições para minha formação estatística.

À Universidade Federal do Rio Grande do Sul, em especial ao Departamento de Estatística, por apoiar meu afastamento. À CAPES, pela concessão da bolsa PICD.

ABSTRACT

In several areas of Science and Quality of Technology, experiments are conducted to study the influence of several factors on characteristics (as "worn-out severity" or "weld quality"), that are recorded as ordered categorical variables.

In this thesis, some aspects of the analysis of experiments with ordinal response are studied as a tool of Quality of Technology. The traditional methods of statistical analysis (Chi-square test and non-parametric tests, for example) are not efficient in this situation, and more sophisticated methods are needed to extract and to interpret the information produced in these experiments. So, two alternative methods are studied in detail: the first is the Accumulation Analysis (AA) technique introduced by Taguchi (1987). The other method is based on the Proportional Odds model proposed by McCullagh (1980), which allows to estimate and to interpret the effects of the factors. A special characteristics of this response is that it might be associated to a continuous latent variable. Thus, the observed data may be seen as a categorization of this continuous variable not directly observable.

These techniques of analysis are illustrated for the single-factor and the multifactor cases, by using two data set presented in the literature: an observational study and an experimental study. The models were adjusted by the procedure LOGISTIC of the software SAS and by the software STATA.

The results obtained point to the need of parametric models to estimate the magnitude and direction of the effects caused by the factors in the characteristic of quality, since the traditional methods and the AA do not extract enough information of the data. In this sense, the Proportional Odds model seems a simple, adequate and efficient method for analyzing data with ordinal response.

RESUMO

Nas diversas áreas da ciência e na Tecnologia da Qualidade são realizados experimentos para estudar a influência de diversos fatores sobre características (tais como "severidade do desgaste" ou "qualidade de uma solda"), que são registradas como variáveis categóricas ordenadas.

Nessa dissertação são estudados aspectos da análise de experimentos com resposta ordinal, como ferramenta da Tecnologia da Qualidade. Os métodos tradicionais de análise estatística (teste Qui-quadrado e testes não-paramétricos, por exemplo) não são eficientes nessa situação, sendo necessários métodos mais elaborados para extrair e interpretar as informações geradas nesses experimentos. Diante disso, foram estudados em detalhe dois métodos alternativos: o primeiro é a técnica da Análise de Acumulação (AA) introduzida por Taguchi (1987) e, o outro método, baseia-se no modelo de Odds Proporcionais proposto por McCullagh (1980), através do qual é possível estimar e interpretar os efeitos dos fatores. Uma característica especial dessa resposta ordinal é a possibilidade de estar associada a uma variável contínua latente. Assim, os dados observados podem ser vistos como uma categorização dessa variável contínua, não diretamente observável.

Essas técnicas de análise são ilustradas para o caso unifatorial e multifatorial, mediante dois conjuntos de dados apresentados na literatura: um estudo observacional e outro experimental. O ajuste dos modelos foi realizado através do procedimento LOGISTIC do pacote estatístico SAS e pelo software STATA.

Com base nos resultados obtidos, verifica-se a necessidade de modelos paramétricos para estimar magnitude e direção dos efeitos que os fatores provocam na característica de qualidade, uma vez que os métodos tradicionais e a AA não extraem informação suficiente dos dados. Nesse sentido, o modelo de Odds Proporcionais parece um método simples, adequado e eficiente para a análise de dados com resposta ordinal.

ÍNDICE

CAPÍTULO 1: INTRODUÇÃO

1.1 QUALIDADE

Uma das idéias básicas de nossa época é a qualidade de produtos, processos e serviços. Ela é definida não somente como a ausência de defeitos ou a conformidade com as especificações, senão muito mais amplamente como "tudo o que alguém faz ao longo de um processo para garantir que um cliente, fora ou dentro da organização, obtenha exatamente aquilo que deseja - em termos de características intrínsecas, custo e atendimento", veja Lobos (1991, p.16). As características intrínsecas são os atributos desejados no produto ou serviço, que podem ser medidos pela ausência de defeitos. Assim, um produto ou serviço de qualidade é aquele que atende perfeitamente (projeto perfeito), de forma confiável (sem defeito), de forma acessível (baixo custo), de forma segura (segurança do cliente) e no tempo certo (entrega no prazo, local e quantidade certos) às necessidades do cliente, veja Campos (1992, p.2)

A qualidade de um produto ou de um processo pode ser avaliada e melhorada sistematicamente e este melhoramento é a base para o enorme crescimento da produtividade exibido pelas empresas japonesas.

A posição dos Estados Unidos como maior potência industrial do mundo tem sido ameaçada nos últimos anos. A seriedade desse problema pode ser constatada através da relação de produtos manufaturados norte-americanos que perderam a liderança no mercado internacional: automóveis, máquinas fotográficas, equipamentos médicos, televisão em cores, pneus radiais, motores elétricos, processadores de alimentos, forno de microondas, equipamentos esportivos, computadores, robótica, ferramentas manuais e mecânicas, etc. Na última década, a participação dessas indústrias norte-americanas no mercado mundial assistiu a um declínio de no mínimo 50%, em grande parte provocado pelo Japão, veja Wheelwright (1984) e Box & Bisgaard (1987).

A acentuada flexibilidade das empresas japonesas - por um lado - mostra a importância da gestão da qualidade. Ela inclui a prática do CQT - Controle da Qualidade Total (TQC - Total Quality Control, em inglês), que pode ser definido como "o controle exercido por

todas as pessoas para a satisfação das necessidades de todas as pessoas", veja Campos (1992, p.15). Aspectos relativos aos conceitos e implantação do TQC e ao papel da Estatística também podem ser obtidos, por exemplo, em Ishikawa (1986).

Diante disso, os dados da Tabela A mostram os níveis de algumas variáveis econômicas, refletindo a vantagem adquirida e mantida pelos japoneses, veja Pilagallo (1993a).

Tabela A - Principais indicadores dos níveis de qualidade e produtividade da indústria brasileira em 1990, em contraste com a média mundial (Europa e USA) e com o Japão.

Indicadores	Brasil 1990	Média Mundial (Europa e USA)	Japão
Índice de rejeição (Quantidade de peças/produtos defeituosos por milhão produzido)	23 mil a 28 mil	200	10
Retrabalho (% de peças/produtos que são corrigidos)	30 %	2 %	0,001 %
Gasto da empresa com assistência técnica (% do valor bruto das vendas durante a garantia do produto)	2,7 %	0,1 %	menos de 0,05 %
Tempo médio de entrega (Entre a chegada do pedido na fábrica e a entrega do produto no cliente)	35 dias	2 a 4 dias	2 dias
Rotatividade do estoque (N° de vezes em que o estoque é renovado/ano)	8	60 a 70	150 a 200
Quebra de máquinas (% de tempo parado)	40 %	15 a 20 %	5 a 8 %
Investimento em pesquisa e desenvolvimento (% sobre o faturamento)	Menor que 1 %	3 a 5 %	8 a 12 %
Treinamento (% das horas/empregado/ano)	Menor que 1 %	5 a 7 %	10 %
N° de níveis hierárquicos (Da diretoria ao operário)	10 a 12	7	3

Fonte: Pilagallo (1993a)

No caso específico do Brasil, o desenvolvimento da qualidade deve enfrentar numerosos obstáculos, entre os quais o tamanho reduzido e a pouca diversificação dos grupos econômicos brasileiros; a falta de financiamento público e privado para empreendimentos de alto risco; e, a ineficiência do sistema educacional e dos instrumentos de coordenação e promoção industrial do Estado, veja Schwartz (1991).

A superação desses obstáculos - que é um dos objetivos básicos do Programa Brasileiro de Qualidade e Produtividade, veja Gandour (1990) e Anônimo (1990) - tem permitido, porém, melhorar nos últimos anos os índices exibidos na Tabela A. Por exemplo, o índice de rejeição caiu para uma média de 11 mil a 15 mil unidades por milhão produzido; o tempo médio de entrega diminuiu para 20 dias e o número de níveis hierárquicos entre a diretoria e o operário atualmente está entre 4 e 8. Essas informações são o resultado de uma pesquisa realizada pela Iman Consultoria Ltda, que revelou uma sensível melhora dos níveis de qualidade e produtividade da indústria brasileira em 1993, embora ainda estejam longe da média mundial, veja Pilagallo (1993a e 1993b).

Os índices atingidos pela indústria japonesa refletem sua grande competitividade, com produtos de alta qualidade a baixo custo. Essa revolução da qualidade é o resultado do trabalho desenvolvido, entre outros, pelo Professor W. Deming, que alertou os dirigentes das empresas para a necessidade de criar um ambiente no qual as ferramentas estatísticas possam ser usadas. Sem um ambiente propício, a discussão dessas técnicas é infrutífera, veja Box & Bisgaard (1987).

No que se segue será estudada uma componente básica desse esforço de modernização, que é a Tecnologia da Qualidade.

1.2 TECNOLOGIA DA QUALIDADE E ESTATÍSTICA

A Tecnologia da Qualidade pode ser definida como o conjunto de atividades administrativas, operacionais, científicas e/ou de engenharia utilizadas para atingir o nível de qualidade pré-especificado para um processo, produto ou serviço. Ela consiste essencialmente na utilização do método científico e do trabalho em equipe para aprender sobre e melhorar os

processos produtivos e seus resultados. Em outras palavras, permite identificar, medir, projetar, melhorar ou controlar a qualidade de um produto ou processo.

Essa aplicação do método científico utiliza duas idéias essenciais: 1) por sua própria natureza, os processos produtivos e os serviços estão constantemente gerando dados e, 2) esses dados podem ser transformados em informação.

A ciência que permite organizar a geração de dados e sua transformação em informação é precisamente a Estatística, que pode ser definida como o estudo da variabilidade e a medição da conseqüente incerteza, para extrair eficientemente a informação necessária em estudos científicos e tecnológicos das mais diversas áreas. Assim, a aplicação da Estatística é parte fundamental do método científico. Ele permite aprender sobre um processo combinando observadores perspicazes por um lado e eventos críticos (carregados de informação) por outro.

Uma primeira forma de aprendizagem é a coleta sistemática de eventos críticos, realizada por observadores que não interferem no processo. Esses estudos - denominados observacionais - permitem aprender mediante a acumulação relativamente automática e a análise de evidências, realizadas com métodos estatísticos. No Capítulo 2 será fornecido um exemplo de estudos observacionais (Exemplo A).

Uma segunda forma de aprendizagem é a experimentação, caracterizada pela intervenção ativa e deliberada do observador, no processo que gera os eventos críticos. A execução e análise dessas intervenções - possibilitadas pela Estatística - permite acelerar enormemente o melhoramento do processo. Um exemplo de estudo experimental na área da Tecnologia da Qualidade aparece no Exemplo B do Capítulo 2 e seguintes.

Nessa Dissertação é demonstrada a utilização da Estatística na Tecnologia da Qualidade. Reciprocamente, essa última também influencia a Estatística, como comprova o surgimento e desenvolvimento dos métodos propostos pelo engenheiro japonês G. Taguchi.

Até aqui foram apresentados aspectos gerais relacionados com o melhoramento da qualidade. No entanto, o objetivo dessa dissertação é apresentar e discutir algumas técnicas estatísticas que têm sido utilizadas para analisar experimentos industriais com resposta categórica ordenada, planejados para melhorar a qualidade de produtos ou processos. Em especial, serão comparados dois métodos importantes para analisar dados desse tipo, quais sejam, o método de Taguchi e um modelo proposto por P. McCullagh, descrevendo suas limitações e alternativas. No que segue, será detalhada a estrutura do trabalho.

1.3 ESTRUTURA DA DISSERTAÇÃO

Nos estudos para o melhoramento da qualidade, com relativa freqüência a resposta de interesse não pode ser medida quantitativamente. No entanto, uma característica especial dessa resposta é a possibilidade de estar associada a uma variável latente contínua. Os dados observados podem ser vistos como uma categorização dessa variável contínua, não observável diretamente. Assim, a característica de qualidade é medida segundo uma variável categórica ordenada.

A análise de respostas ordinais será discutida no Capítulo 2. Inicialmente serão introduzidas definições das escalas de medida utilizadas habitualmente, assim como alguns exemplos de estudos experimentais e observacionais com respostas ordinais. Dois conjuntos de dados apresentados na literatura, denominados de Exemplo A e Exemplo B serão considerados em detalhe. O primeiro é um estudo observacional e o segundo é um experimento para a melhoria da qualidade. Nesse capítulo eles serão utilizados somente para ilustrar as características e limitações dos métodos simples de análise para dados ordinais. Outras técnicas de análise estatística mais elaborada são necessárias e serão desenvolvidas nos capítulos seguintes.

No Capítulo 3, inicialmente serão mencionadas algumas contribuições de Taguchi na área da Tecnologia da Qualidade. Em seguida será apresentada a técnica da Análise de Acumulação apresentada em Taguchi (1987) para examinar experimentos com resposta ordinal. O desenvolvimento formal do método será dividido em dois casos: o correspondente à situação de um único fator explanatório e o caso multifatorial. Os dados observacionais do Exemplo A serão utilizados para exibir a utilidade da técnica no caso unifatorial. Posteriormente, o Método de Taguchi será aplicado aos dados experimentais do Exemplo B. Também serão descritas suas principais limitações metodológicas e algumas modificações recentemente sugeridas na literatura para melhorar seu desempenho.

Tendo em vista as limitações do método de Taguchi para análise de respostas ordinais, no Capítulo 4 será apresentado o modelo de McCullagh, que utiliza as idéias dos modelos de regressão para interpretar as relações entre uma resposta ordinal e as covariáveis. Em particular, será formalmente desenvolvido o modelo de odds proporcionais, com observações sobre as dificuldades de estimação dos parâmetros. Também serão mencionados aspectos computacionais do ajuste desses modelos. A derivação do modelo de odds proporcionais na Seção 4.1.1 tem um caráter mais técnico e porisso o formalismo matemático empregado pode ser dispensado em uma

primeira leitura. Essa sugestão também é válida para a Seção 4.1.2, que trata do método de estimação dos parâmetros do modelo.

Para exemplificar o método, o ajuste de um modelo de odds proporcionais será desenvolvido passo a passo para os dados do Exemplo A, realizando posteriormente o ajuste através de pacotes estatísticos disponíveis. (Detalhes sobre essa implementação computacional encontram-se no Anexo B). A Tabela 4.4 exibe um resumo dos resultados das técnicas de análise empregadas nesse exemplo. Os dados do Exemplo B serão utilizados para ilustrar uma aplicação do modelo de McCullagh em um experimento para a melhoria da qualidade. Esses resultados serão comparados com aqueles obtidos pela técnica de Análise de Acumulação e pelos métodos utilizados por pesquisadores que têm reanalisado os dados do experimento. A Tabela 4.9 contém um resumo dos diferentes resultados obtidos.

O Capítulo 5 exibe algumas considerações finais sobre os métodos examinados, indicando também aspectos que ainda devem ser pesquisados.

Os Anexos contém informações importantes organizadas em duas partes. O Anexo A, composto por quatro seções, trata dos cálculos e demonstrações necessários para o desenvolvimento dos Capítulos 3 e 4. O Anexo B, também com quatro seções, exibe as rotinas dos pacotes estatísticos que permitem ajustar os modelos de odds proporcionais aos dados dos exemplos, bem como os respectivos resultados produzidos por elas.

Finalmente, serão apresentadas as referências bibliográficas utilizadas nessa dissertação.

CAPÍTULO 2: ANÁLISE DE RESPOSTAS ORDINAIS

Nesse capítulo serão abordados aspectos gerais sobre a análise de respostas ordinais. Inicialmente serão apresentadas definições para as escalas de medida usuais, bem como ilustrações de experimentos com resposta categórica ordenada. Em seguida, serão apresentados dois conjuntos de dados denominados de Exemplo A e Exemplo B. Eles serão utilizados ao longo da dissertação para exibir a aplicação de diversas técnicas de análise estatística para essa situação. Os resultados dessas análises estão resumidos na Tabela 4.4 e Tabela 4.9, respectivamente para os Exemplos A e B. Também serão mencionadas algumas técnicas clássicas para análise de dados com resposta ordinal.

2.1 VARIÁVEIS ORDINAIS

Um sistema de medida é um procedimento operacional que utiliza uma regra para atribuir números ou outros rótulos a indivíduos (pessoas, objetos ou eventos). A regra usualmente especifica as categorias de um atributo variável ou algum aspecto quantitativo de uma observação variável, definindo, assim, uma *escala de medida*. Escalas de medida comumente são classificadas como sendo *nominais, ordinais* de *intervalo* e de *razão*, podendo medir variáveis discretas ou contínuas, veja Cureton (1978, p. 764).

Variáveis cuja escala de medida consiste de um conjunto de categorias disjuntas, são denominadas variáveis categóricas. Elas surgem freqüentemente nas mais diversas áreas do conhecimento, tais como ciências sociais, epidemiologia, ecologia, educação, medicina, etc. Por exemplo, o estado de evolução de uma doença pode ser medido como "doença progressiva", "remissão parcial" ou "remissão completa". As variáveis categóricas também surgem em áreas quantitativas, como na engenharia da qualidade industrial. Por exemplo, em um experimento para

melhorar a qualidade de uma solda, podem ser cuidadosamente variados os fatores que possivelmente influenciam o resultado final. A variável resposta "qualidade da solda" pode ser classificada como "ruim", "razoável", "boa" ou "excelente".

Existem muitos tipos de variáveis categóricas, de acordo à escala de medida utilizada. Assim, variáveis categóricas para as quais não existe uma ordem natural dos níveis ou categorias são ditas *nominais*. Em uma escala nominal, os números meramente identificam os indivíduos ou as categorias de um atributo através do qual os indivíduos podem ser classificados. Os números atribuídos aos jogadores de futebol constituem um bom exemplo de escala nominal. Sem perda de informação, letras, palavras ou símbolos arbitrários poderiam ser empregados nesse caso. Exemplos de variáveis nominais são estado civil (solteiro, casado, divorciado, viúvo, desquitado) e religião (católica, protestante, judaica, outra). Para as variáveis nominais, a ordem em que aparecem as categorias deveria ser irrelevante na análise estatística, no sentido de que diferentes permutações na ordem das mesmas deveriam conduzir aos mesmos resultados.

Em muitas variáveis categóricas, contudo, existe uma ordem natural dos seus níveis, mas as distâncias absolutas entre eles são desconhecidas ou nem mesmo estão definidas. Essas variáveis são chamadas de *categóricas ordenadas*. A principal característica de um conjunto de categorias ordenadas é que elas expressam, em ordem crescente ou decrescente, a extensão ou o grau de intensidade de algum fenômeno observável. O exemplo anterior, relativo a melhoria da qualidade da solda, constitui uma aplicação na área da experimentação industrial. Outros exemplos são classe social (baixa, média, alta) e atitude política (liberal, moderado, conservador). Variáveis contínuas medidas através de postos ou escores (denominados ranks em inglês) também são tratadas como categóricas ordenadas.

Métodos estatísticos para a análise de dados categóricos são encontrados na literatura. Dentre as referências mais importantes, têm-se Agresti (1984), Agresti (1990), Bishop, Fienberg and Holland (1975), Moses et al. (1984), Anderson (1984), McCullagh (1980) ou McCullagh & Nelder (1989). No entanto, a utilização adequada de tais métodos depende das características específicas de cada situação.

A seguir serão apresentadas algumas ilustrações de experimentos com respostas ordinais.

2.2 ILUSTRAÇÕES

As ilustrações apresentadas nessa seção são úteis para melhor compreender a natureza da resposta ordinal de certos experimentos.

Ilustração 1: (Ensaio clínico). Para avaliar a eficácia do tratamento da candidíase oral crônica mediante a droga denominada clotrimazole, foi planejado um experimento. Utilizando um sistema de aleatorização foram definidos dois grupos de 10 indivíduos: o grupo controle ao qual foi administrado um placebo e o grupo de pacientes tratados, que receberam a droga.

Os dados mostrados na tabela abaixo ilustram essa questão; eles foram publicados por Kirkpatrick & Alling (1978) e posteriormente analisados por Moses et al. (1984). A ordem de classificação é explicada nessas referências. ■

Tabela 2.A - Freqüências nas categorias de resposta ordenadas no ensaio clínico para tratamento de candidíase oral crônica.

Tratamento	Categoria de Resposta				Total
	1	2	3	4	
Clotrimazole	6	3	1	0	10
Placebo	1	0	0	9	10

Ilustração 2: (Experimento Industrial I). Na National Railway Corporation do Japão foi realizado um experimento, apresentado por Taguchi & Wu (1980) e discutido também por Hamada & Wu (1990). O objetivo do experimento é encontrar os fatores que afetam a viabilidade de um procedimento para soldar duas chapas de aço. A viabilidade é definida como o grau de dificuldade no processo de soldagem de duas chapas, classificado segundo três categorias: fácil, mediano e difícil.

Inicialmente, os pesquisadores estavam interessados em 9 fatores isolados (A-I) e em quatro combinações de dois fatores (AG, AH, AC, GH). O plano experimental escolhido foi um

fatorial fracionado do tipo 2^{9-5}, com uma observação para cada subexperimento (denominado run, em inglês). ■

Ilustração 3: (Experimento Industrial II). Uma companhia fabrica um produto de borracha que deve satisfazer certas especificações, não manchando o painel no qual será fixado. Para determinar os fatores importantes que influenciam as características de qualidade, foram incorporados quatro fatores no delineamento experimental. Eles eram os compostos químicos usados na fabricação do produto. Em cada subexperimento de um delineamento fatorial fracionado do tipo 2^{4-1}, uma unidade do produto foi fixada em um painel de metal pintado, o qual foi submetido a temperatura alta durante 3 dias. O painel foi então inspecionado e suas características foram classificadas em uma das seguintes três categorias de dano: nenhum a muito leve; leve a moderado; moderadamente severo a severo.

Os dados e uma análise inicial desse experimento foram apresentados por Lear & Stanton (1985) e discutidas posteriormente por Hamada & Wu (1990). ■

Ilustração 4: (Experimento Industrial III). Um experimento industrial analisado por Phadke et al. (1983), foi realizado com o objetivo de otimizar o processo de formação das denominadas janelas de contato em circuitos integrados. As janelas de contato facilitam as interconexões entre as portas, fontes e trilhas dos circuitos. O tamanho ideal das janelas deve ficar em torno de 3 μm, sendo importante manter essa dimensão. Janelas não abertas ou muito pequenas resultam na perda de contato com o dispositivo, enquanto as janelas excessivamente largas provocam perdas nas características básicas do dispositivo.

No experimento foram incorporados oito fatores para o controle do tamanho das janelas, denominados A, BD, C, E, F, G, H e I. O delineamento escolhido foi o plano ortogonal L18 proposto por Taguchi, variando o fator A em dois níveis e os demais fatores em três níveis, com 10 repetições em cada um dos 18 subexperimentos. Os dados foram agrupados em cinco categorias de resposta, de acordo com o tamanho das janelas: 1 - janela não aberta; 2 - $(0; 2{,}55)\,\mu m$; 3 - $[2{,}55; 2{,}75)\,\mu m$; 4 - $[2{,}75; 3{,}25)\,\mu m$ e 5 - $> 3{,}25\,\mu m$.

Os dados desse experimento foram reanalisados por Hamada & Wu (1990), Agresti (1986), Nair (1986) e Box & Jones (1986a e1986b). ■

A seguir serão apresentados dois exemplos suplementares de estudos com variáveis ordinais, denominados Exemplo A e Exemplo B. O primeiro deles é um estudo observacional e o segundo um experimento. Eles serão tratados com especial destaque nessa dissertação, sendo objeto de diversos tratamentos estatísticos.

2.3 EXEMPLO A: TAMANHO RELATIVO DE AMÍGDALAS

Em uma investigação procurou-se avaliar a relação entre a presença da bactéria *Streptococcus pyogenes* e o aumento das amígdalas em crianças.

A Tabela 2.1 apresenta o correspondente conjunto de dados, referentes a classificação de 1398 crianças entre 0 a 15 anos de acordo com o tamanho relativo de suas amígdalas e com a característica "portadora" ou "não portadora" de *Streptococcus pyogenes*. A informação foi apresentada por Holmes &Williams (1954) e analisada por Armitage (1955), Armitage (1974) e McCullagh (1980).

Como os dados foram coletados para investigar a natureza e direção de um possível efeito do *Streptococcus pyogenes* no tamanho das amígdalas, o tamanho das amígdalas, com três categorias ordenadas, é considerado como a variável resposta. Entanto, a presença ou ausência de *Streptococcus pyogenes* é um possível fator explanatório. ■

Tabela 2.1 - Freqüências de indivíduos segundo o tamanho relativo das amígdalas e a presença de *Streptococcus pyogenes*.

Streptococcus	Tamanho relativo da amígdala			
pyogenes	Presente mas não aumentada	Aumentada	Grandemente aumentada	Total
Portadoras	19	29	24	72
Não portadoras	497	560	269	1326
Total	516	589	293	1398

Fonte: McCullagh (1980).

A seguir serão explorados alguns resultados obtidos após aplicar métodos estatísticos simples aos dados da Tabela 2.1.

A estatística χ^2 de Pearson pode ser utilizada para avaliar a hipótese de que as proporções de indivíduos com diferentes tamanhos de amígdalas são iguais nos grupos de portadores e não portadores de *Streptococcus pyogenes*. Em outras palavras, a estatística χ^2 permitirá testar se as proporções relativas de indivíduos nas diferentes categorias de resposta permanecem iguais para as duas populações.

Realizando os cálculos do modo usual, o valor da estatística qui-quadrado de Pearson (com 2 graus de liberdade), sob a hipótese nula é $\chi^2_{(2)} = 7{,}8848$, correspondendo a uma probabilidade de significância $p = 0{,}0194$. Isso sugere que as diferenças observadas entre as proporções das respectivas categorias de tamanho de amígdalas, entre os portadores e não portadores, não são exclusivamente devidas a flutuações aleatórias. Os portadores de *Streptococcus pyogenes* com idades entre 0 e 15 anos, parecem apresentar tamanhos relativos de amígdalas diferentes dos não portadores.

Contudo, a estatística χ^2 não informa sobre a direção do efeito da presença de *Streptococcus pyogenes* sobre o tamanho das amígdalas. De fato, essa estatística mede a discrepância geral das proporções esperadas em relação às observadas, mas não leva em conta a ordem das categorias. A decomposição da estatística χ^2 geral é, entanto, um procedimento útil para detectar a existência de tendências (lineares, por exemplo), tais como o crescimento do número de indivíduos portadores com o aumento das amígdalas, veja Snedecor & Cochran (1980, p.206) ou Armitage (1974, p.363).

Denote por P_j a proporção de indivíduos com tamanho de amígdalas correspondente à categoria j $(j = 1,2,3)$, na população de portadores de *Streptococcus pyogenes*. Sejam a_j e $p_j = \frac{a_j}{n_j}$, respectivamente, o número e a proporção observada de indivíduos portadores de *Streptococcus pyogenes* na j-ésima categoria ordenada e seja x_j o escore atribuído a essa categoria, j=1,2,3.

Então, se o tamanho das amígdalas para as crianças contaminadas pelo *Streptococcus pyogenes* tem aumentado mais do que para os não portadores, os valores de P_j tenderiam a aumentar com o crescimento dos valores de x_j de $x = 0$ para $x = 2$.

Um crescimento linear entre as proporções P_j e os escores x_j deverá seguir a relação $P_j = \alpha + \beta x_j$; j=1,2,3.

Sob a hipótese $P_j = \alpha + \beta x_j$, o coeficiente de regressão β indica a alteração na proporção por unidade de mudança em x, enquanto α representa a proporção esperada quando $x = 0$. Esses parâmetros podem ser estimados pelo método dos quadrados mínimos, como segue:

$$\hat{\beta} = \frac{\sum_{j=1}^{K} n_j (p_j - \bar{p})(x_j - \bar{x})}{\sum_{j=1}^{K} n_j (x_j - \bar{x})^2} = \frac{\sum_{j=1}^{K} a_j x_j - \frac{\left(\sum_{j=1}^{K} a_j\right)\left(\sum_{j=1}^{K} n_j x_j\right)}{N}}{\sum_{j=1}^{K} n_j x_j^2 - \frac{\left(\sum_{j=1}^{K} n_j x_j\right)^2}{N}}$$

e

$$\hat{\alpha} = \bar{p} - \hat{\beta}\bar{x},$$

onde $\bar{x} = \frac{1}{N}\sum_{j=1}^{K} n_j x_j$; $\bar{p} = \frac{1}{N}\sum_{j=1}^{K} n_j p_j$ e $N = \sum_{j=1}^{K} n_j$.

No exemplo, N=1398, $\bar{x} = 0{,}8405$, $\bar{p} = 0{,}0515$, $\sum_{j=1}^{3} a_j = 72$,

$\sum_{j=1}^{3} a_j x_j = 77$, $\sum_{j=1}^{3} n_j x_j = 1175$, $\sum_{j=1}^{3} n_j x_j^2 = 1761$,

de tal forma que $\hat{\beta} = 0{,}0213$ e $\hat{\alpha} = 0{,}0336$. Portanto, a reta ajustada $\hat{p}_j = 0{,}0336 + 0{,}0213 x_j$ permite calcular as predições $\hat{p}_j$ de cada categoria. Se p_j e $\hat{p}_j$ estão próximos para todas as categorias ou pelo menos para grande parte delas, isso fornece evidência de que a proporção populacional P_j tende a variar linearmente com os valores de x_j. Por outro lado, se p_j e $\hat{p}_j$ diferem, a associação entre P_j e x_j pode se desviar da relação linear suposta. As diferenças

$(p_j - \hat{p}_j)$ também servem para identificar as categorias que apresentam maiores divergências com respeito à linearidade assumida. Veja a Tabela 2.2.

Tabela 2.2 - Proporções observadas e linearmente preditas, para os portadores e não portadores de *Streptococcus pyogenes*.

Streptococcus	Tamanho relativo da amígdala			
pyogenes	Presente mas não aumentada	Aumentada	Grandemente aumentada	Total
Portadores (a_j)	19	29	24	72
Não portadores	497	560	269	1326
Total (n_j)	516	589	293	1398
X_j	0	1	2	
$p_j = a_j / n_j$	0,0368	0,0492	0,0819	
$p_j - \hat{p}_j$	0,0032	-0,0057	0,0057	

Para testar se a relação entre P_j e x_j é linear utiliza-se a estatística

$$\chi^2_{RES} = \frac{1}{\bar{p}(1-\bar{p})} \sum_{j=1}^{K} n_j (p_j - \hat{p}_j)^2$$

que possui uma distribuição assintótica de χ^2 com $K-2$ graus de liberdade. A hipótese de linearidade deverá ser rejeitada se a estatística χ^2_{RES} assume valores grandes.

Alternativamente, χ^2_{RES} pode ser calculada através da relação

$$\chi^2 = \chi^2_{RES} + \chi^2_{LIN} \qquad (2.1)$$

onde χ^2 é a estatística qui-quadrado de Pearson determinada sob a hipótese de homogeneidade acima descrita e a estatística χ^2_{LIN} possui uma distribuição assintótica de qui-quadrado com 1 grau

de liberdade. Ela corresponde à componente linear e pode ser usada para testar a significância do coeficiente de regressão $\hat{\beta}$, sendo determinada por

$$\chi^2_{LIN} = \hat{\beta}^2 \frac{1}{\bar{p}(1-\bar{p})} \sum_{j=1}^{K} n_j (x_j - \bar{x})^2 .$$

Dessa forma, valores altos de χ^2_{LIN} fornecem evidência contra a hipótese $\beta = 0$, sugerindo também que o crescimento dos valores de x_j é acompanhado por um crescimento de P_j (se β for positivo) ou por um decréscimo P_j (se β for negativo). No exemplo,

$$\chi^2_{LIN} = (0{,}0213)^2 \frac{1}{(0{,}0515)(1-0{,}0515)} 773{,}4285 = 7{,}1927$$

e

$$\chi^2_{RES} = \frac{1}{(0{,}0515)(1-0{,}0515)} 0{,}0338 = 0{,}6921 .$$

Como pode ser observado, a relação $\chi^2 = \chi^2_{LIN} + \chi^2_{RES}$ é satisfeita. Essa decomposição pode ser apresentada através de um quadro como o da Tabela 2.3.

Tabela 2.3 - Decomposição da estatística χ^2 geral segundo a relação $P_j = \alpha + \beta x_j$.

Componente	g.l.	Qui-quadrado	p-value
Linear	1	$\chi^2_{LIN} = 7{,}1927$	$p < 0{,}01$
Residual	1	$\chi^2_{RES} = 0{,}6921$	$p < 0{,}5$
Total	2	$\chi^2 = 7{,}8848$	$p < 0{,}025$

O valor da estatística χ^2_{RES} indica que as diferenças entre as proporções observadas e as linearmente preditas são pequenas, levando a considerar como razoável a hipótese de linearidade.

Por sua vez, através da estatística χ^2_{LIN} rejeita-se a hipótese de que o coeficiente de regressão é zero.

A conclusão que se obtém dessa análise é, assim, que há evidências de que a proporção populacional de crianças portadoras de *Streptococcus pyogenes* cresce linearmente quando o tamanho de amígdalas se movimenta da categoria não aumentada $(x = 0)$ para grandemente aumentada $(x = 2)$. Em outras palavras, a proporção de indivíduos classificados nas categorias superiores de tamanho das amígdalas é maior na população de portadores do que na população de não portadores, o que sugere que as crianças infectadas pelo *Streptococcus pyogenes* apresentam amígdalas aumentadas.

Outro método para analisar esses dados é o teste de Mann-Whitney (também conhecido como teste de Wilcoxon) para amostras independentes, veja Lehmann (1975, p.5), Conover (1980, p.215) Daniel (1978, p.82) ou Snedecor &Cochran (1967, p.130).

Nesse teste, a hipótese nula de que as duas amostras (portadores e não portadores de *Streptococcus pyogenes*) foram retiradas da mesma população corresponde ao fato de que a proporção de indivíduos com tamanho relativo de amígdala na categoria j; j=1,2,3, é igual para os portadores e não portadores. Em contrapartida, a hipótese alternativa afirma que as proporções de indivíduos com amígdalas grandemente aumentadas e aumentadas são maiores no grupo de indivíduos contaminados em relação aos não contaminados.

Para amostras grandes, a distribuição amostral da estatística de teste é aproximadamente normal, o que permite obter o nível de significância p=0,0045, sugerindo que as crianças portadoras de *Streptococcus pyogenes* possuem amígdalas maiores do que as não portadoras.

É importante observar que as técnicas de análise apresentadas não permitem, em geral, estimar o impacto da contaminação por *Streptococcus pyogenes* sobre o tamanho das amígdalas. Porém, freqüentemente, essa estimação de efeitos é um objetivo importante das investigações. Assim, nessas condições, os métodos abordados não são suficientes para obter as respostas desejadas. Para tanto, técnicas de análise mais elaboradas serão discutidas nas próximas seções, depois de apresentar um segundo exemplo básico, correspondente a um estudo experimental.

2.4 EXEMPLO B: QUALIDADE DE CAMISAS TERMOPLÁSTICAS

Nesse exemplo será brevemente descrito e analisado o experimento realizado na empresa Flex Products, Inc., por uma equipe conduzida por J. Quinlan, visando o melhoramento da qualidade de camisas termoplásticas para cabos de velocímetros, veja Quinlan (1985).

As camisas termoplásticas, produzidas por extrusão, são compostas por um forro interno de polipropileno extrusionado, uma camada de fio trançado e um invólucro coextrusionado. A contração excessiva das camisas após a extrusão pode provocar perturbações na ensamblagem e, conseqüentemente, perda de qualidade.

O objetivo do experimento é encontrar as causas do encolhimento das camisas após a extrusão. Procedida a análise do correspondente diagrama de causas e efeitos (denominado também diagrama de Ishikawa ou de espinha de peixe), os engenheiros idenficaram 15 fatores potencialmente responsáveis pelo defeito. Esses fatores, bem como seus níveis ou modalidades de alteração, são mostrados na Tabela 2.4.

Entretanto, como esperava-se que apenas alguns desses fatores fossem realmente ativos, decidiu-se pela execução de um experimento dirigido à triagem dos fatores.

A exploração exaustiva das $2^{15} = 32768$ combinações possíveis dos fatores foi substituída pelo estudo da contração das camisas produzidas nas condições definidas por cada um dos 16 subexperimentos do plano ortogonal L16 proposto por Taguchi.

Assim, para cada subexperimento foram fabricados 3000 pés do produto final, totalizando 48000 pés. A realização do experimento foi simplificada mediante a elaboração de folhas de acompanhamento com a ordem (aleatorizada quando possível) dos subexperimentos. Após amostragem aleatória de segmentos de aproximadamente 600 mm em cada porção de 3000 pés, o subexperimento incluiu testes de embebição quente, em 4 dias consecutivos, para determinar o percentual de contração Y como resposta, para cada segmento. Os resultados obtidos são exibidos na Tabela 2.5.

O método de análise utilizado por Quinlan (1985) consistiu em construir o *quociente sinal-ruído* $SN = -10 \log\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2\right)$ proposto por Taguchi para cada subexperimento, aferindo depois os 16 valores de SN com uma versão da técnica da análise da variância também proposta por esse autor, veja Taguchi (1987, p.625), Box (1988), Nair (1992) e Box, Bisgaard and Fung (1988). Os fatores de impacto estatisticamente significativo, em ordem de importância, foram

E,G,K,A,C,F,D e H. Aqueles com as máximas contribuições à variação total foram E e G, explicando conjuntamente mais de 70% da variabilidade dos quocientes.

Tabela 2.4 - Fatores utilizados no experimento para melhorar a qualidade de camisas termoplásticas de cabos de velocímetros.

Código	Fator	Modalidade	
		1	2
A	Diâmetro externo do forro	Usual	Modificada
B	Moldes do forro	Usual	Modificado
C	Material do forro	Usual	Modificado
D	Velocidade de linha do forro	Usual	80 % da usual
E	Tipo de fio	Usual	Modificado
F	Tensão de trançado	Usual	Modificada
G	Diâmetro do fio	Menor	Usual
H	Tensão do forro	Usual	Maior
I	Temperatura do forro	Ambiente	Pré-aquecida
J	Material do revestimento	Usual	Modificado
K	Tipo de molde do revestimento	Usual	Modificado
L	Temperatura de fusão	Usual	Menor
M	Compactação	Usual	Mais densa
N	Método de resfriamento	Usual	Modificado
O	Velocidade de linha	Usual	70% da usual

Fonte: Quinlan (1985).

Tabela 2.5 - Resultados do experimento para melhorar a qualidade das camisas termoplásticas conduzido por J. Quinlan, seguindo o plano ortogonal L16.

Subexpe-	Fator															Resposta			
rimento	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**T1**	**T2**	**T3**	**T4**
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,49	0,54	0,46	0,45
2	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	0,55	0,60	0,57	0,58
3	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	0,07	0,09	0,11	0,08
4	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1	0,16	0,16	0,19	0,19
5	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	0,13	0,22	0,20	0,23
6	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1	0,16	0,17	0,13	0,12
7	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	0,24	0,22	0,19	0,25
8	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2	0,13	0,19	0,19	0,19
9	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	0,08	0,10	0,14	0,18
10	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1	0,07	0,04	0,19	0,18
11	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	0,48	0,49	0,44	0,41
12	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2	0,54	0,53	0,53	0,54
13	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	0,13	0,17	0,21	0,17
14	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2	0,28	0,26	0,26	0,30
15	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2	0,34	0,32	0,30	0,41
16	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1	0,58	0,62	0,59	0,54

Fonte: Quinlan (1985).

Na reanálise desses dados experimentais, Box (1988) e Box, Bisgaard and Fung (1988) constataram que apenas os fatores E e G influem significativamente na contração das camisas.

Contudo, essa dissertação trata de métodos estatísticos para analisar experimentos com resposta ordinal. Assim, a variável resposta "percentual de contração das camisas termoplásticas após extrusão" será discretizada para sua análise. Um possível argumento para justificar a discretização da resposta nesse experimento é o seguinte: apesar da resposta ser quantitativa, poderia existir algum tipo de dificuldade para medí-la adequadamente (em termos de precisão, custo, tempo, etc.). Dessa forma, poderia ser mais proveitoso observá-la segundo categorias ordenadas.

Assim, o critério adotado foi: para um percentual de contração no intervalo [0; 0,2) foi observada a categoria 1; no intervalo [0,2; 0,4) foi observada a categoria 2 e para o intervalo [0,4; 1) foi observada a categoria 3. Dessa maneira foi construída a nova variável resposta grau de contração das camisas termoplásticas após extrusão, medida segundo três categorias ordenadas. Por sua vez, as categorias 1, 2 e 3 podem ser rotuladas de "leve", "médio" e "forte", respectivamente. A relação física conhecida é que quanto maior o grau de contração das camisas termoplásticas, mais graves serão as perturbações na ensamblagem, afetando negativamente as características de qualidade do produto, sendo natural, então, a necessidade de identificar o impacto dos fatores que provocam a contração das camisas.

A Tabela 2.6 contém os resultados do experimento realizado por Quinlan, expressos mediante a resposta ordinal acima definida. As colunas do experimento original foram reorganizadas de acordo com a ordem padrão proposta por Yates (1937).

Através da estatística χ^2 pode ser testada a hipótese de independência entre as diferentes combinações das modalidades dos fatores com o grau de contração das camisas. O valor observado da estatística de teste foi $\chi^2 = 107{,}03$, com 30 graus de liberdade, correspondendo a um nível de significância p inferior a 10^{-6}. Isso sugere que as diferenças observadas nas proporções das respectivas categorias de grau de contração, entre os 16 subexperimentos, não são exclusivamente devidas ao acaso.

Tabela 2.6 - Freqüências de respostas no experimento para melhorar a qualidade das camisas termoplásticas para cabos de velocímetros.

Subexpe-	Fator															Resposta		
rimento	H	D	-L	B	-J	-F	N	A	-I	-E	M	-C	K	G	-O	1	2	3
1	-	-	+	-	+	+	-	-	+	+	-	+	-	-	+	0	0	4
2	+	-	-	-	-	+	+	-	-	+	+	+	+	-	-	0	0	4
3	-	+	-	-	+	-	+	-	+	-	+	+	-	+	-	4	0	0
4	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	4	0	0
5	-	-	+	+	-	-	+	-	+	+	-	-	+	+	-	1	3	0
6	+	-	-	+	+	-	-	-	-	+	+	-	-	+	+	4	0	0
7	-	+	-	+	-	+	-	-	+	-	+	-	+	-	+	1	3	0
8	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	4	0	0
9	-	-	+	-	+	+	-	+	-	-	+	-	+	+	-	4	0	0
10	+	-	-	-	-	+	+	+	+	-	-	-	-	+	+	4	0	0
11	-	+	-	-	+	-	+	+	-	+	-	-	+	-	+	0	0	4
12	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-	0	0	4
13	-	-	+	+	-	-	+	+	-	-	+	+	-	-	+	3	1	0
14	+	-	-	+	+	-	-	+	+	-	-	+	+	-	-	0	4	0
15	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	0	3	1
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	4

No entanto, com base nessa estatística não é possível determinar as combinações de modalidades dos fatores que ocasionam essas diferenças, sendo também desconsiderada a estrutura ordenada das respostas. É importante ressaltar também que todas as células da tabela de contingência à direita da Tabela 2.6 apresentam freqüências esperadas inferiores a 5, ficando aparente sua inadequacidade para a aplicação do teste de qui-quadrado, veja Everitt (1992, p.39).

Outro método para analisar os resultados do experimento dispostos na Tabela 2.6 consiste em usar o teste não-paramétrico de Kruskal-Wallis, veja Lehmann (1975, p.205), Conover (1980, p.229) ou Daniel (1978, p.200). Ele requer a suposição básica de que os 16 subexperimentos constituem amostras independentes de tamanho $n_i = 4$, $i = 1,2,...,16$ (com observações independentes tanto dentro como entre as amostras). Também é necessário que a variável de interesse seja contínua e medida no mínimo na escala ordinal e, finalmente, que as populações sejam idênticas, exceto para uma possível diferença na posição pelo menos para uma delas.

A hipótese nula do teste afirma que as proporções de camisas termoplásticas nas respectivas categorias do grau de contração são iguais para as 16 combinações dos níveis dos fatores. O valor observado da estatística de teste de Kruskal-Wallis atinge um nível de significância inferior a 10^{-4}, sugerindo que o grau de contração das camisas não é igual para as distintas combinações de modalidades dos fatores. Contudo, a aplicação desse método não permite

identificar quais as combinações que contribuem de maneira mais acentuada para aumentar ou diminuir o grau de contração das camisas, nem para estimar os efeitos desses fatores.

Esse exemplo voltará a ser analisado nos capítulos seguintes, nos quais serão abordados métodos que possibilitam obter essas respostas. A Tabela 4.9 apresenta um resumo dos resultados obtidos através da aplicação de diversos métodos de análise, quanto ao impacto produzido pelos fatores no grau de contração das camisas.

Como pode ser constatado, as técnicas de análise de dados ordinais abordados até aqui apresentam limitações, que serão brevemente comentados na próxima seção. Também serão introduzidas algumas técnicas de análise estatística que permitem superar parcialmente essas dificuldades. No entanto, duas técnicas sofisticadas para análise de dados nessa situação serão apresentadas nos Capítulos 3 e 4.

2.5 MÉTODOS CLÁSSICOS DE ANÁLISE

Inicialmente serão comentadas as principais deficiências das técnicas de análise de dados ordinais já mencionadas. Em seguida serão apresentados alguns métodos estatísticos mais elaborados para o exame desse tipo de dados.

O primeiro método abordado, baseado na estatística χ^2, não informa a direção dos efeitos dos fatores envolvidos no experimento. A decomposição (2.1) é um procedimento útil somente quando existem dois tratamentos a serem comparados, pois, em outra situação, seriam necessárias análises múltiplas desse tipo para identificar o(s) tratamento(s) que influem na resposta.

Outro problema relacionado com a estatística χ^2 é que nos experimentos industriais o número de replicações geralmente é pequeno, acarretando um grande número de células de contingência com freqüências esperadas inferiores a 5.

Os métodos não-paramétricos, por sua vez, também não permitem identificar a direção dos efeitos dos fatores, não detectando assim as combinações que otimizam as respostas.

Por fim, uma limitação comum a todos os métodos acima é que não permitem estimar os efeitos dos fatores. Na maioria dos experimentos, porém, o pesquisador deseja avaliar quantitativamente essas magnitudes, individual e conjuntamente.

Assim sendo, constata-se a necessidade de utilizar técnicas de análise estatística que permitam ao experimentador obter essas informações. Nesse sentido serão apresentados agora

alguns métodos mais elaborados para o exame de dados produzidos em situações experimentais ou observacionais, quando a resposta é medida através de uma resposta categórica ordenada.

Nessa situação, Bross (1958) propôs uma técnica denominada *ridit analysis*, na qual uma função de distribuição empírica é usada para atribuir escores - chamados *ridits* - às categorias ordenadas da resposta.

Nesse caso, a única suposição feita é que as categorias de resposta representam intervalos disjuntos de uma distribuição correspondente a uma variável contínua não observável. A ordem das categorias é conhecida, mas não está disponível uma escala numérica, nem são feitas suposições quanto à forma da distribuição.

Para aplicar essa técnica de análise, deve ser identificada uma distribuição de referência para as categorias de resposta. Assim, a análise deve começar pela definição de uma população ou grupo de referência através de uma fonte externa ou de informação fornecida pelos grupos observados. Veja Flora (1988) ou Fleiss (1973).

Considere agora os dados do Exemplo A. Para o cálculo dos ridits, será utilizada a população de não portadores de *Streptococcus pyogenes* como distribuição de referência.

Na i-ésima população (ou grupo), seja P_{ij} a proporção de indivíduos na j-ésima categoria do tamanho de amígdalas. Assim, os escores ou ridits são definidos por

$$R_j = \sum_{t=1}^{j-1} P_{2t} + \frac{1}{2} P_{2j}.$$

Para os dados do exemplo em questão, resulta

$$R_1 = \frac{1}{2} P_{21} = \frac{1}{2} \frac{497}{1326} = 0{,}1874$$

$$R_2 = P_{21} + \frac{1}{2} P_{22} = \frac{497}{1326} + \frac{1}{2} \frac{560}{1326} = 0{,}5860$$

$$R_3 = P_{21} + P_{22} + \frac{1}{2} P_{23} = \frac{497}{1326} + \frac{560}{1326} + \frac{1}{2} \frac{269}{1326} = 0{,}8986.$$

Conhecidas as freqüências observadas nas mesmas categorias de resposta para qualquer grupo, podem ser calculados os ridits médios dos respectivos grupos, os quais são

interpretados como probabilidades. Para o grupo de referência, o ridit médio é necessariamente igual a $\frac{1}{2}$.

Sejam X e Y, respectivamente, o tamanho relativo das amígdalas de um indivíduo escolhido ao acaso na população formada pelos não portadores e pelos portadores de *Streptococcus pyogenes*. Então, o ridit médio da população de portadores da bactéria é definido por $\overline{R} = \sum_{j=1}^{K} R_j P_{1j}$ e pode ser interpretado como uma estimativa da probabilidade $P[X < Y]$ de que um indivíduo não portador tenha amígdalas menores do que um portador. No exemplo,

$$\overline{R} = 0{,}1874\ \frac{19}{72} + 05860\ \frac{29}{72} + 0{,}8986\ \frac{24}{72} = 0{,}5850.$$

O erro padrão de $\overline{R}$ é aproximadamente $ep(\overline{R}) = \left[2\sqrt{3n_i}\right]^{-1}$, onde n_i é o número de indivíduos observados no i-ésimo grupo. Assim, para os portadores de *Streptococcus pyogenes* $\overline{R} = 0{,}5850$, $ep(\overline{R}) = 0{,}0340$. A significância da diferença entre o ridit médio do grupo de portadores com o grupo de referência (não portadores), pode ser testada através do valor observado da estatística $z = \frac{\overline{R} - 0{,}5}{\text{e.p.}(\overline{R})}$, que possui assintoticamente uma distribuição de probabilidade normal padrão. No exemplo $z = 2{,}4985$, atinge um nível de significância inferior a 5%.

A conclusão que se obtém dessa análise é que o tamanho relativo das amígdalas nos portadores de *Streptococcus pyogenes* parece ser maior.

No contexto da análise de experimentos industriais, foco dessa dissertação, o método parece apresentar dificuldades semelhantes às técnicas discutidas na seção anterior. Quando o número de tratamentos é grande, deverão ser feitas várias comparações emparelhadas entre grupos. Isso freqüentemente dificulta a interpretação.

Ainda, com essa técnica não é possível determinar a magnitude dos efeitos dos fatores. Para tanto, será apresentado agora um primeiro método que permite a estimação dos efeitos, chamado de *modelo de resposta média*.

Agresti (1986) sugere a utilização do modelo de resposta média proposto por Grizzle, Starmer and Koch (1969). Esse método consiste em reduzir as respostas das K categorias

em uma única medida, com base no escore médio $\sum_{j=1}^{K} v_j p_{ij}$, para i=1,2,...,R, onde $\{v_1, v_2, ..., v_K\}$ são escores fixos atribuídos às respectivas categorias de resposta e p_{ij} a proporção observada na categoria j (j=1,2,...,K) do i-ésimo subexperimento (i=1,2,...,R). O modelo estabelece uma relação linear entre a resposta média $\sum_{j=1}^{K} v_j p_{ij}$ e os efeitos dos fatores.

No caso geral, o modelo pode ser escrito como $\underset{\sim}{F} = \underset{\sim}{X}\underset{\sim}{\beta}$, onde $\underset{\sim}{F}$ é o vetor de respostas que contém a média dos escores; $\underset{\sim}{\beta}$ é o vetor de parâmetros de $\underset{\sim}{X}$ é a matriz do delineamento. Assim, a análise consiste em encontrar as estimativas das componentes do vetor $\underset{\sim}{\beta}$, determinando se dessas estimativas resulta uma boa concordância entre as médias observadas e as preditas com base em $\underset{\sim}{X}\underset{\sim}{\beta}$. Se o ajuste é bom; isto é, se as médias preditas estão próximas das observadas, são realizados testes adicionais para verificar se todas as variáveis ou fatores são necessários na especificação do modelo, visando um modelo reduzido que preserve um bom ajuste.

Para ilustrar o método, sejam 0, 2 e 5, respectivamente, os escores atribuídos às categorias de resposta "leve", "médio" e "forte" da variável grau de contração das camisas termoplásticas no experimento do Exemplo B.

A escolha desses escores pode ser parcialmente justificada da seguinte forma: ao grau de contração "leve" foi atribuído o escore 0, pois acredita-se que não afeta substancialmente a qualidade do produto, enquanto que o escore 5 atribuído à categoria "forte" deve-se ao fato de que esse grau de contração poderá produzir um impacto muito grande na qualidade dos cabos de velocímetro.

Então, o modelo de resposta média tem a forma

$$\sum_{j=1}^{K} v_j p_{ij} = \alpha + \beta_A + \beta_B + ... + \beta_{-O}; \qquad i=1,2,...,16$$

onde α é o intercepto e os termos restantes $\beta_A, \beta_B, ..., \beta_{-O}$ são os efeitos principais dos fatores identificados pelos respectivos índices. Esse modelo foi ajustado através do procedimento PROC CATMOD do pacote estatístico SAS, obtendo-se um modelo reduzido. O mesmo inclui os fatores E,G,B,F,D,A,C,K e I, cujos efeitos principais são significantes ao nível 10^{-4}, evidenciando um

impacto significativo no grau de contração das camisas, podendo afetar a qualidade dos cabos de velocímetro.

Os parâmetros do modelo acima são estimados através do método de mínimos quadrados ponderados (WLS), que requer que as freqüências das células sejam positivas. Assim, no PROC CATMOD utilizou-se a opção **addcell=0.01** para adicionar o valor 0,01 às freqüências de cada célula; veja SAS Institute Inc. (1989). As estimativas dos parâmetros do modelo reduzido são apresentadas na Tabela 2.7. Os fatores que permaneceram no modelo reduzido possivelmente produzem um impacto na contração das camisas termoplásticas.

Tabela 2.7 - Estimativas dos parâmetros (efeitos principais) do modelo de resposta média reduzido, ajustado aos dados de Quinlan através do PROC CATMOD do SAS

Parâmetro	Estimativa	Erro Padrão
INTERCEPTO	2.1211	0.0394
D	0.3680	0.0393
B	0.3776	0.0394
-F	-0.3680	0.0393
A	-0.4698	0.0665
-I	0.3975	0.0662
-E	-1.6086	0.0393
-C	0.4698	0.0665
K	-0.3975	0.0662
G	-0.8727	0.0393

A principal vantagem do modelo de resposta média, em relação aos métodos apresentados anteriormente, é que possibilita interpretar magnitude e direção do impacto provocado pelos fatores no grau de contração das camisas, através das estimativas dos parâmetros do modelo. Embora não tenha ocorrido para esse conjunto de dados, uma possível desvantagem dessa técnica reside no fato de que, para alguns valores dos regressores, a resposta média predita pode ficar fora dos valores v_1 e v_K atribuídos às categorias extremas, veja Agresti (1986).

Outro método de análise consiste em combinar as categorias em apenas dois grupos e então analisar esses dados binários através do modelo de regressão logística. Quando a variável resposta tem 3 categorias, por exemplo, duas análises separadas podem ser feitas: uma combinando as duas primeiras categorias e, a outra, combinando as duas últimas categorias, podendo surgir

problemas de estimação, em virtude da natureza dos delineamentos altamente fracionados e face à insuficiência de dados, veja por exemplo Hamada & Wu (1990).

Finalmente, dados com resposta ordinal podem ser analisados através das técnicas propostas por G. Taguchi e P. McCullagh, as quais serão detalhadamente discutidas nos capítulos seguintes.

CAPÍTULO 3: MÉTODO DE TAGUCHI

Os métodos estatísticos empregados por G. Taguchi para melhorar a qualidade de produtos e processos têm despertado muito interesse e controvérsias. Esses métodos englobam o uso de delineamentos fatoriais fracionados e outros planos ortogonais, o delineamento de parâmetros (parameter design, em inglês) para minimizar a sensibilidade aos fatores ambientais e para minimizar a propagação da variabilidade, a definição e utilização de quocientes de sinal-ruído, de funções de perda e técnicas como a análise de acumulação e a denominada em inglês minute analysis, assim como diversas técnicas para análise de dados de sobrevivência.

Essas contribuições de Taguchi à engenharia da qualidade são apresentadas, por exemplo, em Box, Bisgaard and Fung (1987) ou Box (1988). A conclusão obtida por esses autores é que, por um lado, as idéias de Taguchi são importantes e deveriam ser conhecidas pelos engenheiros envolvidos no melhoramento da qualidade. Por outro lado, muitas das técnicas de planejamento e análise empregadas para implementar essas idéias freqüentemente são ineficientes e complicadas, podendo ser substituídas por outras mais adequadas. Uma discussão sobre as técnicas de Taguchi podem também ser encontradas no recente artigo editado por Nair (1992).

Nesse capítulo será somente descrita a técnica proposta por G. Taguchi para analisar dados com resposta ordinal, apresentando-se também os princípios e limitações do método. O Exemplo A será utilizado para ilustrar o desenvolvimento da técnica no caso de um único fator explanatório, enquanto que os dados do Exemplo B serão empregados para aplicar a técnica no caso multifatorial. A Tabela 4.4 e a Tabela 4.9 exibem um resumo dos resultados obtidos através de diversas técnicas de análise aplicadas aos dados desses exemplos.

3.1 ANÁLISE DE ACUMULAÇÃO

A fragilidade das técnicas de análise para resposta ordinal discutidas no capítulo anterior evidencia a necessidade de métodos mais elaborados.

G. Taguchi (1987, p.73) tem introduzido uma técnica denominada Análise de Acumulação (doravante denominada AA) para avaliar dados categóricos produzidos em situações experimentais.

Em linhas gerais, para uma variável reposta com K categorias ordenadas, o método da AA consiste em construir K-1 tabelas de respostas binárias, às quais é aplicada a técnica da análise da variância (analysis of variance ou ANOVA, em inglês). Posteriormente, as K-1 tabelas da ANOVA são combinadas para determinar uma Tabela da Análise de Acumulação.

Para facilitar a compreensão, o desenvolvimento formal do método da AA será dividido em dois casos. No primeiro, a situação corresponde ao caso de um único fator explanatório. O segundo caso corresponde à consideração simultânea de vários fatores.

Referências importantes sobre o método são Taguchi (1987), Nair (1986) e Hamada & Wu (1990). Como nelas não se encontram desenvolvimentos formais, o método será detalhado nas seções seguintes.

3.1.1 CASO UNIFATORIAL

No caso de um único fator explanatório, seja A um fator com I níveis e número constante n de replicações em cada nível. Seja $\underset{\sim}{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iK})^t$ o vetor de freqüências das K categorias ordenadas, observadas no i-ésimo nível do fator A. Então y_{ij} (para i=1,2,...,I e j=1,2,...,K) denota a freqüência observada na j-ésima categoria de resposta, quando o fator A assume o nível i. Os dados podem ser representados mediante a tabela de contingência de I linhas e K colunas, exibida na Tabela 3.1.

Tabela 3.1 - Matriz das freqüências observadas em um experimento com resposta ordenada.

Nível do Fator	Categoria de Resposta					Total
Explanatório	1	2	3	...	K	
1	y_{11}	y_{12}	y_{13}	...	y_{1K}	n
2	y_{21}	y_{22}	y_{23}	...	y_{2K}	n
⋮	⋮	⋮	⋮		⋮	⋮
I	y_{I1}	y_{I2}	y_{I3}	...	y_{IK}	n

Sejam $c_{ij} = \sum_{r=1}^{j} y_{ir}$, i=1,2,...,I as freqüências acumuladas das primeiras j categorias no i-ésimo nível do fator e $c_{.j}$ as médias de c_{ij} em relação aos I níveis, isto é,

$$c_{.j} = \frac{1}{I}\sum_{i=1}^{I} c_{ij} , \ \forall j = 1,2,...,K-1.$$

Como pode ser observado no desenvolvimento que segue, o método da AA imita o procedimento da análise da variância. Assim, são construídas K-1 tabelas de contingência de dimensão Ix2 , nas quais a i-ésima linha da j-ésima tabela contém o vetor $\left(c_{ij}, n - c_{ij}\right)$ de freqüências acumuladas.

Uma forma de analisar essa tabelas é considerar a ocorrência de respostas dicotômicas. Na j-ésima tabela, j=1,2,...,K-1, é atribuído o valor 1 às respostas observadas em uma categoria menor ou igual a j e o valor 0 àquelas observadas em uma categoria superior a j, para cada nível do fator. Assim, no i-ésimo nível do fator, a tabela contém c_{ij} "uns" e $n - c_{ij}$ "zeros". Em cada uma dessas tabelas aplica-se o procedimento padrão da ANOVA para dados binários.

Tabela 3.2 - Estrutura da j-ésima tabela de respostas binárias construída na AA.

Nível do	Resposta Dicotômica	
Fator	1	2
1	c_{1j}	$n - c_{1j}$
2	c_{2j}	$n - c_{2j}$
⋮	⋮	⋮
I	c_{Ij}	$n - c_{Ij}$

Para exemplificar esse procedimento, considere os dados do estudo mencionado no Exemplo A, apresentados na Tabela 2.1, cuja variável resposta é o tamanho relativo das amígdalas, enquanto que presença e ausência de *Streptococcus pyogenes* são os dois níveis do fator explanatório. Observe-se o número desigual de repetições em cada nível do fator, provocado pelo caráter observacional dos dados.

No exemplo em questão, surgem duas tabelas de dados binários. Na primeira, os indivíduos que foram classificados na primeira categoria de resposta assumem valor "1", enquanto que os demais assumem valor "0". Na segunda tabela, os indivíduos da primeira e segunda categorias assumem valor "1" e os da terceira categoria o valor "0".

Tabela 3.3 - Primeira tabela de respostas binárias construída pela AA para os dados do Exemplo A, apresentados da Tabela 2.1.

Streptococcus	Resposta Dicotômica	
pyogenes	1	0
1	19	53
0	497	829

Nota: Os níveis 0 e 1 da variável Streptococcus pyogenes correspondem a portadores e não portadores da bactéria, respectivamente.

Tabela 3.4 - Segunda tabela de respostas binárias construída pela AA para os dados do Exemplo A, apresentados da Tabela 2.1.

Streptococcus	Resposta Dicotômica	
pyogenes	1	0
1	48	24
0	1057	269

Nota: Os níveis 0 e 1 da variável Streptococcus pyogenes correspondem a portadores e não portadores da bactéria, respectivamente.

Para continuar o desenvolvimento, considere novamente o caso experimental $n_i = n$, $i = 1,2,...,I$, mencionado no início da seção.

Seja X_{isj} a variável aleatória que assume o valor 1 se, para a j-ésima tabela, a s-ésima resposta do i-ésimo nível do fator A pertence a uma categoria de resposta menor ou igual a j, para i=1,2,...,I; s=1,2,...,n e j=1,2,...,K-1. Então

$$c_{ij} = \sum_{r=1}^{j} y_{ir} = \sum_{s=1}^{n} x_{isj} \qquad \text{e} \qquad p_{ij} = \frac{c_{ij}}{n}$$

representam, respectivamente, o número e a proporção de "uns" no i-ésimo nível do fator A, na tabela j.

A proporção observada de "uns" na j-ésima tabela é definida por

$$\overline{p}_j = \frac{1}{nI} \sum_{i=1}^{I} \sum_{s=1}^{n} x_{isj} = \frac{1}{nI} \sum_{i=1}^{I} c_{ij} = \frac{\sum_{i=1}^{I} np_{ij}}{nI} = \frac{\sum_{i=1}^{I} c_{ij}}{nI} = \frac{c_{.j}}{n}.$$

A soma total de quadrados da j-ésima tabela é dada por

$$SQ_{TOT.j} = \sum_{i=1}^{I} \sum_{s=1}^{n} \left(X_{isj} - \overline{p}_j\right)^2 = \sum_{i=1}^{I} c_{ij} - 2\overline{p}_j \sum_{i=1}^{I} c_{ij} + nI\overline{p}_j^2 = \frac{Ic_{.j}\left(n - c_{.j}\right)}{n}.$$

Ainda para a j-ésima tabela, a soma de quadrados dentro dos níveis do fator A é, analogamente,

$$SQ_{DENTRO,j} = \sum_{i=1}^{I} \sum_{s=1}^{n} \left(X_{isj} - p_{ij}\right)^2 = n \sum_{i=1}^{I} \left[\frac{c_{ij}}{n} - 2p_{ij}\frac{c_{ij}}{n} + p_{ij}^2\right]$$

$$= n \sum_{i=1}^{I} p_{ij}\left(1 - p_{ij}\right) = \frac{1}{n} \sum_{i=1}^{I} c_{ij}\left(n - c_{ij}\right).$$

Por sua vez, a soma de quadrados devida ao fator A (ou soma de quadrados entre os níveis desse fator) pode ser obtida através de

$$SQ_{A,j} = SQ_{ENTRE,j} = \sum_{i=1}^{I}\sum_{s=1}^{n}\left(p_{ij} - \bar{p}_j\right)^2 = \sum_{i=1}^{I} n\left(p_{ij} - \bar{p}_j\right)^2$$

$$= \sum_{i=1}^{I} n\left[\frac{c_{ij}}{n} - \frac{c_{.j}}{n}\right]^2 = \frac{1}{n}\sum_{i=1}^{I}\left(c_{ij} - c_{.j}\right)^2 = \frac{1}{n}\sum_{i=1}^{I}\left(c_{ij}^2 - c_{.j}^2\right).$$

Portanto, para a j-ésima tabela produzida pela AA, j=1,2,...,K-1, a tabela da ANOVA está exibida na Tabela 3.5, como segue:

Tabela 3.5 - Tabela da ANOVA da j-ésima tabela construída pela Análise de Acumulação, para o caso de um único fator explanatório.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Entre, j	$I-1$	$\frac{1}{n}\sum_{i=1}^{I}\left(c_{ij}^2 - c_{.j}^2\right)$
Dentro, j	$I(n-1)$	$\frac{1}{n}\sum_{i=1}^{I} c_{ij}\left(n - c_{ij}\right)$
Total, j	$nI-1$	$\frac{1}{n} I\, c_{.j}\left(n - c_{.j}\right)$

A AA parece apresentar uma grave limitação com respeito ao modelo postulado, embora o objetivo principal não seja a modelagem dos efeitos dos fatores. O modelo linear assumido na Tabela 3.5 é

$$x_{isj} = \delta_j + \omega_{ij} + e_{isj} \qquad i = 1,2,...,I;\ \ s = 1,2,...,n$$

ou

$$c_{ij} = \xi_j + \tau_{ij} + e_{ij} \qquad i = 1,2,...,I$$

onde $\xi_j = n\delta_i$ e $\tau_{ij} = n\omega_{ij}$. Note que c_{ij}, a freqüência de observações classificadas até a j-ésima categoria pode ser escrita como $c_{ij} = \sum_{r=1}^{j} y_{ir} = \sum_{s=1}^{n} x_{isj}$. Ainda, ξ_j é o parâmetro comum (média geral) a todos os tratamentos na j-ésima tabela, τ_{ij} é o efeito do i-ésimo tratamento na j-ésima tabela e e_{ij} é a componente aleatória, com média zero e variância σ_{ij}^2.

As observações do i-ésimo tratamento são denotadas por $\underset{\sim}{y_i} = \left(y_{i1}, y_{i2}, \ldots, y_{iK}\right)^t$ e verificam $\sum_{j=1}^{K} y_{ij} = n$. Se esse vetor possui uma distribuição multinomial $M\left(n;\ \underset{\sim}{\pi_i}\right)$ para cada tratamento, onde $\underset{\sim}{\pi_i} = \left(\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK}\right)^t$, então a variável aleatória $C_{ij} = \sum_{s=1}^{n} Y_{is}$ segue uma distribuição binomial com parâmetros $\left(n;\ \gamma_{ij}\right)$, [condicional ao número de observações ou replicações fixo do i-ésimo tratamento], veja a Seção A4 do Anexo A.

Para a j-ésima tabela, seja $\tau_{ij} = \mu_{ij} - \xi_j$ o desvio do tratamento i em relação à média geral, tal que $\sum_{i=1}^{I} \tau_{ij} = 0$. Assim, o modelo

$$c_{ij} = \xi_j + \tau_{ij} + e_{ij} \qquad \forall\, i = 1,2,\ldots,I$$

com $\sum_{i=1}^{I} \tau_{ij} = 0$; $E\left[e_{ij}\right] = 0$ e $\operatorname{Var} e_{ij} = \sigma_{ij}^2 = n\gamma_{ij}\left(1 - \gamma_{ij}\right)$ é um de "efeitos fixos". A média do i-ésimo tratamento é

$$\mu_{ij} = E\left[C_{ij}\right] = \xi_j + \tau_{ij} \qquad i=1,2,\ldots,I.$$

Nesse momento pode ser constatado o primeiro problema associado ao modelo postulado pela AA. Para j fixo; isso é, na j-ésima tabela de respostas dicotômicas, $\gamma_{ij} \neq \gamma_{i'j}$ para todo $i' \neq i$. Conseqüentemente, $\operatorname{Var} e_{ij} = \sigma_{ij}^2 \neq \sigma_{i'j}^2 = \operatorname{Var} e_{i'j}$, de tal forma que na ANOVA realizada na j-ésima tabela, a hipótese de homocedasticidade dos erros não é satisfeita, devido às características da resposta binária.

A tabela da AA utilizada por Taguchi (1987) é uma composição das K-1 tabelas da ANOVA para respostas dicotômicas. Isso é, as somas de quadrados da AA são obtidas adicionando as respectivas somas de quadrados das K-1 tabelas, ponderadas pelo fator $\left[\bar{p}_j\left(1-\bar{p}_j\right)\right]^{-1} = \dfrac{n^2}{c_{.j}\left(n-c_{.j}\right)}$. Por exemplo, a soma de quadrados do efeito do fator A é obtida pela adição das respectivas somas de quadrados das K-1 tabelas, ponderadas pelo fator $\left[\bar{p}_j\left(1-\bar{p}_j\right)\right]^{-1}$. Essa ponderação é necessária pois, sob a hipótese de que o fator não tem efeito, as esperanças de $SQ_{A,j}$ são distintas: elas precisam ser ponderadas antes de serem agrupadas. Uma ponderação com pesos $\left[\bar{p}_j\left(1-\bar{p}_j\right)\right]^{-1}$ parece adequada, pois esse valor é o inverso da variância binomial da j-ésima tabela, de tal forma que as tabelas que apresentarem grande variabilidade receberão pesos reduzidos na composição das somas de quadrados finais. Assim,

$$SQ_A = \sum_{j=1}^{K-1} \frac{1}{\bar{p}_j\left(1-\bar{p}_j\right)} SQ_{A,j} = \sum_{j=1}^{K-1} \frac{n^2}{c_{.j}\left(n-c_{.j}\right)} \frac{1}{n} \sum_{i=1}^{I} \left(c_{ij}-c_{.j}\right)^2$$

$$= n \sum_{j=1}^{K-1} \sum_{i=1}^{I} \frac{\left(c_{ij}-c_{.j}\right)^2}{c_{.j}\left(n-c_{.j}\right)}.$$

De forma análoga, a soma total de quadrados resulta da adição das correspondentes somas de quadrados das K-1 tabelas, ponderadas pelo fator $\left[\bar{p}_j\left(1-\bar{p}_j\right)\right]^{-1}$, como segue:

$$SQ_{TOTAL} = \sum_{j=1}^{K-1} \frac{1}{\bar{p}_j\left(1-\bar{p}_j\right)} SQ_{TOT,j}$$

$$= \sum_{j=1}^{K-1} \frac{n^2}{c_{.j}\left(n-c_{.j}\right)} \frac{1}{n} I\, c_{.j}\left(n-c_{.j}\right) = nI(K-1).$$

Já a soma de quadrados do erro pode ser obtida pela diferença

$$SQ_{ERRO} = SQ_{TOTAL} - SQ_A$$

$$= nI(K-1) - n \sum_{j=1}^{K-1} \sum_{i=1}^{I} \frac{\left(c_{ij} - c_{.j}\right)^2}{c_{.j}\left(n - c_{.j}\right)}$$

$$= n \sum_{j=1}^{K-1} \left[I - \frac{1}{c_{.j}\left(n - c_{.j}\right)} \sum_{i=1}^{I} \left(c_{ij} - c_{.j}\right)^2 \right]$$

$$= n \sum_{j=1}^{K-1} \frac{I\, c_{.j}\left(n - c_{.j}\right) - \sum_{i=1}^{I} c_{ij}^2 + I\, c_{.j}^2}{c_{.j}\left(n - c_{.j}\right)}$$

$$= n \sum_{j=1}^{K-1} \frac{n \sum_{i=1}^{I} c_{ij} - \sum_{i=1}^{I} c_{ij}^2}{c_{.j}\left(n - c_{.j}\right)} = n \sum_{j=1}^{K-1} \sum_{i=1}^{I} \frac{c_{ij}\left(n - c_{ij}\right)}{c_{.j}\left(n - c_{.j}\right)}.$$

Tabela 3.6 - Tabela da AA para o experimento com um fator.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Fator A	$(K-1)(I-1)$	$SQ_A = n \sum_{j=1}^{K-1} \sum_{i=1}^{I} \frac{\left(c_{ij} - c_{.j}\right)^2}{c_{.j}\left(n - c_{.j}\right)}$
ERRO	$I\,(K-1)(n-1)$	$SQ_{ERRO} = n \sum_{j=1}^{K-1} \sum_{i=1}^{I} \frac{c_{ij}\left(n - c_{.j}\right)}{c_{.j}\left(n - c_{.j}\right)}$
TOTAL	$nI\,(K-1) - K + 1$	$SQ_{TOTAL} = nI\,(K-1)$

A extensão dos modelos postulados nas K-1 tabelas, para a tabela global da AA não é imediata. Embora formalmente não seja realizada a análise da variância nessa tabela, seus princípios são adotados para decidir se um determinado fator é importante. Mais uma vez constata-se que as suposições básicas da ANOVA, normalidade, independência e homocedasticidade dos erros não são satisfeitas.

Para j fixo; isso é, na j-ésima tabela de resposta binária, já foi mostrado que para todo $i \neq i'$, $\sigma_{ij}^2 \neq \sigma_{i'j}^2$. Mas essa tabela é construída pela adição das freqüências de respostas nas categorias menores ou iguais a j. Assim, para a tabela $j' \neq j$, $\gamma_{ij} \neq \gamma_{ij'}$ e, portanto, $\sigma_{ij}^2 \neq \sigma_{ij'}^2$. Esse fato compromete também a suposição de independência entre os c_{ij} exigidos pela ANOVA. No que diz respeito a normalidade, ela poderia ser obtida através da convergência da distribuição binomial para a distribuição normal. Contudo, isso depende criticamente do número de replicações dos subexperimentos, que freqüentemente não é suficiente para esse objetivo.

Como pode ser visto em Nair (1986), Taguchi sugere considerar $I(n-1)(K-1)$ e $(I-1)(K-1)$, respectivamente, como os graus de liberdade da soma de quadrados do erro e do efeito do fator A. Esses graus de liberdade parecem inapropriados, pois são obtidos através da soma dos respectivos graus de liberdade das K-1 tabelas da ANOVA definidas pela AA. Esse problema será comentado, de forma genérica, na Seção 3.1.3 que trata da técnica da AA no caso multifatorial.

Para testar a hipótese de que o fator não tem efeito, a estatística $F_A = \dfrac{QM_A}{QM_{ERRO}}$ pode ser utilizada, onde o numerador representa a quadrado médio do efeito do fator A, enquanto o denominador é o quadrado médio do erro, definidos do modo usual, veja Nair (1986) e Hamada & Wu (1990).

Para ilustrar o método, a AA será aplicada aos dados do Exemplo A.

3.1.2 EXEMPLO A

Nesse caso, são obtidos os resultados seguintes.

Primeira Tabela (j=1): $\bar{p}_j = \dfrac{1}{I}\sum_{i=1}^{I}\dfrac{c_{ij}}{n_i} = \dfrac{1}{I}\sum_{i=1}^{I} p_{ij}$

$$\bar{p}_j = \frac{1}{2}(0{,}2639 + 0{,}3748) = 0{,}3194$$

onde, $p_{11} = \frac{c_{11}}{n_1} = \frac{19}{72} = 0{,}2639$ e $p_{21} = \frac{c_{21}}{n_2} = \frac{497}{1326} = 0{,}3748$.

Assim, $\left[\bar{p}_1(1-\bar{p}_1)\right]^{-1} = 4{,}6005$

e

$$SQ_{TOT,1} = (1-\bar{p}_1)^2 \sum_{i=1}^{2} c_{i1} + \bar{p}_1^2 \left[n_1 - \sum_{i=1}^{2} c_{i1} \right] = 329{,}0050$$

$$SQ_{A,1} = \sum_{i=1}^{1}\sum_{s=1}^{n_i} (p_{i1} - \bar{p}_1)^2 = \sum_{i=1}^{2} n_i (p_{i1} - \bar{p}_1)^2 = 4{,}3002$$

$$SQ_{DENTRO,1} = \sum_{i=1}^{1}\sum_{s=1}^{n_i} (x_{is1} - p_{i1})^2 = \sum_{i=1}^{2} \left[c_{i1}(1-p_{i1})^2 + (n_i - c_{i1}) p_{i1}^2 \right] = 324{,}7048.$$

Tabela 3.7 - ANOVA para a primeira tabela de respostas binárias construída pela AA, para os dados do Exemplo A.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Entre	1	4,3002
Dentro	1396	324,7048
Total	1397	329,0050

Segunda Tabela (j=2): De modo análogo, resultam

$p_{12} = 0{,}6667$ e $p_{22} = 0{,}7971$.

$\bar{p}_2 = 0{,}7319$ e $\left[\bar{p}_2(1-\bar{p}_2)\right]^{-1} = 5{,}0963$.

Assim,

$$SQ_{TOT,2} = (1-\bar{p}_2)^2 \sum_{i=1}^{2} c_{i2} + \bar{p}_2^2 \left[n_2 - \sum_{i=1}^{2} c_{i2} \right] = 236{,}3782$$

$$SQ_{A,2} = \sum_{i=1}^{I}\sum_{s=1}^{n_i}\left(p_{i2} - \overline{p}_2\right)^2 = \sum_{i=1}^{2} n_i\left(p_{i2} - \overline{p}_2\right)^2 = 5{,}9491$$

$$SQ_{DENTRO,2} = \sum_{i=1}^{2}\left[c_{i2}\left(1 - p_{i2}\right)^2 + \left(n_i - c_{i2}\right)p_{i2}^2\right] = 230{,}4291.$$

Tabela 3.8 - ANOVA para a segunda tabela de respostas binárias construída pela AA, para os dados do Exemplo A.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Entre	1	5,9491
Dentro	1396	230,4291
Total	1397	236,3782

Para o Exemplo A, as somas de quadrados da AA, compostas pelas respectivas somas de quadrados de cada tabela, ponderadas pelo termo $\left[\overline{p}_j\left(1-\overline{p}_j\right)\right]^{-1}$. são obtidas a seguir. A Tabela 3.9 apresenta a tabela global da AA para esse exemplo.

$$SQ_{TOTAL} = \sum_{j=1}^{K-1} \frac{1}{\overline{p}_j\left(1-\overline{p}_j\right)} SQ_{TOT,j}$$

$$= 4{,}6005\ 329{,}0050 + 5{,}0963\ 236{,}3782 = 2718{,}2475.$$

$$SQ_A = SQ_{ENTRE} = \sum_{j=1}^{K-1} \frac{1}{\overline{p}_j\left(1-\overline{p}_j\right)} SQ_{A,j}$$

$$= 4{,}6005\ 4{,}3002 + 5{,}0963\ 5{,}9491 = 50{,}1014.$$

$$SQ_{ERRO} = SQ_{DENTRO} = \sum_{j=1}^{K-1} \frac{1}{\overline{p}_j\left(1-\overline{p}_j\right)} SQ_{DENTRO,j}$$

$$= 4{,}6005\ 324{,}7048 + 5{,}0963\ 230{,}4291 = 2668{,}1461.$$

Tabela 3.9 - Tabela da AA para os dados do Exemplo A.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
Entre	2	50,1014
Dentro	2792	2668,1461
Total	2794	2718,2475

É importante perceber que o número de observações é diferente para cada linha da tabela; ou seja, para portadores e não portadores de *Streptococcus pyogenes*. Nessa situação, não é possível determinar o número de graus de liberdade, conforme sugestão de Taguchi apresentada na Tabela 3.6. No entanto, seguindo os mesmos princípios, os graus de liberdade associados as somas de quadrados da tabela da AA são obtidos através da adição dos respectivos graus de liberdade das duas tabelas da ANOVA para respostas binárias. Sob a hipótese de que o *Streptococcus pyogenes* não afeta o tamanho relativo das amígdalas, o valor observado da estatística $F_A = 26,21$ é comparado com a distribuição de probabilidade F de Snedecor-Fisher, com 2 graus de liberdade no numerador e 2792 no denominador. O nível de significância atingido foi $2,87 \times 10^{-7}$, sugerindo que o *Streptococcus pyogenes* provoca um impacto no tamanho relativo das amígdalas.

Observe, contudo, que o número total de graus de liberdade nesse exemplo é 2794, enquanto que existem apenas 1398 observações. Alguns problemas relacionados com os graus de liberdade da AA e com a estatística de teste F_A serão abordados nas próximas seções.

3.1.3 CASO MULTIFATORIAL

Para estender a AA para o caso multifator, pode-se considerar um experimento usando um delineamento fatorial com R subexperimentos. Cada subexperimento consiste de n observações da resposta, segundo as K categorias ordenadas.

Seja y_{ir} a freqüência da r-ésima categoria para o i-ésimo subexperimento. A extensão da AA na situação de múltiplos fatores consiste em realizar K-1 análises de variância (multiway ANOVA, em inglês) sobre os fatores. De forma semelhante ao caso de um fator, os dados da j-ésima tabela da ANOVA (j=1,2,...,K-1) consistem de $\sum_{r=1}^{j} y_{ir}$ "uns" e $n - \sum_{r=1}^{j} y_{ir}$ "zeros', para o i-ésimo subexperimento.

Por exemplo, para um delineamento fatorial fracionado do tipo 2^{3-1}, as observações podem ser representadas como na Tabela 3.10, a qual corresponde à fração principal gerada pela relação definitória I=ABC para a estrutura de confundimento.

Tabela 3.10 - Frequências nas categtorias de respostas para o delineamento fatorial do tipo fracionado 2^{3-1}.

Subexpe-	Fator			Categoria de resposta					Total
rimento	A	B	C	1	2	3	...	K	
1	+	+	+	y_{11}	y_{12}	y_{13}	...	y_{1K}	n
2	+	-	-	y_{21}	y_{22}	y_{23}	...	y_{2K}	n
3	-	+	-	y_{31}	y_{32}	y_{33}	...	y_{3K}	n
4	-	-	+	y_{41}	y_{42}	y_{43}	...	y_{4K}	n

Seja $c_{ij} = \sum_{r=1}^{j} y_{ir}$, para todo i=1,2,...,R e j=1,2,...,K-1, onde $R = 2^{3-1}$ é o número de subexperimentos. Então, de maneira semelhante ao caso de um fator, a AA utiliza K-1 tabelas de contingência do tipo Rx2, sobre as quais serão realizadas as ANOVAS para dados binários. A j-ésima tabela é semelhante a Tabela 3.2, como pode ser observado na Tabela 3.11.

Tabela 3.11 - Número de respostas observadas até a j-ésima categoria (inclusive) no delineamento 2^{3-1}.

Subexperimento	Fator			Resposta Dicotômica	
	A	B	C	1	0
1 (abc)	+	+	+	c_{1j}	$n - c_{1j}$
2 (a)	+	-	-	c_{2j}	$n - c_{2j}$
3 (b)	-	+	-	c_{3j}	$n - c_{3j}$
4 (c)	-	-	+	c_{4j}	$n - c_{4j}$

A j-ésima tabela da ANOVA para essa situação é idêntica a Tabela 3.5, onde R=I. Adotando a notação de Yates usada por Box, Hunter and Hunter (1978) para os delineamentos fatoriais, obtém-se as somas de quadrados dos respectivos fatores. Então, a, b, c e abc representam as combinações dos níveis dos fatores (subexperimentos) e também os respectivos totais entre as n replicações desses subexperimentos. Por exemplo, na Tabela 3.11 abc representa o subexperimento 1, definido pela combinação do nível + dos fatores A, B e C, mas representa também o número total de respostas (c_{1j}) observadas até a categoria j (inclusive), dentre as n replicações.

Defina a variável aleatória X_{isj}, que assume o valor 1, se para a j-ésima tabela, a s-ésima resposta do i-ésimo subexperimento pertence a uma categoria menor ou igual a j (para i=1,2,...,I; s=1,2,...,n e j=1,2,...,K-1) e o valor 0 em caso contrário. Assim, podem ser obtidos os totais dos respectivos subexperimentos, como segue:

$$abc = \sum_{s=1}^{n} X_{1sj} = \sum_{r=1}^{j} y_{1r} = c_{1j}$$

$$a = \sum_{s=1}^{n} X_{2sj} = \sum_{r=1}^{j} y_{2r} = c_{2j}$$

$$b = \sum_{s=1}^{n} X_{3sj} = \sum_{r=1}^{j} y_{3r} = c_{3j}$$

$$c = \sum_{s=1}^{n} X_{4sj} = \sum_{r=1}^{j} y_{4r} = c_{4j}..$$

Na maioria dos experimentos, o objetivo é examinar magnitude e direção dos efeitos dos fatores para determinar sua possível influência. Para a estimação dos efeitos é usual a utilização de contrastes. Por exemplo, na j-ésima tabela construída pela AA, o efeito do fator A é estimado por

$$A,j = \bar{y}_{A^+,j} - \bar{y}_{A^-,j} = \frac{abc + a}{2n} - \frac{b + c}{2n}$$

$$= \frac{1}{2n}(abc + a - b - c) = \frac{1}{2n}\left(c_{1j} + c_{2j} - c_{3j} - c_{4j}\right).$$

As interpretações sobre a magnitude desses efeitos geralmente podem ser confirmadas através da análise da variância. É necessário, assim, determinar as somas de quadrados associadas a esses efeitos, as quais podem ser obtidas através dos respectivos contrastes.

A soma de quadrados de um contraste $C = \sum_i c_i y_i$ é dada por

$$SQ_C = \frac{\left[\sum_{i=1}^{l} c_i y_i\right]^2}{n \sum_{i=1}^{l} c_i^2},$$

veja Montgomery (1991, p.69).

Portanto, para o delineamento fatorial fracionado do tipo 2^{3-1}, as somas de quadrados dos efeitos principais são

$$SQ_{A,j} = \frac{\left(c_{1j} + c_{2j} - c_{3j} - c_{4j}\right)^2}{4\,n}$$

$$SQ_{B,j} = \frac{\left(c_{1j} - c_{2j} + c_{3j} - c_{4j}\right)^2}{4\,n}$$

e

$$SQ_{C,j} = \frac{\left(c_{1j} - c_{2j} - c_{3j} + c_{4j}\right)^2}{4\,n}.$$

A análise da variância para a j-ésima tabela produzida pela AA pode ser efetuada com o auxílio da Tabela 3.12, mostrada abaixo.

Tabela 3.12 - Análise da Variância para a j-ésima tabela da AA, para o delineamento fatorial fracionado do tipo 2^{3-1} (I=ABC).

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
A.j	1	$SQ_{A,j} = \frac{\left(c_{1j} + c_{2j} - c_{3j} - c_{4j}\right)^2}{4\,n}$
B.j	1	$SQ_{B,j} = \frac{\left(c_{1j} - c_{2j} + c_{3j} - c_{4j}\right)^2}{4\,n}$
C.j	1	$SQ_{C,j} = \frac{\left(c_{1j} - c_{2j} - c_{3j} + c_{4j}\right)^2}{4\,n}$
ERRO.j	4(n-1)	$SQ_{ERRO,j} = SQ_{TOT,j} - SQ_{A,j} - SQ_{B,j} - SQ_{C,j}$
TOTAL.j	4n - 1	$SQ_{TOT,j} = \frac{4}{n} c_{.j}\left(n - c_{.j}\right)$

Assim, as somas de quadrados da tabela da AA no caso multifatorial são obtidas de forma análoga ao caso de um fator, através da adição das somas de quadrados de cada uma das K-1 tabelas, com ponderações $\left[\overline{p}_j\left(1-\overline{p}_j\right)\right]^{-1}$. Para o delineamento fatorial do tipo 2^{3-1}, esses resultados são desenvolvidos na Seção A1 do Anexo A e apresentados na Tabela 3.13. Observe que essa tabela é artificial; isso é, os valores de I=2 e R=4 não foram substituídos para que a estrutura dos graus de liberdade da AA possa ser usada posteriormente.

Os graus de liberdade associados às somas de quadrados são inapropriados, pois foram obtidos através da soma dos respectivos graus de liberdade das K-1 tabelas da ANOVA para respostas binárias definidas pela AA. Observe que nesse experimento existem 4 subexperimentos com n replicações, perfazendo um total de 4n observações. No entanto, o

número de graus de liberdade associados a soma total de quadrados na tabela da AA é $nR(K-1)-K+1=4n(K-1)-K+1$. Esse problema também é verificado nas demais somas de quadrados da AA, não parecendo lógico existir um número de graus de liberdade maior do que a quantidade global de observações. Aparentemente, por esses motivos os "testes informais" sugeridos por Taguchi (1987) conduzem, muito freqüentemente, a decisões incorretas sobre a significância dos fatores.

Tabela 3.13 - Tabela da AA para o delineamento fatorial fracionado do tipo 2^{3-1}.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
A	(I-1)(K-1)	$SQ_A=\frac{n}{4}\sum_{j=1}^{K-1}\frac{\left(c_{1j}+c_{2j}-c_{3j}-c_{4j}\right)^2}{c_j\left(n-c_j\right)}$
B	(I-1)(K-1)	$SQ_B=\frac{n}{4}\sum_{j=1}^{K-1}\frac{\left(c_{1j}-c_{2j}+c_{3j}-c_{4j}\right)^2}{c_j\left(n-c_j\right)}$
C	(I-1)(K-1)	$SQ_C=\frac{n}{4}\sum_{j=1}^{K-1}\frac{\left(c_{1j}-c_{2j}-c_{3j}+c_{4j}\right)^2}{c_j\left(n-c_j\right)}$
ERRO	nR(K-1)-K+1- -3(I-1)(K-1)	$SQ_{ERRO}=SQ_{TOTAL}-SQ_A-SQ_B-SQ_C$
TOTAL	nR(K-1)-K+1	$SQ_{TOTAL}=4n(K-1)$

Nota: I=2 é o número de níveis de cada fator; R=4 é o número de subexperimentos; K é o número de categorias de resposta e n é o número de replicações em cada subexperimento.

Uma explicação para a "multiplicação" dos graus de liberdade não foi encontrada na literatura. Nair (1986), observa que a inadequacidade dessas quantidades não representa uma deficiência séria, pois o objetivo principal da AA não é a realização de testes formais de significância, mas sim a descrição da importância dos fatores.

No caso multifatorial, outro problema da AA, possivelmente mais grave, está relacionado à própria natureza dos dados. Na ANOVA para dados categóricos ordenados, a ortogonalidade do delineamento não assegura a independência das estatísticas de teste para diferentes fatores. Na AA, por sua vez, não existe independência entre as somas de quadrados. Isso pode ser verificado, por exemplo, no delineamento fatorial do tipo 2^{3-1}, cuja soma total de

quadrados é igual à constante $SQ_{TOTAL} = 4n(K-1)$. Então, a soma de quadrados do erro depende dos efeitos dos demais fatores e, conseqüentemente, a estatística $F_A = \frac{QM_A}{QM_{ERRO}}$ utilizada para testar a hipótese de que o efeito do fator A é nulo não é adequada, pois sua distribuição depende dos parâmetros correspondentes aos fatores B e C, veja Hamada & Wu (1990).

Segundo Nair (1986), sob a hipótese nula de que não existe efeito dos fatores, o valor esperado do QM_{ERRO} é aproximadamente igual a 1, de tal forma que não existe nenhuma vantagem no seu uso. Por outro lado, quando um ou mais fatores possuem efeitos consideráveis, o QM_{ERRO} pode ser substancialmente menor do que 1, inflacionando o valor da estatística F.

Para ilustrar o método da AA no caso multifatorial, será novamente considerado o Exemplo B.

3.1.4 EXEMPLO B

Os dados do Exemplo B podem ser analisados com a metodologia da AA. Para tanto, serão usados os dados apresentados na Tabela 2.6, correspondentes aos resultados do experimento segundo o critério de discretização discutido na Seção 2.4.

O plano ortogonal L16 adotado por Quinlan nesse experimento, coincide essencialmente com o plano fatorial fracionado do tipo 2_{III}^{15-11}; sua identidade pode ser exibida mediante a mudança do sinal de algumas colunas da matriz do delineamento, veja Box, Bisgaard and Fung (1988).

A AA para esse experimento gera K-1=2 tabelas de respostas dicotômicas, para as quais devem ser conduzidas as análises da variância. Assim, para a j-ésima tabela os totais resultantes da execução dos subexperimentos são definidos por $c_{1j}, c_{2j}, \ldots, c_{16j}$ e correspondem a primeira coluna de resposta dicotômica. Conseqüentemente, para a primeira tabela construída pela AA, foram obtidas as respostas binárias apresentadas na Tabela 3.14, que incorpora a correspondência entre colunas detalhada por Box, Bisgaard and Fung (1988).

Tabela 3.14 - Freqüências de respostas para a primeira tabela de respostas binárias construída pela AA, no experimento para melhorar a qualidade das camisas termoplásticas.

Subexpe-	Fator															Resposta Dicotômica	
rimento	H	D	-L	B	-J	-F	N	A	-I	-E	M	-C	K	G	-O	1	0
1	-	-	+	-	+	+	-	-	+	+	-	+	-	-	+	0	4
2	+	-	-	-	-	+	+	-	-	+	+	+	+	-	-	0	4
3	-	+	-	-	+	-	+	-	+	-	+	+	-	+	-	4	0
4	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	4	0
5	-	-	+	+	-	-	+	-	+	+	-	-	+	+	-	1	3
6	+	-	-	+	+	-	-	-	-	+	+	-	-	+	+	4	0
7	-	+	-	+	-	+	-	-	+	-	+	-	+	-	+	1	3
8	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	4	0
9	-	-	+	-	+	+	-	+	-	-	+	-	+	+	-	4	0
10	+	-	-	-	-	+	+	+	+	-	-	-	-	+	+	4	0
11	-	+	-	-	+	-	+	+	-	+	-	-	+	-	+	0	4
12	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-	0	4
13	-	-	+	+	-	-	+	+	-	-	+	+	-	-	+	3	1
14	+	-	-	+	+	-	-	+	+	-	-	+	+	-	-	0	4
15	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	0	4
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	4

As somas de quadrados para a primeira tabela de respostas binárias construída pela AA são obtidas de forma semelhante ao caso do delineamento fatorial fracionado do tipo 2^{3-1} apresentado acima. Como pode ser observado, nesse experimento foram realizadas n=4 replicações em cada um dos R=16 subexperimentos. Portanto,

$$SQ_{TOT,1} = \frac{R\, c_{.1}(n - c_{.1})}{n} = 15{,}8594$$

com Rn-1 graus de liberdade, onde $c_{.1} = \frac{1}{R}\sum_{i=1}^{R} c_{i1} = \frac{29}{16} = 1{,}8125$.

As somas de quadrados dos efeitos principais dos fatores podem ser obtidas através dos contrastes. Assim, a ANOVA para a primeira tabela de respostas dicotômicas construída pela AA é apresentada na Tabela 3.15. Por sua vez, a soma de quadrados do erro é obtida pela diferença

$$SQ_{ERRO} = 15{,}8594 - 13{,}6094 = 2{,}2500.$$

Tabela 3.15 - ANOVA para a primeira tabela de respostas binárias construída pela AA para os dados da Tabela 2.6, relativos ao experimento para melhorar a qualidade das camisas termoplásticas.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
A.1	1	0.7656
B.1	1	0.1406
-C.1	1	0.7656
D.1	1	0.1406
-E.1	1	5.6406
-F.1	1	0.1406
G.1	1	2.6406
H.1	1	0.1406
-I.1	1	1.2656
-J.1	1	0.1406
K.1	1	1.2656
-L.1	1	0.1406
M.1	1	0.1406
N.1	1	0.1406
-O.1	1	0.1406
ERRO.1	48	2.2500
TOT.1	63	15.8594

De forma análoga obtém-se a ANOVA para a segunda tabela de respostas dicotômicas, apresentada na Tabela 3.16.

Tabela 3.16 - Freqüências de respostas para a segunda tabela de respostas binárias construída pela AA, no experimento para melhorar a qualidade das camisas termoplásticas.

Subexpe-	Fator															Resposta Dicotômica	
rimento	H	D	-L	B	-J	-F	N	A	-I	-E	M	-C	K	G	-O	1	0
1	-	-	+	-	+	+	-	-	+	+	-	+	-	-	+	0	4
2	+	-	-	-	-	+	+	-	-	+	+	+	+	-	-	0	4
3	-	+	-	-	+	-	+	-	+	-	+	+	-	+	-	4	0
4	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	4	0
5	-	-	+	+	-	-	+	-	+	+	-	-	+	+	-	4	0
6	+	-	-	+	+	-	-	-	-	+	+	-	-	+	+	4	0
7	-	+	-	+	-	+	-	-	+	-	+	-	+	-	+	4	0
8	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	4	0
9	-	-	+	-	+	+	-	+	-	-	+	-	+	+	-	4	0
10	+	-	-	-	-	+	+	+	+	-	-	-	-	+	+	4	0
11	-	+	-	-	+	-	+	+	-	+	-	-	+	-	+	0	4
12	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-	0	4
13	-	-	+	+	-	-	+	+	-	-	+	+	-	-	+	4	0
14	+	-	-	+	+	-	-	+	+	-	-	+	+	-	-	4	0
15	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	3	1
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0	4

Tabela 3.17 - ANOVA para a segunda tabela de respostas binárias construída pela AA para os dados da Tabela 2.6, relativos ao experimento para melhorar a qualidade das camisas termoplásticas.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
A.2	1	0.3906
B.2	1	1.8906
-C.2	1	0.3906
D.2	1	0.3906
-E.2	1	6.8906
-F.2	1	0.3906
G.2	1	1.8906
H.2	1	0.1406
-I.2	1	0.1406
-J.2	1	0.1406
K.2	1	0.1406
-L.2	1	0.1406
M.2	1	0.1406
N.2	1	0.1406
-O.2	1	0.1406
ERRO.2	48	0.7500
TOT.2	63	14.1094

Portanto, as somas de quadrados da AA, definidas pela adição das respectivas somas de quadrados das K-1 tabelas de respostas binárias, ponderadas pelo fator $\frac{n^2}{c_{.j}\left(n - c_{.j}\right)}$, são:

$$SQ_{TOTAL} = Rn(K-1) = 16\ 4\ (3-1) = 128$$

$$SQ_{ERRO} = \sum_{j=1}^{K-1} \frac{n^2}{c_{.j}\left(n - c_{.j}\right)} SQ_{ERRO.j} = \frac{16}{3{,}9648}\ 2{,}25 + \frac{16}{3{,}5273}\ 0{,}75 = 12{,}4818$$

e

$$SQ_A = \sum_{j=1}^{K-1} \frac{n^2}{c_{.j}\left(n - c_{.j}\right)} SQ_{A.j} = \frac{16}{3{,}9648}\ 0{,}7656 + \frac{16}{3{,}5273}\ 0{,}3906 = 4{,}8615.$$

As somas de quadrados dos efeitos principais dos demais fatores são obtidos de forma idêntica. A AA para esse experimento é apresentada na Tabela 3.18.

Como mencionado anteriormente, Taguchi sugere o uso da estatística $F_A = \frac{QM_A}{QM_{ERRO}}$ para testar a hipótese de ausência de impacto para o fator A. Sob essa hipótese, a distribuição de probabilidade usada como referência nesses testes é a F de Snedecor-Fisher com 2 graus de liberdade no numerador e 96 no denominador. O quadrado médio do erro da AA para os resultados desse experimento assume o valor 0,1300.

Tabela 3.18 - Tabela da AA para o experimento para melhorar a qualidade das camisas termoplásticas.

Fonte de Variação	Graus de Liberdade	Soma de Quadrados
A	2	4.8615
B	2	9.1433
-C	2	4.8615
D	2	2.3394
-E	2	54.0184
-F	2	2.3394
G	2	19.2320
H	2	1.2054
-I	2	5.7453
-J	2	1.2054
K	2	5.7453
-L	2	1.2054
M	2	1.2054
N	2	1.2054
-O	2	1.2054
ERRO	96	12.4818
TOTAL	126	128

A estatística de teste F_A apresenta sérios problemas, sendo completamente inadequada. Contudo, com o intuito de determinar os fatores que a AA detecta como possivelmente responsáveis pela contração das camisas, os testes serão realizados no sentido informal. Para todos os fatores, o nível de significância atingido foi inferior a 1%. Assim, adotando esse critério, conclui-se que os 15 fatores envolvidos no experimento possivelmente produzem um impacto significante na contração das camisas termoplásticas.

No caso do Exemplo B, então, a AA não parece um método eficiente para determinar os fatores que provocam maior impacto no grau de contração das camisas. A principal causa disso é a própria estatística de teste, que possui um denominador inferior a 1, fazendo com que os valores observados de F sejam grandes.

Para estudar melhor o comportamento da técnica da AA no que diz respeito aos testes de significância, proponho estimar os efeitos dos fatores seguindo os princípios da AA, construindo posteriormente o gráfico probabilístico normal dos efeitos. É adequado salientar que apesar de sua simplicidade, essa observação não se encontra na literatura disponível.

Para tanto, a estimação dos efeitos será efetuada de maneira similar ao procedimento usado para determinar as somas de quadrados. Assim, no experimento para melhorar a qualidade das camisas termoplásticas, a estimativa do efeito do fator A, por exemplo, é obtida como segue: na j-ésima tabela de resposta binária, o efeito do fator A é

$$A,j = \frac{1}{2n}\left(-c_{1j} - c_{2j} - \cdots - c_{8j} + c_{9j} + c_{10j} + \cdots + c_{16j}\right).$$

Assim,

$$A,1 = \frac{1}{8}(-18 + 11) = -0{,}8750$$

e

$$A,2 = \frac{1}{8}(-24 + 19) = -0{,}6250.$$

Então, o efeito do fator A é estimado mediante a soma dos efeitos em cada tabela, ponderados por $\sqrt{\frac{n^2}{c_{.j}\left(n - c_{.j}\right)}}$, o que resulta em

$$A = -0{,}8750\ \sqrt{4{,}0355}\ -\ 0{,}6250\ \sqrt{4{,}5360} = -3{,}0889.$$

As estimativas assim obtidas são apresentadas na Tabela 3.19. Através do gráfico probabilístico normal para esses efeitos constata-se que apenas os fatores E e G parecem possuir impacto significativo no grau de contração das camisas termoplásticas.

Tabela 3.19 - Estimativas do efeito dos fatores no experimento para melhorar a qualidade das camisas termoplásticas, segundo os princípios da AA.

Fator	Estimativa do efeito pela AA		
	1ª. tabela	2ª. tabela	Global
A	-0,875	-0,625	-3,0889
B	-0,375	1,375	2,1751
-C	-0,875	-0,625	-3,0889
D	-0,375	-0,625	-2,0844
-E	-2,375	-2,625	-10,3617
-F	-0,375	-0,625	-2,0844
G	1,625	1,375	6,1928
H	0,375	-0,375	-0,0453
-I	-1,125	-0,375	-3,0586
-J	0,375	-0,375	-0,0453
K	-1,125	-0,375	-3,0586
-L	0,375	-0,375	-0,0453
M	0,375	-0,375	-0,0453
N	0,375	-0,375	-0,0453
-O	0,375	-0,375	-0,0453

Cabe observar que mesmo adotando um critério duvidoso para obter o efeito global dos fatores, o gráfico probabilístico normal aparenta ser mais eficiente do que a própria AA, na verificação do caráter ativo de fatores experimentais. Esse fato parece ilustrar, mais uma vez, a inadequacidade do método da AA para a análise de experimentos com resposta ordinal. Na próxima seção serão brevemente apresentadas algumas sugestões recentes para melhorar o desempenho dessa técnica.

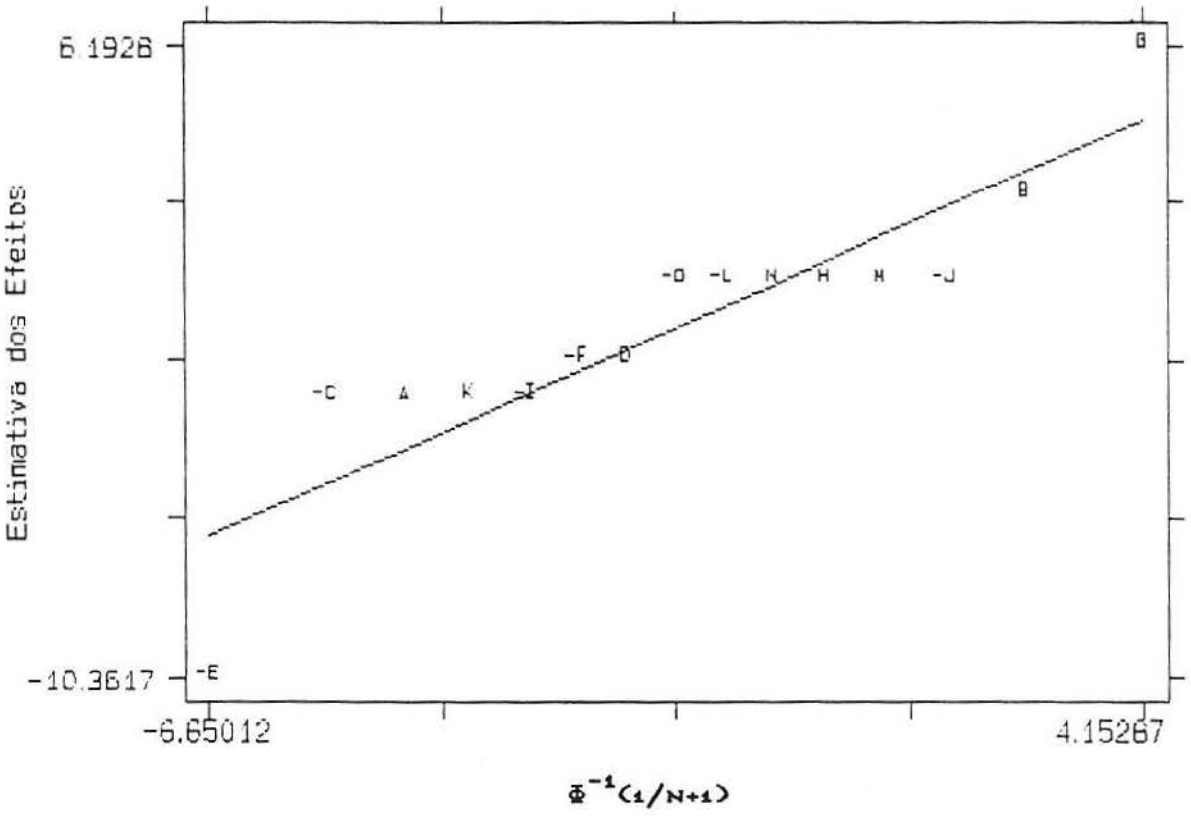

Figura 3.1 - Gráfico probabilístico normal dos efeitos principais dos fatores, estimados segundo os princípios da AA para os dados do Exemplo B.

3.2 MODIFICAÇÕES

Na seção anterior foram brevemente abordados os problemas que a AA apresenta quanto ao modelo postulado, os graus de liberdade das estatísticas e os testes de significância construídos com elas. Em resumo, constatou-se que as suposições básicas da ANOVA não são satisfeitas, que o número de graus de liberdade é maior do que o total de observações e que a estatística F parece inadequada para testar a significância dos fatores.

Para contornar esses problemas da AA, Nair (1986) sugere uma modificação na estatística F_A, no sentido de utilizar apenas a soma de quadrados SQ_A para testar a significância dos fatores. No caso multifatorial, essa estatística modificada, denotada por T_A, elimina uma das formas pelas quais F_A depende dos outros fatores. Isso pode ser percebido, por exemplo, no delineamento fatorial do tipo 2^{3-1}, onde a distribuição de F_A depende dos fatores B e C, pois não existe independência entre as somas de quadrados.

Considerando o caso - descrito anteriormente - de um único fator, a estatística T é

$$T_A = n \sum_{j=1}^{K-1} \frac{\left(c_{ij} - c_{.j}\right)^2}{c_{.j}\left(n - c_{.j}\right)}.$$

Supondo que os totais das colunas da Tabela 3.1 são fixos, seja $\underset{\sim}{R}_i$ a matriz condicional de covariâncias do vetor $\underset{\sim}{y}_i = \left(y_{i1}, y_{i2}, \ldots, y_{iK}\right)^t$. Então, sob a hipótese de que o fator A é inativo, as matrizes $\underset{\sim}{R}_1, \underset{\sim}{R}_2, \ldots, \underset{\sim}{R}_I$ são identicamente iguais a $\underset{\sim}{R}$.

A estatística T pode ser definida pela forma quadrática $T = \sum_{i=1}^{I} \underset{\sim}{y}_i^t \underset{\sim}{A}^t \underset{\sim}{A} \underset{\sim}{y}_i$, para $\underset{\sim}{A}$ quadrada. Por sua vez, a matriz $\underset{\sim}{A}^t \underset{\sim}{A}$ pode ser decomposta como $\underset{\sim}{A}^t \underset{\sim}{A} = \underset{\sim}{Q}\,\underset{\sim}{L}\,\underset{\sim}{Q}^t$, onde $\underset{\sim}{L}$ é uma matriz diagonal de dimensão (K-1)x(K-1) dos autovalores de $\underset{\sim}{A}^t \underset{\sim}{A}$ e $Q = \left(\underset{\sim}{q}_1, \underset{\sim}{q}_2, \ldots, \underset{\sim}{q}_{K-1}\right)$

é uma matriz de ordem Kx(K-1) cujas colunas são os autovalores associados. Além disso, a matriz particionada $\underset{\sim}{\overline{Q}} = \left[\underset{\sim}{1} \mid \underset{\sim}{Q}\right]$ tem a propriedade que $\underset{\sim}{\overline{Q}}^t \underset{\sim}{R}\, \underset{\sim}{\overline{Q}}$ é proporcional a matriz identidade.

Seja $Z_j^2 = \sum_{i=1}^{I} \left(\underset{\sim}{q}_j^t \underset{\sim}{y}_i \right)^2$ e sejam $\lambda_1, \lambda_2, \ldots, \lambda_{K-1}$ os autovalores de $\underset{\sim}{A}^t \underset{\sim}{A}$. Então,

$$T = \sum_{i=1}^{I} \underset{\sim}{y}_i^t \underset{\sim}{A}^t \underset{\sim}{A}\, \underset{\sim}{y}_i = \sum_{i=1}^{I} \left(\underset{\sim}{y}_i^t \underset{\sim}{Q}\, \underset{\sim}{L}\, \underset{\sim}{Q}^t \underset{\sim}{y}_i \right)$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{K-1} \lambda_j \left[\underset{\sim}{q}_j^t \underset{\sim}{y}_i \right]^2 = \sum_{j=1}^{K-1} \lambda_j Z_j^2 .$$

Observa-se assim que a estatística modificada da AA foi decomposta em K-1 componentes. Ela é uma soma ponderada, onde os pesos são os autovalores da matriz $\underset{\sim}{A}^t \underset{\sim}{A}$. As duas primeiras componentes, quais sejam, Z_1^2 e Z_2^2 são usadas para testar efeitos de posição e de dispersão, respectivamente. Entretanto, os pesos λ_j decrescem rapidamente com j, o que leva a pensar que a AA fornece fundamentalmente um teste de posição. Nair (1986) sugere a utilização separada de Z_1^2 e Z_2^2 para aumentar o poder de detecção de efeitos de posição e dispersão.

Um procedimento alternativo para detectar efeitos de posição e dispersão é o uso de dois conjuntos de escores $\underset{\sim}{1}$ e $\underset{\sim}{d}$, os quais dependem apenas das freqüências marginais das categorias ordenadas. A partir desses conjuntos, determina-se as somas de quadrados de posição e de dispersão, respectivamente denotadas por $SQ_A\left(\underset{\sim}{1}\right)$ e $SQ_A\left(\underset{\sim}{d}\right)$, onde o índice diz respeito ao fator A considerado. Assim,

$$SQ_A\left(\underset{\sim}{1}\right) = \frac{1}{n} \sum_{i=1}^{I} \left[\underset{\sim}{1}^t \left(\underset{\sim}{y}_i - \underset{\sim}{y}_. \right) \right]^2$$

e

$$SQ_A\left(\underset{\sim}{d}\right) = \frac{1}{n}\sum_{i=1}^{I}\left[\underset{\sim}{d}^t\left(\underset{\sim}{y}_i - \underset{\sim}{y}_.\right)\right]^2,$$

onde $\underset{\sim}{y}_. = \frac{1}{I}\sum_{i=1}^{I}\underset{\sim}{y}_i$.

Pode ser mostrado que, sob a hipótese nula, as somas de quadrados $SQ_A\left(\underset{\sim}{l}\right)$ e $SQ_A\left(\underset{\sim}{d}\right)$ são aproximadamente distribuídas segundo a distribuição de probabilidades χ^2_{I-1}, onde I denota o número de níveis do fator A, veja Nair (1986).

As modificações propostas por Nair (1986), também discutidas por Hamada & Wu (1986) e Hamada & Wu (1990), constituem um avanço na melhoria da técnica da AA. Contudo, os testes de posição e de dispersão definidos a partir da decomposição da estatística modificada da AA não são o tema central da presente dissertação. Eles foram apresentados com o intuito de informar a direção em que estão seguindo as pesquisas para o refinamento da técnica.

No próximo capítulo será apresentado outro método para análise de experimentos com resposta ordinal, o qual permite estimar a magnitude e direção dos efeitos dos fatores em estudo.

CAPÍTULO 4: MODELO DE McCULLAGH

Nos capítulos precedentes constatou-se a necessidade de modelos que permitam analisar e interpretar as relações entre uma resposta ordinal e variáveis explicativas. Nesse capítulo será discutido o *modelo de odds proporcionais* proposto por McCullagh (1980), que faz parte de uma classe de modelos de regressão para analisar a dependência entre uma variável categórica ordenada e um conjunto de covariáveis.

A principal motivação para os modelos propostos é a possibilidade de existir uma resposta variável, latente e contínua, usualmente não observável. Em outras palavras, os dados observados são uma categorização de uma variável aleatória contínua subjacente. Um caso conhecido é o dos bioensaios, onde a variável latente corresponde ao nível de tolerância de uma droga, sobre o qual se assume uma distribuição contínua na população. O nível de tolerância não pode ser observado diretamente, mas sua elevação é manifestada através do crescimento da probabilidade de sobrevivência.

A existência de uma variável latente não é fundamental para a validade do modelo, mas se ela existe, a interpretação dos parâmetros se torna clara e direta, veja McCullagh (1980).

Antes de apresentar uma formulação do modelo proposto por P. McCullagh, é conveniente fazer algumas observações sobre a estrutura desse capítulo.

A Seção 4.1 contém o desenvolvimento formal para o modelo de odds proporcionais. Inicialmente são considerados os aspectos fundamentais para a compreensão do método, bem como o enfoque dado por alguns pesquisadores. Esses conceitos e definições são vitais e, assim, serão utilizados ao longo da dissertação. Na seqüência, a Seção 4.1.1 tem um caráter mais técnico. A construção passo a passo do modelo é útil para os leitores interessados nos detalhes matemáticos do desenvolvimento.

Com respeito à Seção 4.1.2, que trata do método de estimação de parâmetros, valem as mesmas considerações. Inicialmente, é importante que o leitor entenda as idéias

intuitivas do processo iterativo de estimação, para depois, se necessário, poder desenvolver o formalismo, consultando os anexos e/ou as referências citadas.

Por sua vez, a Seção 4.1.3 é de importância, pois trata de alguns problemas ligados à estimação de parâmetros que possivelmente acontecem na prática.

Uma discussão sobre os recursos computacionais disponíveis para o ajuste de modelos de odds prorporcionais é apresentada na Seção 4.2. São mencionadas as principais características dos pacotes disponíveis e algumas dificuldades encontradas.

Por fim, na Seção 4.3 são explorados dois exemplos com dados reais, aos quais foram ajustados modelos de odds proporcionais. O primeiro, na Seção 4.3.1, corresponde aos dados do Exemplo A. O processo de ajuste foi detalhado passo a passo, incluindo o procedimento iterativo de estimação de parâmetros, para ilustrar a construção do modelo. Os resultados são comparados com aqueles obtidos através dos pacotes estatísticos disponíveis. A Tabela 4.4 exibe um resumo dos resultados das técnicas de análise estatística aplicadas a esses dados. Os dados do Exemplo B são analisados na Seção 4.3.2, também mediante o ajuste de um modelo de odds proporcionais. Os resultados são comparados com aqueles produzidos pela técnica da AA e pelos métodos estatísticos empregados por outros analistas (veja a Tabela 4.9).

Na continuação, portanto, será apresentado o desenvolvimento formal do modelo, com observações sobre a estimação de parâmetros e aplicações a exemplos.

4.1 MODELO DE ODDS PROPORCIONAIS

Seguindo a formalização apresentada por Hastie & Tibshirani (1987) e Hastie, Botha and Schnitzler (1989), seja Z a variável latente, com função de distribuição $F_\eta(z) = F(z-\eta)$, onde η é um parâmetro de posição. Seja $\underset{\sim}{x}$ o vetor de covariáveis e $\eta\left(\underset{\sim}{x}\right) = \underset{\sim}{\beta}^t \underset{\sim}{x}$. Assim, a distribuição condicional de $Z \mid \underset{\sim}{x}$ é dada por $F\left(z - \underset{\sim}{\beta}^t \underset{\sim}{x}\right)$. A variável contínua subjacente Z não pode ser observada, mas regiões de valores de Z são conhecidas através da categorização em intervalos reais $(-\infty,\theta_1]$, $(\theta_1,\theta 2],\ldots,(\theta_{k-1},+\infty)$, onde os θ_j são

parâmetros desconhecidos. Isso induz a variável aleatória observada Y, que assume os valores $Y = j$ se e somente se $Z \in (\theta_{j-1}, \theta_j]$. Assim, as probabilidades acumuladas são definidas por

$$\gamma_j(\underset{\sim}{x}) = P[Y \leq j \mid \underset{\sim}{x}] = F(\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}).$$

Qualquer distribuição unimodal simétrica pode ser postulada para a variável latente Z, produzindo geralmente resultados similares. Contudo, a suposição de uma distribuição logística $F(z) = \frac{e^Z}{1+e^Z}$ tem se mostrado bastante adequada, principalmente pela facilidade de cálculo. Na existência de razões que levem a acreditar que a distribuição subjacente é assimétrica, as funções de ligação loglog ou complementar loglog podem ser usadas, veja Hastie, Botha and Schnitzler (1989).

É importante observar que são necessárias suposições apenas sobre a distribuição condicional da variável categórica ordenada, dado o conjunto de covariáveis.

Seja Y a variável dependente cujas categorias de resposta ordenadas são denotadas por 1, 2,..., k e $\underset{\sim}{x} = (x_1, x_2, \ldots, x_p)^t$ o vetor coluna p-dimensional de covariáveis. Sejam também $\pi_1(\underset{\sim}{x}), \pi_2(\underset{\sim}{x}), \ldots, \pi_k(\underset{\sim}{x})$ as probabilidades das k categorias de resposta ordenadas, quando o vetor de covariáveis assume o valor $\underset{\sim}{x}$; ou seja,

$$P[Y = j \mid \underset{\sim}{x}] = \pi_j(\underset{\sim}{x}), \quad \forall j = 1, 2, \ldots, k.$$

Então,

$$\gamma_j(\underset{\sim}{x}) = P[Y \leq j \mid \underset{\sim}{x}] = \pi_1(\underset{\sim}{x}) + \pi_2(\underset{\sim}{x}) + \ldots + \pi_j(\underset{\sim}{x}).$$

Nesse ponto é conveniente introduzir uma definição. Se B é um evento aleatório, um jogador que atribua probabilidade P(B) estará disposto a apostar em favor do resultado B de maneira que os dois ganhos possíveis estejam na razão $\frac{P(B)}{1-P(B)}$, denominada quociente de

disparidades (odds ratio, em inglês). Nessa dissertação será utilizada a expressão simplificada odds para referir o quociente $\frac{P(B)}{1-P(B)}$. Assim, o odds do evento $\left[Y \leq j \mid \underset{\sim}{x} \right]$ é definido pela razão $\frac{\gamma_j\left(\underset{\sim}{x}\right)}{1-\gamma_j\left(\underset{\sim}{x}\right)}$, para todo $1 \leq j < k$.

No modelo de odds proporcionais introduzido por McCullagh (1980), a distribuição da variável contínua subjacente é assumida ser a distribuição logística; ou seja, $F(z) = \frac{e^z}{1+e^z}$. Assim, o modelo especifica que

$$\frac{\gamma_j\left(\underset{\sim}{x}\right)}{1-\gamma_j\left(\underset{\sim}{x}\right)} = \frac{F\left(\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right)}{1-F\left(\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right)} = \frac{\dfrac{\exp\left\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right\}}{1+\exp\left\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right\}}}{1-\dfrac{\exp\left\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right\}}{1+\exp\left\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right\}}} = \exp\left\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right\} \qquad (4.1)$$

para todo $j = 1, 2, \ldots, k-1$, onde os parâmetros desconhecidos θ_j satisfazem $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{k-1}$, $\theta_0 \equiv -\infty$ e $\theta_k \equiv +\infty$. O vetor de parâmetros $\underset{\sim}{\beta} = \left[\beta_1, \beta_2, \ldots, \beta_p\right]^t$ representa os coeficientes de regressão a serem estimados, enquanto que o vetor $\underset{\sim}{\theta} = \left[\theta_1, \theta_2, \ldots, \theta_{k-1}\right]^t$ contém os pontos de corte desconhecidos.

Esse modelo pode ser escrito de uma forma equivalente, como

$$P\left[Y \leq j \mid \underset{\sim}{x} \right] = \frac{\exp\left\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right\}}{1+\exp\left\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right\}}, \qquad \forall\ 1 \leq j < k.$$

Para dois vetores de covariáveis $\underset{\sim}{x}_1$ e $\underset{\sim}{x}_2$ distintos, o quociente

$$\frac{\dfrac{\gamma_j(\underset{\sim}{x}_1)}{1-\gamma_j(\underset{\sim}{x}_1)}}{\dfrac{\gamma_j(\underset{\sim}{x}_2)}{1-\gamma_j(\underset{\sim}{x}_2)}} = \frac{\exp\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}_1\}}{\exp\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}_2\}} = \exp\{\underset{\sim}{\beta}^t(\underset{\sim}{x}_2 - \underset{\sim}{x}_1)\}$$

assume um valor fixo que independe da categoria j. Ele pode ser interpretado como a chance relativa de ter um escore menor ou igual a j, entre dois indivíduos com diferentes valores das covariáveis.

O modelo (4.1) pode ser escrito, ainda, como

$$\text{logit}\{\gamma_j(\underset{\sim}{x})\} = \log\frac{\gamma_j(\underset{\sim}{x})}{1-\gamma_j(\underset{\sim}{x})} = \theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\ , \qquad 1 \le j < k\,.$$

Quando existem apenas duas categorias de resposta, esse modelo é equivalente ao modelo logístico linear para dados binários proposto por Cox (1970). Nesse caso particular, o modelo também corresponde à estrutura log-linear, veja McCullagh (1980).

Anderson & Philips (1981) apresentam essas idéias na forma de um modelo geral, definido por

$$P[Y \le j \mid \underset{\sim}{x}] = \psi(\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}) \qquad 0 \le j \le k\,, \tag{4.2}$$

onde $\psi(.)$ é uma função de distribuição completamente especificada. O modelo da equação (4.2) é um modelo linear generalizado, mas versões não lineares também podem ser obtidas.

Outras suposições sobre a distribuição condicional de $Z \mid \underset{\sim}{x}$ são possíveis. Por exemplo, a escolha da distribuição Normal $N\left(\underset{\sim}{\beta}^t \underset{\sim}{x},\ 1\right)$ conduz ao modelo de probitos

$$\gamma_j\left(\underset{\sim}{x}\right) = P\left[Y \le j \mid \underset{\sim}{x}\right] = \phi\left(\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\right)$$

e, portanto, a

$$\phi^{-1}\left(\gamma_j\left(\underset{\sim}{x}\right)\right) = \theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}\ ,$$

onde $\phi(.)$ denota a função de distribuição da normal padrão.

A seguir será abordada a situação em que os indivíduos da população de interesse possuem apenas um dos k atributos $A_1, A_2, \ldots, A_k$. Se uma amostra aleatória com n indivíduos for extraída com reposição, então o número de indivíduos observados que possuem o atributo A_k é dado pela distribuição multinomial

$$P\left[Y_1 = y_1,\ Y_2 = y_2, \ldots,\ Y_k = y_k\right] = \frac{n!}{y_1!\, y_2! \ldots y_k!}\, \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_k^{y_k} \qquad (4.3)$$

$$\text{com} \quad \sum_{j=1}^{k} y_j = n \quad \text{e} \quad \sum_{j=1}^{k} \pi_j = 1 \quad .$$

Nesse caso, a distribuição multinomial é conseqüência do método de amostragem e pode ser obtida como uma aproximação para a distribuição correspondente à amostragem sem reposição de populações grandes.

Uma outra maneira de se obter uma distribuição multinomial é a seguinte: sejam $Y_1, Y_2, \ldots, Y_k$ variáveis aleatórias independentes com distribuição de Poisson com médias $\mu_1, \mu_2, \ldots, \mu_k$, respectivamente. Então, a distribuição conjunta condicional de $Y_1, Y_2, \ldots, Y_k$ dado $\sum_{j=1}^{k} Y_j = m$ é dada pela equação (4.3), com $\pi_j = \frac{\mu_j}{\sum_{j=1}^{n} \mu_j}$.

O desenvolvimento do modelo de odds proporcionais não aparece na literatura, exceto alguns resultados importantes apresentados por McCullagh & Nelder (1989) e Hastie &

Tibshirani (1987). Assim, com o intuito de facilitar sua compreensão, a derivação do modelo será detalhada.

4.1.1 DERIVAÇÃO

Considere uma variável resposta Y com k categorias ordenadas, denotadas por 1, 2,..., k. Seja $\underset{\sim}{x} = \left(x_1, x_2, \ldots, x_p\right)^t$ o vetor coluna p-dimensional constituído pelos regressores ou covariáveis. As combinações entre os diferentes níveis ou categorias dos regressores definem subpopulações distintas. Assim, $\pi_1\left(\underset{\sim}{x}\right), \pi_2\left(\underset{\sim}{x}\right), \ldots, \pi_k\left(\underset{\sim}{x}\right)$ são as probabilidades de resposta nas categorias 1, 2,..., k, respectivamente, quando o vetor de covariáveis assume o valor $\underset{\sim}{x}$. Portanto, as probabilidades de resposta das respectivas categorias, para os distintos valores possíveis do vetor $\underset{\sim}{x}$, quais sejam, $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n$, podem ser dispostas como na Tabela 4.1 abaixo.

Tabela 4.1 - Probabilidades de resposta nas categorias da variável Y, para os distintos valores do vetor de covariáveis.

Valor de $\underset{\sim}{x}$	Categorias de Resposta			
	1	2	...	k
$\underset{\sim}{x}_1$	$\pi_1\left(\underset{\sim}{x}_1\right)$	$\pi_2\left(\underset{\sim}{x}_1\right)$	...	$\pi_k\left(\underset{\sim}{x}_1\right)$
$\underset{\sim}{x}_2$	$\pi_1\left(\underset{\sim}{x}_2\right)$	$\pi_2\left(\underset{\sim}{x}_2\right)$		$\pi_2\left(\underset{\sim}{x}_2\right)$
⋮	⋮	⋮		⋮
$\underset{\sim}{x}_n$	$\pi_1\left(\underset{\sim}{x}_n\right)$	$\pi_2\left(\underset{\sim}{x}_n\right)$	...	$\pi_2\left(\underset{\sim}{x}_n\right)$

Assim, as freqüências de resposta nas diferentes categorias, dentre as n subpopulações, podem ser apresentadas na Tabela 4.2, como segue:

Tabela 4.2 - Freqüências de resposta nas categorias da variável Y, para os distintos valores do vetor de covariáveis $\underset{\sim}{x}$.

Valor de $\underset{\sim}{x}$	Categorias de Resposta			
	1	2	...	k
$\underset{\sim}{x}_1$	y_{11}	y_{12}	...	y_{1k}
$\underset{\sim}{x}_2$	y_{21}	y_{22}	...	y_{2k}
⋮	⋮	⋮		⋮
$\underset{\sim}{x}_n$	y_{n1}	y_{n2}	...	y_{nk}

Por simplicidade de notação, seja $\pi_j\left(\underset{\sim}{x}_i\right) = \pi_{ij}$, para todo $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, k$. Esse abuso de notação é justificado, pois não existe possibilidade de confusão entre os dois termos, uma vez que as subpopulações são determinadas pelo condicionamento da variável resposta nos diferentes valores assumidos pelo vetor $\underset{\sim}{x}$.

Assim, para $i = 1, 2, \ldots, n$ sejam $\underset{\sim}{y}_i = \left(y_{i1}, y_{i2}, \ldots, y_{ik}\right)^t$ e $\underset{\sim}{\pi}_i = \left(\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik}\right)^t$ os vetores-coluna k-dimensionais que contém as freqüências observadas e as probabilidades de resposta da i-ésima subpopulação.

Supondo que os vetores $\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_n$ são independentes, com distribuição multinomial; isto é,

$$P\left[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{ik} = y_{ik}\right] = \frac{n_i!}{y_{i1}!\, y_{i2}! \ldots y_{ik}!} \prod_{j=1}^{k} \left(\pi_{ij}^{y_{ij}}\right)$$

com as restrições $\sum_{j=1}^{k} y_{ij} = n_i$ e $\sum_{j=1}^{k} \pi_{ij} = 1$, a função de verossimilhança é dada por

$$L\left(\underset{\sim}{\pi}\,;\underset{\sim}{y}\right) = \prod_{i=1}^{n} \frac{n_i!}{y_{i1}!\, y_2! \ldots y_{ik}!} \prod_{j=1}^{k} \left(\pi_{ij}^{y_{ij}}\right),$$

onde $\underset{\sim}{\pi} = \left(\underset{\sim}{\pi}_1,\ \underset{\sim}{\pi}_2,\ldots,\ \underset{\sim}{\pi}_n\right)^t$ e $\underset{\sim}{y} = \left(\underset{\sim}{y}_1,\ \underset{\sim}{y}_2,\ldots,\ \underset{\sim}{y}_n\right)^t$ possuem dimensão $nk \times 1$.

A contribuição que o i-ésimo vator de freqüências observadas para o logaritmo da função de verossimilhança é

$$l\left(\underset{\sim}{\pi}_i\ ;\ \underset{\sim}{y}_i\right) = \sum_{j=1}^{k} y_{ij} \log\, \pi_{ij} = \underset{\sim}{y}_i^t \log\ \underset{\sim}{\pi}_i$$

com as restrições $\underset{\sim}{y}_i^t\, \underset{\sim}{1}_k = n_i$ e $\underset{\sim}{\pi}_i^t\, \underset{\sim}{1}_k = 1$ ou $\sum_{j=1}^{k} y_{ij} = n_i$ e $\sum_{j=1}^{k} \pi_{ij} = 1$, para todo $i = 1, 2,\ldots, n$.

A matriz de variâncias e covariâncias da i-ésima subpopulação é definida por $\Sigma_i = n_i\left[\text{diag}\left\{\underset{\sim}{\pi}_i\right\} - \underset{\sim}{\pi}_i\, \underset{\sim}{\pi}_i^t\right]$, com dimensão $k \times k$ e de posto $k-1$. Uma inversa generalizada simples é definida por $\Sigma_i^{-} = \text{diag}\left\{\frac{1}{n_i\, \underset{\sim}{\pi}_i}\right\}$, veja McCullagh & Nelder (1989, p.168).

O modelo de odds proporcionais especifica que

$$\text{logit}\left\{\gamma_j\left(\underset{\sim}{x}_i\right)\right\} = \theta_j - \underset{\sim}{\beta}^t\, \underset{\sim}{x}_i$$

para todo $1 \le j < k$ e $1 \le i < n$; onde $\underset{\sim}{x}_i = \left[x_{i1},\ x_{i2},\ldots,\ x_{ip}\right]^t$ representa o valor que o vetor de covariáveis $\underset{\sim}{x}$ assume na i-ésima subpopulação. Assim, o modelo possui $p^* = p + k - 1$ parâmetros a serem estimados, os quais constituem os vetores $\underset{\sim}{\theta} = \left[\theta_1,\ \theta_2,\ldots,\ \theta_{k-1}\right]^t$ e

$\underset{\sim}{\beta} = \left[\beta_1, \beta_2, \ldots, \beta_p\right]^t$. Seja, também, $\underset{\sim}{\beta}^* = \left[\theta_1, \theta_2 \ldots, \theta_{k-1}, \beta_1, \beta_2, \ldots, \beta_p\right]^t$ o vetor dos $p^* = p + k - 1$ parâmetros do modelo. Para tanto, é necessário derivar o núcleo da função log-verossimilhança $l\left(\underset{\sim}{\pi} ; \underset{\sim}{y}\right)$ com respeito ao vetor de parâmetros $\underset{\sim}{\beta}^*$.

Como existe independência entre os vetores de observações $\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_n$, esse núcleo é definido como a soma das n contribuições de cada vetor observado; ou seja,

$$l\left(\underset{\sim}{\pi} ; \underset{\sim}{y}\right) = \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log \pi_{ij}$$

sob as restrições $\sum_{j=1}^{k} y_{ij} = n_i$ e $\sum_{j=1}^{k} \pi_{ij} = 1$, para todo $i = 1, 2, \ldots, n$. Assim, podem ser tomadas as derivadas do núcleo da função log-verossimilhança de cada subpopulação $l\left(\underset{\sim}{\pi} ; \underset{\sim}{y}_i\right)$, somando-as posteriormente.

As expressões dessas derivadas são complexas e, por esse motivo, são obtidas na Seção A2 do Anexo A.

Na i-ésima subpopulação, $i = 1, 2, \ldots, n$, seja $\underset{\sim}{\gamma}_i = \underset{\sim}{L}\,\underset{\sim}{\pi}_i = \left[\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{ik-1}\right]^t$ o vetor de probabilidades acumuladas, onde $\underset{\sim}{L}$ é uma matriz triangular inferior de ordem $(k-1) \times k$, cujos elementos subdiagonais são iguais a 1. Uma inversa generalizada de $\underset{\sim}{L}$ é definida por

$$\underset{\sim}{L}^{-} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{bmatrix} \qquad (4.4)$$

Seja $\underset{\sim}{\eta}_i = \underset{\sim}{\theta} - \left[\underset{\sim}{\beta}^t \underset{\sim}{x}_i \right] \underset{\sim}{1}_{k-1}$ o vetor de dimensão $(k-1)\times 1$, o qual contém os valores ajustados para a i-ésima subpopulação. Então, para todo $i = 1, 2, \ldots, n$, a derivada do núcleo da função log-verossimilhança, com respeito aos parâmetros do modelo, é definida por

$$\frac{\partial l\left(\underset{\sim}{\pi}_i \,;\, \underset{\sim}{y}_i\right)}{\partial \underset{\sim}{\beta}^*} = \frac{\partial \underset{\sim}{\eta}_i}{\partial \underset{\sim}{\beta}^*} \frac{\partial \underset{\sim}{\gamma}_i}{\partial \underset{\sim}{\eta}_i} \frac{\partial l\left(\underset{\sim}{\pi}_i \,;\, \underset{\sim}{y}_i\right)}{\partial \underset{\sim}{\gamma}_i} = \underset{\sim}{D}_i^t \underset{\sim}{u}_i$$

onde $\underset{\sim}{u}_i$ e $\underset{\sim}{D}_i$ são definidos pelas equações (A2.1) e (A2.3), respectivamente, veja a Seção A2 do Anexo A para maiores detalhes.

Agora, seja $\underset{\sim}{\gamma} = \underset{\sim}{L}^* \underset{\sim}{\pi}$ o vetor de probabilidades acumuladas para todas as subpopulações. A matriz $\underset{\sim}{L}^* = \underset{n}{I} \otimes \underset{\sim}{L}$, onde $\otimes$ denota o produto de Kronecker, transforma o vetor $\underset{\sim}{\pi}$ no vetor de dimensão $n(k-1)\times 1$ que contém as probabilidades acumuladas. O vetor $\underset{\sim}{\gamma}$ também pode ser escrito como $\underset{\sim}{\gamma} = \left[\underset{\sim}{\gamma}_1, \underset{\sim}{\gamma}_2, \ldots, \underset{\sim}{\gamma}_n \right]^t$, onde os $\underset{\sim}{\gamma}_i$ são definidos pela equação $\underset{\sim}{\gamma}_i = \underset{\sim}{L}\,\underset{\sim}{\pi}_i$. Uma inversa generalizada da matriz $\underset{\sim}{L}^*$ é dada por $\underset{\sim}{L}^{*-} = \underset{n}{I} \otimes \underset{\sim}{L}^-$, onde $\underset{\sim}{L}^-$ definida em (4.4) é a inversa generalizada da matriz $\underset{\sim}{L}$.

Como conseqüência da independência dos vetores de observações $\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_n$, a matriz de covariâncias do vetor $\underset{\sim}{y}$ é definida como uma matriz diagonal de blocos $\Sigma = \text{bdiag}\{\Sigma_1, \Sigma_2, \ldots, \Sigma_n\}$, sendo Σ_i a matriz de covariâncias da i-ésima subpopulação. Por sua vez, uma inversa generalizada de Σ é $\Sigma^- = \text{bdiag}\{\Sigma_1^-, \Sigma_2^-, \ldots, \Sigma_n^-\}$, onde $\Sigma_i^- = \text{diag}\left\{ \frac{1}{n_i \pi_{i1}}, \frac{1}{n_i \pi_{i2}}, \ldots, \frac{1}{n_i \pi_{ik}} \right\}$, para $i = 1, 2, \ldots, n$.

As equações de verossimilhança para os parâmetros do modelo são definidas pelas derivadas parciais $\dfrac{\partial l\left(\underset{\sim}{\pi}\,;\,\underset{\sim}{y}\right)}{\partial \underset{\sim}{\beta}^{*}} = \underset{\sim}{0}$. Pela regra da cadeia,

$$\frac{\partial l\left(\underset{\sim}{\pi}\,;\,\underset{\sim}{y}\right)}{\partial \underset{\sim}{\beta}^{*}} = \frac{\partial \underset{\sim}{\eta}}{\partial \underset{\sim}{\beta}^{*}}\,\frac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\eta}}\,\frac{\partial l\left(\underset{\sim}{\pi}\,;\,\underset{\sim}{y}\right)}{\partial \underset{\sim}{\gamma}} \tag{4.5}$$

onde $\underset{\sim}{\eta}$ é o vetor de dimensão $n(k-1)\times 1$, cujos elementos são os valores ajustados. Então, $\underset{\sim}{\eta} = \left[\underset{\sim}{\eta}_1,\ \underset{\sim}{\eta}_2,\ldots,\ \underset{\sim}{\eta}_n\right]^t$, de tal forma que

$$\frac{\partial \underset{\sim}{\eta}}{\partial \underset{\sim}{\beta}^{*}} = \begin{bmatrix} \dfrac{\partial \underset{\sim}{\eta}_1}{\partial \theta_1} & \dfrac{\partial \underset{\sim}{\eta}_2}{\partial \theta_1} & \cdots & \dfrac{\partial \underset{\sim}{\eta}_n}{\partial \theta_1} \\ \dfrac{\partial \underset{\sim}{\eta}_1}{\partial \theta_2} & \dfrac{\partial \underset{\sim}{\eta}_2}{\partial \theta_2} & \cdots & \dfrac{\partial \underset{\sim}{\eta}_n}{\partial \theta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial \underset{\sim}{\eta}_1}{\partial \beta_p} & \dfrac{\partial \underset{\sim}{\eta}_2}{\partial \beta_p} & \cdots & \dfrac{\partial \underset{\sim}{\eta}_n}{\partial \beta_p} \end{bmatrix}_{p^{*}\times n(k-1)}$$

$$= \left[\begin{array}{cccc:cccc:c:cccc} 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 & & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 & & 0 & 0 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -x_{11} & -x_{11} & \cdots & -x_{11} & -x_{21} & -x_{21} & \cdots & -x_{21} & & -x_{n1} & -x_{n1} & \cdots & -x_{n1} \\ -x_{12} & -x_{12} & \cdots & -x_{12} & -x_{22} & -x_{22} & \cdots & -x_{22} & & -x_{n2} & -x_{n2} & \cdots & -x_{n2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ -x_{1p} & -x_{1p} & \cdots & -x_{1p} & -x_{2p} & -x_{2p} & \cdots & -x_{2p} & & -x_{np} & -x_{np} & \cdots & -x_{np} \end{array}\right]$$

Assim, para $\underset{\sim}{X}_i = \dfrac{\partial \underset{\sim}{\eta}_i}{\partial \underset{\sim}{\beta}^*}$ definida em (A2.2), seja $\underset{\sim}{X}^* = \dfrac{\partial \underset{\sim}{\eta}}{\partial \underset{\sim}{\beta}^*}$, a qual pode ser escrita como

$$\underset{\sim}{X}^* = \frac{\partial \underset{\sim}{\eta}}{\partial \underset{\sim}{\beta}^*} = \left[\underset{\sim}{X}_1 \mid \underset{\sim}{X}_2 \mid \cdots \mid \underset{\sim}{X}_n \right]. \tag{4.6}$$

Por sua vez, $\underset{\sim}{C} = \dfrac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\eta}}$ é

$$\underset{\sim}{C} = \frac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\eta}} = \begin{bmatrix} \dfrac{\partial \underset{\sim}{\gamma}_1}{\partial \underset{\sim}{\eta}_1} & \dfrac{\partial \underset{\sim}{\gamma}_2}{\partial \underset{\sim}{\eta}_1} & \cdots & \dfrac{\partial \underset{\sim}{\gamma}_n}{\partial \underset{\sim}{\eta}_1} \\ \dfrac{\partial \underset{\sim}{\gamma}_1}{\partial \underset{\sim}{\eta}_2} & \dfrac{\partial \underset{\sim}{\gamma}_2}{\partial \underset{\sim}{\eta}_2} & \cdots & \dfrac{\partial \underset{\sim}{\gamma}_n}{\partial \underset{\sim}{\eta}_2} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial \underset{\sim}{\gamma}_1}{\partial \underset{\sim}{\eta}_p} & \dfrac{\partial \underset{\sim}{\gamma}_2}{\partial \underset{\sim}{\eta}_p} & \cdots & \dfrac{\partial \underset{\sim}{\gamma}_n}{\partial \underset{\sim}{\eta}_p} \end{bmatrix}$$

ou seja,

$$\underset{\sim}{C} = \text{bdiag}\left\{ \underset{\sim}{C}_1, \ \underset{\sim}{C}_2, \ldots, \ \underset{\sim}{C}_n \right\} \tag{4.7}$$

onde $\underset{\sim}{C}_i = \dfrac{\partial \underset{\sim}{\gamma}_i}{\partial \underset{\sim}{\eta}_i} = \text{diag}\{\gamma_{i1}(1-\gamma_{i1}),\ \gamma_{i2}(1-\gamma_{i2}),\ldots,\ \gamma_{ik-1}(1-\gamma_{ik-1}),\}$, para todo $i = 1, 2,\ldots, n$.

Cabe ressaltar que para todo $i \neq i'$, a matriz de dimensão $(k-1)\times(k-1)$ $\dfrac{\partial \underset{\sim}{\gamma}_i}{\partial \underset{\sim}{\eta}_{i'}}$ é nula, pois

$\dfrac{\partial \gamma_{ij}}{\partial \eta_{i'j'}} = 0$ para todo $i \neq i'$ ou para $j \neq j'$.

Seja $\underset{\sim}{N}$ a matriz de ordem $nk \times nk$, definida por $\underset{\sim}{N} = \text{bdiag}\left\{\underset{\sim}{N}_1,\ \underset{\sim}{N}_2,\ldots,\ \underset{\sim}{N}_n\right\}$,

onde $\underset{\sim}{N}_i$ é a matriz diagonal de dimensão $k \times k$ que contém o número de observações da i-ésima subpopulação na diagonal principal; ou seja,

$$\underset{\sim}{N}_i = \begin{bmatrix} n_i & 0 & \cdots & 0 \\ 0 & n_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_i \end{bmatrix} \quad \text{e} \quad \underset{\sim}{N} = \begin{bmatrix} \underset{\sim}{N}_1 & 0 & \cdots & 0 \\ 0 & \underset{\sim}{N}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \underset{\sim}{N}_n \end{bmatrix}.$$

Então, o vetor $\underset{\sim}{u} = \dfrac{\partial l\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)}{\partial \underset{\sim}{\gamma}}$ é definido por

$$\underset{n(k-1)\times 1}{\underset{\sim}{u}} = \underset{\sim}{L}^{*-t}\ \underset{\sim}{N}\ \underset{\sim}{\Sigma}^{-}\left[\underset{\sim}{y} - \underset{\sim}{N}\ \underset{\sim}{\pi}\right] = \begin{bmatrix} \underset{\sim}{u}_1 \\ \underset{\sim}{u}_2 \\ \vdots \\ \underset{\sim}{u}_n \end{bmatrix} \tag{4.8}$$

onde $\underset{\sim}{u}_i$ está definido pela equação (A2.1), para todo $i = 1, 2,\ldots, n$.

Definindo $\underset{\sim}{D} = \underset{\sim}{C}\,\underset{\sim}{X}^{*t}$, a equação (4.5) pode ser reescrita como

$$\frac{\partial l\left(\underset{\sim}{\pi}\,;\,\underset{\sim}{y}\right)}{\partial\,\underset{\sim}{\beta}^{*}} = \underset{\sim}{D}^{t}\,\underset{\sim}{u}\ ;\ \text{ou seja,}$$

$$\underset{p^{*}xl}{\frac{\partial l\left(\underset{\sim}{\pi}\,;\,\underset{\sim}{y}\right)}{\underset{\sim}{\beta}^{*}}} = \left[\underset{\sim}{C}\,\underset{\sim}{X}^{*t}\right]^{t}\underset{\sim}{u} = \underset{\sim}{X}^{*}\,\underset{\sim}{C}^{t}\,\underset{\sim}{u} = \underset{\sim}{D}^{t}\,\underset{\sim}{u} \tag{4.9}$$

$$= \left[\underset{\sim}{X}_1 \mid \underset{\sim}{X}_2 \mid \cdots \mid \underset{\sim}{X}_n\right] \text{bdiag}\left\{\underset{\sim}{C}_1,\ \underset{\sim}{C}_2,\ldots,\ \underset{\sim}{C}_n\right\}\underset{\sim}{u}$$

$$= \left[\underset{\sim}{X}_1\,\underset{\sim}{C}_1 \mid \underset{\sim}{X}_2\,\underset{\sim}{C}_2 \mid \cdots \mid \underset{\sim}{X}_n\,\underset{\sim}{C}_n\right]\begin{bmatrix}\underset{\sim}{u}_1\\ \underset{\sim}{u}_2\\ \vdots\\ \underset{\sim}{u}_n\end{bmatrix}$$

$$= \underset{\sim}{X}_1\,\underset{\sim}{C}_1\,\underset{\sim}{u}_1 + \underset{\sim}{X}_2\,\underset{\sim}{C}_2\,\underset{\sim}{u}_2 + \ \ldots\ + \underset{\sim}{X}_n\,\underset{\sim}{C}_n\,\underset{\sim}{u}_n$$

$$\frac{\partial l\left(\underset{\sim}{\pi}\,;\,\underset{\sim}{y}\right)}{\partial\,\underset{\sim}{\beta}^{*}} = \sum_{i=1}^{n}\underset{\sim}{X}_i\,\underset{\sim}{C}_i\,\underset{\sim}{u}_i = \sum_{i=1}^{n}\underset{\sim}{D}_i^{t}\,\underset{\sim}{u}_i\ . \tag{4.10}$$

Portanto, o vetor de derivadas parciais da função log-verossimilhança, com relação aos parâmetros do modelo, é

$$\frac{\partial l\left(\underset{\sim}{\pi}\,;\,\underset{\sim}{y}\right)}{\partial\,\underset{\sim}{\beta}^{*}} = \sum_{i=1}^{n}\underset{\sim}{D}_i^{t}\,\underset{\sim}{u}_i\ ;$$

ou seja,

$$\frac{\partial l\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)}{\partial \underset{\sim}{\beta}^{*}} = \begin{bmatrix} -\sum_{i=1}^{n}\gamma_{i1}\left(1-\gamma_{i1}\right)\left[\dfrac{y_{i2}-n_i\pi_{i2}}{\pi_{i2}} - \dfrac{y_{i1}-n_i\pi_{i1}}{\pi_{i1}}\right] \\ -\sum_{i=1}^{n}\gamma_{i2}\left(1-\gamma_{i2}\right)\left[\dfrac{y_{i3}-n_i\pi_{i3}}{\pi_{i3}} - \dfrac{y_{i2}-n_i\pi_{i2}}{\pi_{i2}}\right] \\ \vdots \\ -\sum_{i=1}^{n}\gamma_{i\,k-1}\left(1-\gamma_{i\,k-1}\right)\left[\dfrac{y_{ik}-n_i\pi_{ik}}{\pi_{ik}} - \dfrac{y_{i\,k-1}-n_i\pi_{i\,k-1}}{\pi_{i\,k-1}}\right] \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sum_{i=1}^{n}\sum_{j=1}^{k-1}x_{i1}\gamma_{ij}\left(1-\gamma_{ij}\right)\left[\dfrac{y_{i\,j+1}-n_i\pi_{i\,j+1}}{\pi_{i\,j+1}} - \dfrac{y_{ij}-n_i\pi_{ij}}{\pi_{ij}}\right] \\ \sum_{i=1}^{n}\sum_{j=1}^{k-1}x_{i2}\gamma_{ij}\left(1-\gamma_{ij}\right)\left[\dfrac{y_{i\,j+1}-n_i\pi_{i\,j+1}}{\pi_{i\,j+1}} - \dfrac{y_{ij}-n_i\pi_{ij}}{\pi_{ij}}\right] \\ \vdots \\ \sum_{i=1}^{n}\sum_{j=1}^{k-1}x_{ip}\gamma_{ij}\left(1-\gamma_{ij}\right)\left[\dfrac{y_{i\,j+1}-n_i\pi_{i\,j+1}}{\pi_{i\,j+1}} - \dfrac{y_{ij}-n_i\pi_{ij}}{\pi_{ij}}\right] \end{bmatrix}.$$

As equações de verossimilhança são obtidas fazendo $\dfrac{\partial l\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)}{\partial \underset{\sim}{\beta}^{*}} = \underset{p^{*}\times 1}{\underset{\sim}{0}}$. Essas equações são funções não lineares de $\underset{\sim}{\beta}^{*}$ e só podem ser resolvidas através de um processo iterativo. Para tanto, nesta situação pode ser adotado o método de mínimos quadrados iterativamente reponderados (iterative reweighted least squares ou IRLS, em inglês), que será discutido a seguir.

4.1.2 ESTIMAÇÃO DOS PARÂMETROS

O procedimento de mínimos quadrados iterativamente reponderados pode ser derivado do método de "scoring" de Fisher. Ele é discutido em diversos textos da literatura, dentre os quais destacam-se Green (1984), Aitkin et al. (1989), Dobson (1983), Agresti (1990) e McCullagh & Nelder (1989).

O princípio do método IRLS, consiste em realizar sucessivos ajustes de um modelo linear, através da regressão ponderada de uma variável dependente ajustada, nas covariáveis. A variável dependente ajustada pode ser interpretada como uma "variável auxiliar", a qual possibilita sair de um sistema de equações como funções não lineares dos parâmetros, para um sistema de equações como funções lineares de $\underset{\sim}{\beta}^*$. Em outras palavras, em cada etapa do processo iterativo busca-se resolver um novo sistema de equações, definido por

$$\underset{\sim}{\beta}^{*(s+1)} = \left[\underset{\sim}{X}^* \underset{\sim}{W}^{(s)} \underset{\sim}{X}^{*t} \right]^{-1} \underset{\sim}{X}^* \underset{\sim}{W}^{(s)} \underset{\sim}{z}^{(s)} .$$

Esse sistema define as equações normais correspondentes ao ajuste de um modelo linear, via mínimos quadrados ponderados, sendo $\underset{\sim}{X}^*$ a matriz do modelo, $\underset{\sim}{z}$ o vetor de variáveis dependentes (ajustadas) e $\underset{\sim}{W}$ a matriz de pesos. Assim, o que de fato está sendo feito é uma regressão linear (nos parâmetros) de $\underset{\sim}{z}$ em $\underset{\sim}{X}^*$, ponderada pela matriz $\underset{\sim}{W}$.

A taxa de convergência da solução $\underset{\sim}{\hat{\beta}}^*$ para $\underset{\sim}{\beta}^*$ depende da matriz de informação, definida por

$$\underset{\sim}{I} = E\left[-\frac{\partial^2 l\left(\underset{\sim}{\pi} ; \underset{\sim}{y} \right)}{\partial \underset{\sim}{\beta}^* \, \partial \underset{\sim}{\beta}^{*t}} \right] = E\left[\frac{\partial l\left(\underset{\sim}{\pi} ; \underset{\sim}{y} \right)}{\partial \underset{\sim}{\beta}^*} \left(\frac{\partial l\left(\underset{\sim}{\pi} ; \underset{\sim}{y} \right)}{\partial \underset{\sim}{\beta}^*} \right)^t \right] . \qquad (4.11)$$

Na Seção A3 do Anexo A é demonstrado que, para o modelo de odds proporcionais, a matriz de informação pode ser estimada por

$$\underset{\sim}{I} = \underset{\sim}{D}^t \underset{\sim}{A}\,\underset{\sim}{D} = \underset{\sim}{D}^t \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\underset{\sim}{\Sigma}^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-}\,\underset{\sim}{D} = \sum_{i=1}^{n} \underset{\sim}{D}_i^t\,\underset{\sim}{A}_i\,\underset{\sim}{D}_i \,. \tag{4.12}$$

No modelo em questão, o vetor $\underset{\sim}{z}$ de dimensão $n(k-1)\times 1$, está definido como

$$\underset{\sim}{z} = \underset{\sim}{\eta} + \left(\frac{\partial\,\underset{\sim}{\eta}}{\partial\,\underset{\sim}{\mu}}\right)^t \left[\underset{\sim}{y} - \underset{\sim}{\mu}\right]$$

onde $\underset{\sim}{\eta} = \underset{\sim}{X}^{*t}\,\underset{\sim}{\beta}$ e $\underset{\sim}{\mu} = \underset{\sim}{N}\,\underset{\sim}{\pi}$. Por sua vez,

$$\underset{nk\times(k-1)}{\frac{\partial\,\underset{\sim}{\eta}}{\partial\,\underset{\sim}{\mu}}} = \frac{\partial\,\underset{\sim}{\gamma}}{\partial\,\underset{\sim}{\mu}}\,\frac{\partial\,\underset{\sim}{\eta}}{\partial\,\underset{\sim}{\gamma}}.$$

Mas, no modelo de odds proporcionais, $\log\dfrac{\gamma_{ij}}{1-\gamma_{ij}} = \eta_{ij}$ e, então, para todo $i = 1, 2,\ldots,n$ e todo $j = 1, 2,\ldots,k-1$

$$\frac{\partial\,\eta_{ij}}{\partial\,\gamma_{ij}} = \frac{1}{\dfrac{\gamma_{ij}}{1-\gamma_{ij}}}\,\frac{1\left(1-\gamma_{ij}\right)-\gamma_{ij}(-1)}{\left(1-\gamma_{ij}\right)^2} = \frac{1}{\gamma_{ij}\left(1-\gamma_{ij}\right)}.$$

Como para todo $i' \neq i$ ou $j' \neq j$, $\dfrac{\partial\,\eta_{ij}}{\partial\,\gamma_{i'j'}} = 0$, a matriz de derivadas $\dfrac{\partial\,\underset{\sim}{\eta}}{\partial\,\underset{\sim}{\gamma}}$ é diagonal, podendo ser escrita como $\text{bdiag}\left\{\underset{\sim}{C}_1^{-1},\ \underset{\sim}{C}_2^{-1},\ldots,\ \underset{\sim}{C}_n^{-1}\right\}$, onde os elementos da matriz diagonal $\underset{\sim}{C}_i^{-1}$ são definidos por

$$\underset{\sim}{C_i}^{-1} = \text{diag}\left\{\frac{1}{\gamma_{i1}(1-\gamma_{i1})}, \frac{1}{\gamma_{i2}(1-\gamma_{i2})}, \ldots, \frac{1}{\gamma_{ik-1}(1-\gamma_{ik-1})}\right\}.$$

Portanto, está provado que $\dfrac{\partial \underset{\sim}{\eta}}{\partial \underset{\sim}{\gamma}} = \left[\dfrac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\eta}}\right]^{-1} = \underset{\sim}{C}^{-1}$.

Mas,

$$\frac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\mu}} = \frac{\partial \underset{\sim}{\pi}}{\partial \underset{\sim}{\mu}} \frac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\pi}} \qquad \text{e} \qquad \underset{\sim}{\mu} = \underset{\sim}{N}\,\underset{\sim}{\pi} \qquad \Rightarrow \qquad \underset{\sim}{\pi} = \underset{\sim}{N}^{-1}\,\underset{\sim}{\mu}.$$

Portanto,

$$\frac{\partial \underset{\sim}{\pi}}{\partial \underset{\sim}{\mu}} = \left[\underset{\sim}{N}^{-1}\right]^{t} = \underset{\sim}{N}^{-1}.$$

Ainda, como $\underset{\sim}{\gamma} = \underset{\sim}{L}^{*}\,\underset{\sim}{\pi} \qquad \Rightarrow \qquad \dfrac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\pi}} = \underset{\sim}{L}^{*t}$.

Assim, $\dfrac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\mu}} = \underset{\sim}{N}^{-1}\,\underset{\sim}{L}^{*t}$ e, conseqüentemente, $\dfrac{\partial \underset{\sim}{\eta}}{\partial \underset{\sim}{\mu}} = \underset{\sim}{N}^{-1}\,\underset{\sim}{L}^{*t}\,\underset{\sim}{C}^{-1}$.

Dessa maneira, a variável dependente ajustada, para o modelo de odds proporcionais, é dada por

$$\underset{n(k-1)\times 1}{\underset{\sim}{z}} = \underset{\sim}{\eta} + \left(\frac{\partial \underset{\sim}{\eta}}{\partial \underset{\sim}{\mu}}\right)^{t}\left[\underset{\sim}{y} - \underset{\sim}{\mu}\right] = \underset{\sim}{X}^{*t}\,\underset{\sim}{\beta}^{*} + \left[\underset{\sim}{N}^{-1}\,\underset{\sim}{L}^{*t}\,\underset{\sim}{C}^{-1}\right]^{t}\left[\underset{\sim}{y} - \underset{\sim}{N}\,\underset{\sim}{\pi}\right]$$

$$\underset{\sim}{z} = \underset{\sim}{X}^{*t}\,\underset{\sim}{\beta}^{*} + \underset{\sim}{C}^{-1}\,\underset{\sim}{L}^{*}\,\underset{\sim}{N}^{-1}\left[\underset{\sim}{y} - \underset{\sim}{N}\,\underset{\sim}{\pi}\right].$$

A matriz de pesos $\underset{\sim}{W}$, por sua vez, é definida por

$$\underset{\sim}{W} = \left(\frac{\partial \underset{\sim}{\mu}}{\partial \underset{\sim}{\eta}}\right)\Sigma^{-}\left(\frac{\partial \underset{\sim}{\mu}}{\partial \underset{\sim}{\eta}}\right)^{t} = \underset{\sim}{C}^{t}\,\underset{\sim}{A}\,\underset{\sim}{C}$$

como é mostrado a seguir:

$$\frac{\partial \underset{\sim}{\mu}}{\partial \underset{\sim}{\eta}} = \frac{\partial \underset{\sim}{\gamma}}{\partial \underset{\sim}{\eta}}\frac{\partial \underset{\sim}{\mu}}{\partial \underset{\sim}{\gamma}} = \underset{\sim}{C}\,\underset{\sim}{L}^{*-t}\,\underset{\sim}{N}$$

pois, $\dfrac{\partial \underset{\sim}{\mu}}{\partial \underset{\sim}{\gamma}} = \dfrac{\partial \underset{\sim}{\pi}}{\partial \underset{\sim}{\gamma}}\dfrac{\partial \underset{\sim}{\mu}}{\partial \underset{\sim}{\pi}} = \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}$, onde $\underset{\sim}{\mu} = \underset{\sim}{N}\,\underset{\sim}{\pi}$ e $\dfrac{\partial \underset{\sim}{\mu}}{\partial \underset{\sim}{\pi}} = \underset{\sim}{N}^{t} = \underset{\sim}{N}$.

Ainda, o sistema de equações $\underset{\sim}{\gamma} = \underset{\sim}{L}^{*}\,\underset{\sim}{\pi}$ possui uma solução $\underset{\sim}{\pi} = \underset{\sim}{L}^{*-}\,\underset{\sim}{\gamma} + \left[\underset{\sim}{I}_{nk} - \underset{\sim}{L}^{*-}\,\underset{\sim}{L}^{*}\right]\underset{\sim}{h}$, para todo vetor $\underset{\sim}{h}$ de dimensão $nk \times 1$. Assim, $\dfrac{\partial \underset{\sim}{\pi}}{\partial \underset{\sim}{\gamma}} = \underset{\sim}{L}^{*-t}$ e, como $\underset{\sim}{C}$ e $\underset{\sim}{N}$ são matrizes diagonais, está provado que $\underset{\sim}{W} = \underset{\sim}{C}^{t}\,\underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\Sigma^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-}\,\underset{\sim}{C} = \underset{\sim}{C}^{t}\,\underset{\sim}{A}\,\underset{\sim}{C}$.

Retornando ao procedimento iterativo IRLS, na s-ésima etapa a solução para $\underset{\sim}{\beta}^{*}$ é obtida por

$$\underset{\sim}{\beta}^{*(s+1)} = \left[\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(s)}\,\underset{\sim}{X}^{*t}\right]^{-1}\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(s)}\,\underset{\sim}{z}^{(s)}$$

onde $\underset{\sim}{W}^{(s)}$ e $\underset{\sim}{z}^{(s)}$ são os valores de $\underset{\sim}{W}$ e $\underset{\sim}{z}$ avaliadas em $\underset{\sim}{\beta}^{*(s)}$. Em outras palavras, no s-ésimo ciclo está sendo realizada uma regressão da nova variável dependente $\underset{\sim}{z}^{(s)}$ em $\underset{\sim}{X}^{*}$, ponderada pela matriz $\underset{\sim}{W}^{(s)}$, de tal forma que se produz uma nova estimativa para $\underset{\sim}{\beta}^{*}$, dada por $\underset{\sim}{\beta}^{*(s+1)}$.

Note que o produto das matrizes $\underset{\sim}{X}^{*}\,\underset{\sim}{W}\,\underset{\sim}{X}^{*t}$ define a matriz de informação para os parâmetros do modelo de odds proporcionais, pois $\underset{\sim}{D} = \underset{\sim}{C}\,\underset{\sim}{X}^{*t}$; isso é:

$$\underset{\sim}{X}^{*}\,\underset{\sim}{W}\,\underset{\sim}{X}^{*t} = \underset{\sim}{X}^{*}\,\underset{\sim}{C}^{t}\,\underset{\sim}{A}\,\underset{\sim}{C}\,\underset{\sim}{X}^{*t} = \underset{\sim}{D}^{t}\,\underset{\sim}{A}\,\underset{\sim}{D} = \underset{\sim}{I}.$$

O próximo passo é determinar $\underset{\sim}{X}^{*}\,\underset{\sim}{W}\,\underset{\sim}{z}$ e a inversa da matriz de informação $\underset{\sim}{X}^{*}\,\underset{\sim}{W}\,\underset{\sim}{X}^{*t}$, de tal forma que o produto dessas quantidades determina o valor de estimativa de $\underset{\sim}{\beta}^{*}$ do ciclo corrente.

O problema agora é o ponto inicial do procedimento iterativo; ou seja, o começo do ciclo. Uma maneira simples de iniciar o processo consiste em usar os dados $\underset{\sim}{y}$ como primeira estimativa de $\underset{\sim}{\mu}$ e, assim, determinar a primeira estimativa da matriz de pesos $\underset{\sim}{W}$. Conseqüentemente, a primeira estimativa de $\underset{\sim}{\beta}^{*}$ pode ser calculada. O procedimento continua até que a diferença entre duas estimativas consecutivas seja suficientemente pequena.

A matriz de covariância assintótica de $\hat{\underset{\sim}{\beta}}^{*}$ é a inversa da matriz de informação, estimada por

$$\widehat{\mathrm{Cov}}\left(\hat{\underset{\sim}{\beta}}^{*}\right) = \left[\underset{\sim}{X}^{*}\,\hat{\underset{\sim}{W}}\,\underset{\sim}{X}^{*t}\right]^{-1}$$

onde $\hat{\underset{\sim}{W}}$ é a matriz $\underset{\sim}{W}$ avaliada em $\hat{\underset{\sim}{\beta}}^{*}$.

O algoritmo para ajuste de modelos lineares generalizados foi apresentado por Agresti (1990, p.447-451). Entretanto, McCullagh & Nelder (1989, p.40) e Dobson (1983, p.30) também ilustram o desenvolvimento do método IRLS para o ajuste de modelos lineares generalizados a partir do método de Newton-Raphson. Outros detalhes também podem ser encontrados no artigo publicado por Green (1984), que contém uma discussão.

4.1.3 DIFICULDADES DE ESTIMAÇÃO

A unicidade das estimativas de máxima verossimilhança para os parâmetros do modelo de odds proporcionais, depende basicamente da identificabilidade do modelo e da concavidade da função de verossimilhança, veja McCullagh (1980).

O problema estatístico de identificabilidade surge em diversas áreas, quando o modelo estatístico não está completamente especificado. Na teoria dos modelos lineares, esse problema está associado à estimabilidade de parâmetros lineares. Ele aparece também na área do Planejamento de Experimentos, particularmente na utilização dos métodos de confundimento (confounding, em inglês). Quando uma estrutura de confundimento é usada, a identificabilidade de certos parâmetros (correspondentes, por exemplo, a interações de segunda ordem) é sacrificada para proceder à estimação e testagem de parâmetros de interesse maior, que permanecem estimáveis (efeitos principais e interações de primeira ordem, por exemplo), veja Basu (1983).

Nesse momento é útil introduzir uma definição, apresentada por Basu (1983). Definições similares são apresentadas em Bunke & Bunke (1986, p.44) e Graybill (1976, p.483).

Seja U uma variável aleatória com função de distribuição F_θ pertencente a uma família $\mathcal{F} = \{F_\theta : \theta \in \Omega\}$ de funções de distribuição indexadas por um parâmetro θ, que pode ser vetorial. Diz-se que θ é não identificável por U se existe no mínimo um par (θ, θ'), onde θ e $\theta' \in \Omega$ e $\theta \neq \theta'$, tal que $F_\theta(u) = F_{\theta'}(u)$ para todo u. Caso contrário, θ é dito identificável.

A não existência de estimativas dos parâmetros no modelo de odds proporcionais é um problema computacional causado freqüentemente pela natureza dos delineamentos e pela paucidade dos dados.

No caso específico dos delineamentos fracionados, veja Box, Hunter and Hunter (1978, p.374), a estrutura de confundimento faz com que alguns parâmetros do modelo se tornem não identificáveis - geralmente as interações de maior ordem. Esse fato é explorado na geração da estrutura de confundimento adotada, que torna não identificáveis parâmetros (correspondentes, por exemplo, a interações de ordem elevada), que não podem ser detectados ou estão associados com variáveis inertes.

Assumindo que o problema de identificabilidade foi eliminado (por exemplo, pela imposição de restrições adequadas), deve ser investigada a concavidade da função de verossimilhança.

Se essa função é estritamente côncava, então possui apenas um máximo. A concavidade da função está garantida se a matriz de derivadas parciais de segunda ordem for negativa definida, com autovalores distantes de zero. Pode ser mostrado que quando o tamanho da amostra aumenta $(n \to \infty)$, a probabilidade do máximo ser único converge para 1, veja McCullagh (1980).

Na prática, contudo, quando a tabela de contingência possui muitas células vazias, os valores dos parâmetros podem divergir. No contexto de experimentos industriais com resposta categórica ordenada, esse fato decorre do pequeno número de replicações utilizadas.

Koch et al. (1990) apresentaram um método para resolver esse problema de convergência. Ele consiste basicamente em realizar um alisamento da matriz de observações. O modelo é então ajustado a nova tabela de contingência, que agora não contém células vazias. Uma definição e aplicação desse procedimento será realizada na Seção 4.3.2, no ajuste de um modelo de odds proporcionais aos dados do Exemplo B. Contudo, como o parâmetro de alisamento parece ser arbitrário, a validade e a eficiência desse método precisam ser melhor analisadas.

4.2. ASPECTOS COMPUTACIONAIS

O objetivo dessa seção é discutir alguns aspectos dos procedimentos computacionais disponíveis para o ajuste de modelos de odds proporcionais, tais como suas características e dificuldades.

McCullagh (1979) apresentou um conjunto de programas escritos na linguagem FORTRAN, denominado PLUM. Eles foram obtidos no Centro de Computação da Universidade de Dortmund por cortesia dos professores M. Schumacher e A. Infante. Com esses programas é possível ajustar modelos da classe

$$\text{link}\left\{\gamma_{ij}\right\} = \frac{\left[\theta_j - \beta_i' \underset{\sim}{x}_i\right]}{\sigma_i}$$

onde link é uma função de ligação do tipo logit, probit, cauchit ou complementar log-log. As quantidades $\beta_i^t x_i$ e σ_i são chamadas, respectivamente, de posição e escala da amostra i.

Considere que a distribuição logística é postulada para a variável contínua subjacente. Então, o modelo de odds proporcionais pode ser obtido mediante a escolha da função de ligação $\log it\{\gamma_{ij}\}$ e assumindo $\sigma_i = 1$ para todo i = 1, 2,..., n.

Contudo, foram encontradas algumas dificuldades para implementar esse programa. Não foi viável, em tempo hábil, eliminar os erros de compilação possivelmente ocasionados pelas diferenças existentes entre a versão do FORTRAN usada originalmente e aquela disponível atualmente. Outra dificuldade, não menos importante, foi a falta de documentação sobre esses programas, que impossibilitou a compreensão dos mesmos.

Uma ferramenta muito poderosa para o ajuste de modelos de odds proporcionais é o procedimento LOGISTIC do pacote estatístico SAS, veja SAS Institute Inc. (1989, p.1071). Mediante o método de máxima verossimilhança, é possível ajustar modelos de regressão logístico linear para dados com resposta binária ou categórica ordenada. Para modelos com resposta binária estão disponíveis os métodos de diagnóstico de regressão desenvolvidos por Pregibon (1981).

Três tipos de função de ligação estão disponíveis no procedimento LOGISTIC, quais sejam, logit, normit, ou complementar log-log. As variáveis explanatórias podem ser categóricas ou contínuas e o procedimento incorpora também os métodos para seleção de modelos: backward, forward e stepwise, em inglês.

Quando a resposta é categórica ordenada, o modelo de odds proporcionais pode ser obtido com a escolha da função de ligação logit. Assim, o modelo ajustado pelo procedimento LOGISTIC tem a forma $\log it\{\gamma_{ij}\} = \theta_j + \beta^t x_i$. Observe que a parametrização adotada nessa dissertação é $\log it\{\gamma_{ij}\} = \theta_j - \beta^t x_i$.

As estimativas de máxima verossimilhança dos parâmetros do modelo são calculadas através do processo iterativo IRLS descrito na Seção 4.1.2.

O procedimento LOGISTIC apresenta três critérios para avaliar o ajuste do modelo. Cabe destacar, aqui, a estatística $-2 \log L(\pi ; y)$ que pode ser usada para testar a

hipótese nula de que todas as covariáveis do modelo têm impacto nulo. Sob H_0, essa estatística tem uma distribuição de qui-quadrado, sendo que o procedimento informa o nível de significância atingido. As outras duas estatísticas, denominadas Critério de Informação de Akaike (AIC) e Critério de Schwartz (SC), são usadas para comparar diferentes modelos ajustados aos mesmos dados. São úteis quando um método automático de seleção de modelos é usado (stepwise, por exemplo).

Também está disponível um procedimento baseado em escores para testar a suposição de linhas paralelas. Sob a hipótese nula de que a suposição de odds proporcionais é adequada, a estatística de teste tem uma distribuição assintótica de qui-quadrado com $p(k-2)$ graus de liberdade.

Na próxima seção, o PROC LOGISTIC será utilizado para ajustar modelos de odds proporcionais aos dados do Exemplo A e do Exemplo B. Os passos necessários são mostrados no Anexo B, juntamente com os resultados produzidos pelo procedimento. Contudo, detalhes dos comentários acima e exemplos são apresentados em SAS Institute Inc. (1989).

Ainda relacionado com o pacote estatístico SAS, existe também o procedimento LOGIST, desenvolvido por Harrel (1986), que pode ser utilizado através do SAS. No entanto, é um procedimento suplementar do qual não dispomos atualmente.

O pacote STATA é outro software estatístico que pode ser utilizado para ajustar modelos de odds proporcionais, veja Computing Resource Center (1992, vol.3, p.77). Para tanto, o procedimento para ajuste é o comando OLOGIT, que incorpora uma opção para seleção de modelos pelos métodos forward e backward. Também está disponível uma opção para obter as probabilidades preditas pelo modelo ajustado. Os parâmetros são estimados pelo método de máxima verossimilhança, sendo que o processo iterativo de maximização é o de Newton-Raphson.

O ajuste de modelos de odds proporcionais pelo OLOGIT é simples. Isso pode ser observado para o ajuste do modelo aos dados do Exemplo A; veja a Seção 4.3.1 e a Seção B2 do Anexo B. Contudo, o PROC LOGISTIC do SAS apresenta mais recursos para a escolha e avaliação do ajuste de modelos. Uma limitação comum aos dois pacotes em questão é a ausência de procedimentos para diagnóstico e análise de resíduos, úteis para a redução do modelo.

Na próxima seção serão apresentadas duas aplicações do modelo de odds proporcionais; os ajustes serão realizados através dos pacotes estatísticos SAS e STATA.

4.3. APLICAÇÕES

4.3.1 EXEMPLO A

O Exemplo A apresentado no Capítulo 2 será utilizado aqui para ilustrar a aplicação do modelo de odds proporcionais. O ajuste do modelo será discutido em detalhes, confrontando posteriormente os resultados obtidos com aqueles produzidos pelos software estatísticos SAS e STATA, bem como com outras técnicas de análise.

Sob a suposição de amostragem multinomial em cada linha da tabela, o núcleo da função log-verossimilhança é

$$l\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right) = \sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij} \log \pi_{ij}\ ,$$

com as restrições $\sum_{j=1}^{3} \pi_{ij} = 1$ e $\sum_{j=1}^{3} y_{ij} = n_i$, para $i = 1, 2$. Os vetores $\underset{\sim}{y}$ e $\underset{\sim}{\pi}$ são

$$\underset{\sim}{y} = \begin{bmatrix} \underset{\sim}{y_1} \\ \cdots \\ \underset{\sim}{y_2} \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 19 \\ 29 \\ 24 \\ 497 \\ 560 \\ 269 \end{bmatrix} \qquad \underset{\sim}{\pi} = \begin{bmatrix} \underset{\sim}{\pi_1} \\ \cdots \\ \underset{\sim}{\pi_2} \end{bmatrix} = \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \\ \pi_{21} \\ \pi_{22} \\ \pi_{23} \end{bmatrix}.$$

O modelo a ser ajustado é

$$\log \frac{\gamma_{ij}}{1-\gamma_{ij}} = \theta_j - \beta x_i\ , \qquad i = 1, 2; \qquad j = 1, 2, 3;$$

onde $x_2 = 0$ para o grupo de não portadores de *Streptococcus pyogenes* e $x_1 = 1$ para os portadores. Assim, os parâmetros do modelo são os interceptos θ_1 e θ_2 e o coeficiente angular β, que compõem o vetor de parâmetros $\underset{\sim}{\beta}^* = (\theta_1, \theta_2, \beta)^t$.

O vetor de probabilidades acumuladas $\underset{\sim}{\gamma} = (\gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22})^t$ pode ser escrito como $\underset{\sim}{\gamma} = \underset{\sim}{L}^* \underset{\sim}{\pi}$, onde

$$\underset{\sim}{L}^* = \underset{\sim}{I}_2 \otimes \underset{\sim}{L} = \begin{bmatrix} 1 & 0 & 0 & \vdots & 0 & 0 & 0 \\ 1 & 1 & 0 & \vdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \vdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \vdots & 1 & 0 & 0 \\ 0 & 0 & 0 & \vdots & 1 & 1 & 0 \end{bmatrix} \quad \text{e} \quad \underset{\sim}{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

De (4.9) e (4.10) resulta $\dfrac{\partial l\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)}{\partial \underset{\sim}{\beta}^*} = \underset{\sim}{D}^t \underset{\sim}{u} = \sum_{i=1}^{n} \underset{\sim}{D}_i^t \underset{\sim}{u}_i$, onde

$$\underset{\sim}{D}_i^t = \begin{bmatrix} \gamma_{i1}(1-\gamma_{i1}) & 0 \\ 0 & \gamma_{i2}(1-\gamma_{i2}) \\ -x_i\gamma_{i1}(1-\gamma_{i1}) & -x_i\gamma_{i2}(1-\gamma_{i2}) \end{bmatrix}$$

e

$$\underset{\sim}{u}_i = \begin{bmatrix} \dfrac{y_{i1} - n_i\pi_{i1}}{\pi_{i1}} - \dfrac{y_{i2} - n_i\pi_{i2}}{\pi_{i2}} \\ \dfrac{y_{i2} - n_i\pi_{i2}}{\pi_{i2}} - \dfrac{y_{i3} - n_i\pi_{i3}}{\pi_{i3}} \end{bmatrix}.$$

Assim, as equações de verossimilhança definidas por $\frac{\partial l\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)}{\partial \underset{\sim}{\beta}^{*}} = \underset{\sim}{0}$, são

$$\begin{bmatrix} -\sum_{i=1}^{2}\gamma_{i1}(1-\gamma_{i1})\left(\frac{y_{i2}-n_i\pi_{i2}}{\pi_{i2}}-\frac{y_{i1}-n_i\pi_{i1}}{\pi_{i1}}\right) \\ -\sum_{i=1}^{2}\gamma_{i2}(1-\gamma_{i2})\left(\frac{y_{i3}-n_i\pi_{i3}}{\pi_{i3}}-\frac{y_{i2}-n_i\pi_{i2}}{\pi_{i2}}\right) \\ \sum_{i=1}^{2}\sum_{j=1}^{2}x_i\gamma_{ij}(1-\gamma_{ij})\left(\frac{y_{ij+1}-n_i\pi_{ij+1}}{\pi_{ij+1}}-\frac{y_{ij}-n_i\pi_{ij}}{\pi_{ij}}\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Substituindo os valores assumidos pelas covariáveis e os valores observados da resposta, as equações tomam a forma

$$-\left\{\gamma_{11}(1-\gamma_{11})\left[\frac{29-72\,\pi_{12}}{\pi_{12}}-\frac{19-72\,\pi_{11}}{\pi_{11}}\right]+\gamma_{21}(1-\gamma_{21})\left[\frac{560-1326\,\pi_{22}}{\pi_{22}}-\frac{497-1326\,\pi_{21}}{\pi_{21}}\right]\right\}=0$$

$$-\left\{\gamma_{12}(1-\gamma_{12})\left[\frac{24-72\,\pi_{13}}{\pi_{13}}-\frac{29-72\,\pi_{12}}{\pi_{12}}\right]+\gamma_{22}(1-\gamma_{22})\left[\frac{269-1326\,\pi_{23}}{\pi_{23}}-\frac{560-1326\,\pi_{22}}{\pi_{22}}\right]\right\}=0$$

$$\left\{\gamma_{11}(1-\gamma_{11})\left[\frac{29-72\,\pi_{12}}{\pi_{12}}-\frac{19-72\,\pi_{11}}{\pi_{11}}\right]+\gamma_{12}(1-\gamma_{12})\left[\frac{24-72\,\pi_{13}}{\pi_{13}}-\frac{29-72\,\pi_{12}}{\pi_{12}}\right]\right\}=0$$

A matriz de informação, definida pela equação (A3.1) é $\underset{\sim}{I} = \underset{\sim}{D}^t \underset{\sim}{A}\,\underset{\sim}{D}$, onde $\underset{\sim}{D} = \underset{\sim}{C}\,\underset{\sim}{X}^{*t}$ e $\underset{\sim}{A} = E\left[\underset{\sim}{u}\,\underset{\sim}{u}^t\right] = \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\underset{\sim}{\Sigma}^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-}$. As matrizes $\underset{\sim}{C}$ e $\underset{\sim}{X}^*$ estão definidas em (4.7) e (4.6), respectivamente. No exemplo

$$\underset{\sim}{X}^* = \frac{\partial\,\underset{\sim}{\eta}}{\partial\,\underset{\sim}{\beta}^*} = \left[\underset{\sim}{X}_1 \;\vdots\; \underset{\sim}{X}_2\right] = \begin{bmatrix} 1 & 0 & \vdots & 1 & 0 \\ 0 & 1 & \vdots & 0 & 1 \\ \cdots & \cdots & \vdots & \cdots & \cdots \\ -1 & -1 & \vdots & 0 & 0 \end{bmatrix}$$

e

$$\underset{\sim}{\eta} = \underset{\sim}{X}^{*t}\,\underset{\sim}{\beta}^* = \begin{bmatrix} \underset{\sim}{\eta}_1 \\ \underset{\sim}{\eta}_2 \end{bmatrix} = \begin{bmatrix} \theta_1 - \beta \\ \theta_2 - \beta \\ \cdots\cdots \\ \theta_1 \\ \theta_2 \end{bmatrix} \quad \text{onde} \quad \underset{\sim}{\eta}_i = \begin{bmatrix} \theta_1 - \beta x_i \\ \theta_2 - \beta x_i \end{bmatrix}.$$

Ainda, $\underset{\sim}{C} = \text{bdiag}\left\{\underset{\sim}{C}_1,\ \underset{\sim}{C}_2\right\}$, onde

$$\underset{\sim}{C}_i = \frac{\partial\,\underset{\sim}{\gamma}_i}{\partial\,\underset{\sim}{\eta}_i} = \begin{bmatrix} \gamma_{i1}(1-\gamma_{i1}) & 0 \\ 0 & \gamma_{i2}(1-\gamma_{i2}) \end{bmatrix}.$$

Assim,

$$\underset{\sim}{D} = \begin{bmatrix} \gamma_{11}(1-\gamma_{11}) & 0 & \vdots & 0 & 0 \\ 0 & \gamma_{12}(1-\gamma_{12}) & \vdots & 0 & 0 \\ \cdots\cdots & \cdots\cdots & \vdots & \cdots\cdots & \cdots\cdots \\ 0 & 0 & \vdots & \gamma_{21}(1-\gamma_{21}) & 0 \\ 0 & 0 & \vdots & 0 & \gamma_{22}(1-\gamma_{22}) \end{bmatrix} \begin{bmatrix} 1 & 0 & \vdots & -1 \\ 0 & 1 & \vdots & -1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & \vdots & 0 \\ 0 & 1 & \vdots & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \gamma_{11}(1-\gamma_{11}) & 0 & -\gamma_{11}(1-\gamma_{11}) \\ 0 & \gamma_{12}(1-\gamma_{12}) & -\gamma_{12}(1-\gamma_{12}) \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \gamma_{21}(1-\gamma_{21}) & 0 & 0 \\ 0 & \gamma_{22}(1-\gamma_{22}) & 0 \end{bmatrix}.$$

Mas, $\underset{\sim}{A} = \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\Sigma^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-}$ sendo que $\underset{\sim}{L}^{*-}$ é uma inversa generalizada de $\underset{\sim}{L}^{*} = \underset{\sim}{I}_2 \otimes \underset{\sim}{L}$, dada por

$$\underset{\sim}{L}^{*-} = \underset{\sim}{I}_2 \otimes \underset{\sim}{L}^{-} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \vdots & 0 & 0 \\ -1 & 1 & \vdots & 0 & 0 \\ 0 & -1 & \vdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \vdots & 1 & 0 \\ 0 & 0 & \vdots & -1 & 1 \\ 0 & 0 & \vdots & 0 & -1 \end{bmatrix}.$$

A matriz Σ^{-} definida por $\Sigma^{-} = \text{bdiag}\left\{\Sigma_1^{-},\ \Sigma_2^{-}\right\}$, onde

$$\Sigma_i^{-} = \begin{bmatrix} \dfrac{1}{n_i\pi_{i1}} & 0 & 0 \\ 0 & \dfrac{1}{n_i\pi_{i2}} & 0 \\ 0 & 0 & \dfrac{1}{n_i\pi_{i3}} \end{bmatrix},$$

é uma inversa generalizada da matriz de covariâncias Σ_i. Então,

$$\Sigma^{-} = \begin{bmatrix} \frac{1}{72\,\pi_{11}} & 0 & 0 & \vdots & 0 & 0 & 0 \\ 0 & \frac{1}{72\,\pi_{12}} & 0 & \vdots & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{72\,\pi_{13}} & \vdots & 0 & 0 & 0 \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \vdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & 0 & 0 & \vdots & \frac{1}{1326\,\pi_{21}} & 0 & 0 \\ 0 & 0 & 0 & \vdots & 0 & \frac{1}{1326\,\pi_{22}} & 0 \\ 0 & 0 & 0 & \vdots & 0 & 0 & \frac{1}{1326\,\pi_{23}} \end{bmatrix}.$$

Por sua vez, $\underset{\sim}{N} = \text{bdiag}\left\{\underset{\sim}{N}_1,\ \underset{\sim}{N}_2\right\}$, onde $\underset{\sim}{N}_i = I_3 \otimes n_i$, para $i = 1,\ 2$, resultando

$$\underset{\sim}{N} = \begin{bmatrix} 72 & 0 & 0 & \vdots & 0 & 0 & 0 \\ 0 & 72 & 0 & \vdots & 0 & 0 & 0 \\ 0 & 0 & 72 & \vdots & 0 & 0 & 0 \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \vdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & 0 & 0 & \vdots & 1326 & 0 & 0 \\ 0 & 0 & 0 & \vdots & 0 & 1326 & 0 \\ 0 & 0 & 0 & \vdots & 0 & 0 & 1326 \end{bmatrix}.$$

Conseqüentemente, a matriz $\underset{\sim}{A}$ para o exemplo em questão assume a forma

$$\underset{\sim}{A} = \begin{bmatrix} 72\dfrac{\pi_{11}+\pi_{12}}{\pi_{11}\pi_{12}} & -72\dfrac{1}{\pi_{12}} & \vdots & 0 & 0 \\ -72\dfrac{1}{\pi_{12}} & 72\dfrac{\pi_{12}+\pi_{13}}{\pi_{12}\pi_{13}} & \vdots & 0 & 0 \\ \cdots\cdots\cdots & \cdots\cdots\cdots & \vdots & \cdots\cdots\cdots & \cdots\cdots\cdots \\ 0 & 0 & \vdots & 1326\dfrac{\pi_{21}+\pi_{22}}{\pi_{21}\pi_{22}} & -1326\dfrac{1}{\pi_{21}} \\ 0 & 0 & \vdots & -1326\dfrac{1}{\pi_{21}} & 1326\dfrac{\pi_{22}+\pi_{23}}{\pi_{22}\pi_{23}} \end{bmatrix}.$$

Portanto, para ajustar o modelo de odds proporcionais aos dados do Exemplo A, a matriz de informação definida em (A3.1) é

$$\underset{\sim}{I} = \underset{\sim}{D}^{t}\,\underset{\sim}{A}\,\underset{\sim}{D} = \begin{bmatrix} I_{11} & I_{21} & I_{31} \\ I_{21} & I_{22} & I_{32} \\ I_{31} & I_{32} & I_{33} \end{bmatrix},$$

sendo que os elementos I_{rs}, $r = 1, 2, 3$ e $s = 1, 2, 3$, são

$$I_{11} = 72\,\gamma_{11}^{2}(1-\gamma_{11})^{2}\left[2+\frac{1-\pi_{11}}{\pi_{11}}+\frac{1-\pi_{12}}{\pi_{12}}\right]+\ 1326\,\gamma_{21}^{2}(1-\gamma_{21})^{2}\left[2+\frac{1-\pi_{21}}{\pi_{21}}+\frac{1-\pi_{22}}{\pi_{22}}\right]$$

$$I_{21} = -72\,\gamma_{12}(1-\gamma_{12})\,\gamma_{11}(1-\gamma_{11})\left[1+\frac{1-\pi_{12}}{\pi_{12}}\right]-\ 1326\,\gamma_{21}(1-\gamma_{21})\,\gamma_{22}(1-\gamma_{22})\left[1+\frac{1-\pi_{21}}{\pi_{21}}\right]$$

$$I_{22} = 72\,\gamma_{12}^{2}(1-\gamma_{12})^{2}\left[2+\frac{1-\pi_{12}}{\pi_{12}}+\frac{1-\pi_{13}}{\pi_{13}}\right]+\ 1326\,\gamma_{22}^{2}(1-\gamma_{22})^{2}\left[2+\frac{1-\pi_{22}}{\pi_{22}}+\frac{1-\pi_{23}}{\pi_{23}}\right]$$

$$I_{31} = -72\,\gamma_{11}^{2}(1-\gamma_{11})^{2}\left[2+\frac{1-\pi_{11}}{\pi_{11}}+\frac{1-\pi_{12}}{\pi_{12}}\right]+\ 72\,\gamma_{11}(1-\gamma_{11})\,\gamma_{12}(1-\gamma_{12})\left[1+\frac{1-\pi_{12}}{\pi_{12}}\right]$$

$$I_{32} = 72\,\gamma_{11}(1-\gamma_{11})\,\gamma_{12}(1-\gamma_{12})\left[1+\frac{1-\pi_{12}}{\pi_{12}}\right] - 72\,\gamma_{12}^2(1-\gamma_{12})^2\left[2+\frac{1-\pi_{12}}{\pi_{12}}+\frac{1-\pi_{13}}{\pi_{13}}\right]$$

$$I_{33} = 72\,\gamma_{11}^2(1-\gamma_{11})^2\left[2+\frac{1-\pi_{11}}{\pi_{11}}+\frac{1-\pi_{12}}{\pi_{12}}\right] - 2\left\{72\,\gamma_{11}(1-\gamma_{11})\,\gamma_{12}(1-\gamma_{12})\left[1+\frac{1-\pi_{12}}{\pi_{12}}\right]\right\} +$$
$$+ 72\,\gamma_{12}^2(1-\gamma_{12})^2\left[2+\frac{1-\pi_{12}}{\pi_{12}}+\frac{1-\pi_{13}}{\pi_{13}}\right] .$$

A solução das equações de verossimilhança é obtida através do procedimento de mínimos quadrados iterativamente reponderados descrito na Seção 4.1.2. Assim, no s-ésimo ciclo resulta a estimativa

$$\underset{\sim}{\beta}^{*(s+1)} = \left[\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(s)}\,\underset{\sim}{X}^{*t}\right]^{-1}\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(s)}\,\underset{\sim}{z}^{(s)} .$$

Para o exemplo em questão, $\underset{\sim}{z} = \underset{\sim}{X}^{*t}\,\underset{\sim}{\beta}^{*} + \underset{\sim}{C}^{-1}\,\underset{\sim}{L}^{*}\,\underset{\sim}{N}^{-1}\left[\underset{\sim}{y} - \underset{\sim}{N}\,\underset{\sim}{\pi}\right]$, onde

$$\underset{\sim}{\eta} = \underset{\sim}{X}^{*t}\,\underset{\sim}{\beta}^{*} = \begin{bmatrix}\theta_1-\beta\\ \theta_2-\beta\\ \theta_1\\ \theta_2\end{bmatrix}$$

e

$$\underset{\sim}{C} = \begin{bmatrix}\gamma_{11}(1-\gamma_{11}) & 0 & 0 & 0\\ 0 & \gamma_{12}(1-\gamma_{12}) & 0 & 0\\ 0 & 0 & \gamma_{21}(1-\gamma_{21}) & 0\\ 0 & 0 & 0 & \gamma_{22}(1-\gamma_{22})\end{bmatrix}.$$

Portanto, a variável dependente ajustada é

$$\underset{\sim}{z} = \begin{bmatrix} \theta_1 - \beta + \dfrac{1}{\gamma_{11}(1-\gamma_{11})}\left[\dfrac{19}{72} - \pi_{11}\right] \\ \theta_2 - \beta + \dfrac{1}{\gamma_{12}(1-\gamma_{12})}\left[\dfrac{19+29}{72} - (\pi_{11}+\pi_{12})\right] \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \theta_1 - \dfrac{1}{\gamma_{21}(1-\gamma_{21})}\left[\dfrac{497}{1326} - \pi_{21}\right] \\ \theta_2 - \dfrac{1}{\gamma_{22}(1-\gamma_{22})}\left[\dfrac{497+560}{1326} - (\pi_{21}+\pi_{22})\right] \end{bmatrix}.$$

Já a matriz de pesos, definida por $\underset{\sim}{W} = \underset{\sim}{C}^t \underset{\sim}{A} \underset{\sim}{C}$ é dada por

$$\underset{\sim}{W} = \begin{bmatrix} W_{11} & W_{12} & 0 & 0 \\ W_{12} & W_{22} & 0 & 0 \\ 0 & 0 & W_{33} & W_{34} \\ 0 & 0 & W_{34} & W_{44} \end{bmatrix}$$

com os elementos W_{rs} definidos abaixo:

$$W_{11} = 72\ \frac{\pi_{11}+\pi_{12}}{\pi_{11}\pi_{12}}\ \gamma_{11}^2(1-\gamma_{11})^2$$

$$W_{12} = W_{21} = -72\ \frac{1}{\pi_{12}}\ \gamma_{11}(1-\gamma_{11})\ \gamma_{12}(1-\gamma_{12})$$

$$W_{22} = 72\ \frac{\pi_{12}+\pi_{13}}{\pi_{12}\pi_{13}}\ \gamma_{12}^2(1-\gamma_{12})^2$$

$$W_{33} = 1326\ \frac{\pi_{21} + \pi_{22}}{\pi_{21}\pi_{22}}\ \gamma_{21}^2(1-\gamma_{21})^2$$

$$W_{34} = W_{43} = -1326\ \frac{1}{\pi_{21}}\ \gamma_{21}(1-\gamma_{21})\,\gamma_{22}(1-\gamma_{22})$$

e

$$W_{44} = 1326\ \frac{\pi_{22} + \pi_{23}}{\pi_{22}\pi_{23}}\ \gamma_{22}^2(1-\gamma_{22})^2 .$$

Cabe salientar que a matriz de informação definida por $\underset{\sim}{I} = \underset{\sim}{D}^t\ \underset{\sim}{A}\ \underset{\sim}{D}$, também pode ser escrita como $\underset{\sim}{I} = \underset{\sim}{X}^*\ \underset{\sim}{W}\ \underset{\sim}{X}^{*t}$. Usando o vetor de observações $\underset{\sim}{y}$ como estimativa inicial do vetor $\underset{\sim}{\mu}$, dá-se início ao processo iterativo de estimação. Os ciclos do processo são os seguintes:

Ciclo 0: $\underset{\sim}{\mu}^{(0)} = \underset{\sim}{y} \quad \Rightarrow \quad \pi^{(0)} = \underset{\sim}{N}^{-1}\,\underset{\sim}{y} = \left[\frac{19}{72}, \frac{29}{72}, \frac{24}{72}, \frac{497}{1326}, \frac{560}{1326}, \frac{269}{1326}\right]$

$$\underset{\sim}{X}^*\,\underset{\sim}{W}^{(0)}\,\underset{\sim}{X}^{*t} = \begin{bmatrix} 383{,}7021 & -141{,}7751 & -9{,}3240 \\ -141{,}7751 & 264{,}8422 & -4{,}0906 \\ -9{,}3240 & -4{,}0906 & 13{,}4146 \end{bmatrix}$$

$$\underset{\sim}{z}^{(0)} = \underset{\sim}{X}^{*t}\,\underset{\sim}{\beta}^{*(0)} + \left[\underset{\sim}{C}^{(0)}\right]^{-1}\,\underset{\sim}{L}^*\,\underset{\sim}{N}^{-1}\left[\underset{\sim}{y} - \underset{\sim}{N}\,\pi^{(0)}\right]$$

$$\underset{\sim}{z}^{(0)} = \underset{\sim}{X}^{*t}\,\underset{\sim}{\beta}^{*(0)} = \underset{\sim}{\eta}^{(0)} = \begin{bmatrix} \log\dfrac{\gamma_{11}}{1-\gamma_{11}} \\ \log\dfrac{\gamma_{12}}{1-\gamma_{12}} \\ \log\dfrac{\gamma_{21}}{1-\gamma_{21}} \\ \log\dfrac{\gamma_{22}}{1-\gamma_{22}} \end{bmatrix} = \begin{bmatrix} -1{,}0259 \\ 0{,}6931 \\ -0{,}5116 \\ 1{,}3685 \end{bmatrix}$$

$$\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(0)}\,\underset{\sim}{z}^{(0)} = \left[-393{,}8734 \quad 430{,}9635 \quad 6{,}7303\right]^{t}.$$

Então,

$$\underset{\sim}{\beta}^{*(1)} = \left[\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(0)}\,\underset{\sim}{X}^{*t}\right]^{-1}\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(0)}\,\underset{\sim}{z}^{(0)} = \left[-0{,}5090 \quad 1{,}3635 \quad 0{,}5637\right]^{t}.$$

Ciclo 1: A partir da estimativa $\underset{\sim}{\beta}^{*(1)}$, obtém-se

$$\underset{\sim}{\eta}^{(1)} = \underset{\sim}{X}^{*t}\,\underset{\sim}{\beta}^{*(1)} = \begin{bmatrix} \theta_1^{(1)} - \beta^{(1)} \\ \theta_2^{(1)} - \beta^{(1)} \\ \theta_1^{(1)} \\ \theta_2^{(1)} \end{bmatrix} = \begin{bmatrix} -1{,}0727 \\ 0{,}7998 \\ -0{,}5090 \\ 1{,}3635 \end{bmatrix}$$

Mas, $\eta_{ij} = \log \dfrac{\gamma_{ij}}{1-\gamma_{ij}} \quad \Rightarrow \quad \gamma_{ij} = \dfrac{\exp\{\eta_{ij}\}}{1+\exp\{\eta_{ij}\}}$

$$\Rightarrow \quad \gamma_{ij}^{(1)} = \frac{\exp\left\{\eta_{ij}^{(1)}\right\}}{1+\exp\left\{\eta_{ij}^{(1)}\right\}}.$$

Assim, $\underset{\sim}{\gamma}^{(1)} = \begin{bmatrix} 0{,}2549 \\ 0{,}6899 \\ 0{,}3754 \\ 0{,}7963 \end{bmatrix}$ e $\underset{\sim}{\pi}^{(1)} = \underset{\sim}{L}^{*-}\,\underset{\sim}{\gamma}^{(1)} = \begin{bmatrix} 0{,}2594 \\ 0{,}4350 \\ 0{,}3101 \\ 0{,}3754 \\ 0{,}4209 \\ 0{,}2037 \end{bmatrix}$.

Então,

$$\underset{\sim}{W}^{(1)} = \begin{bmatrix} 16{,}1595 & -6{,}7254 & 0 & 0 \\ -6{,}7254 & 18{,}2025 & 0 & 0 \\ 0 & 0 & 367{,}4003 & -134{,}3424 \\ 0 & 0 & -134{,}3424 & 254{,}1622 \end{bmatrix}$$

e

$$\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(1)}\,\underset{\sim}{X}^{*t} = \begin{bmatrix} 383{,}5598 & & \\ -141{,}0678 & 272{,}3647 & \\ -9{,}4341 & -11{,}4771 & 20{,}9112 \end{bmatrix}$$

$$\left[\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(1)}\,\underset{\sim}{X}^{*t}\right]^{-1} = \begin{bmatrix} 0{,}0033 & & \\ 0{,}0018 & 0{,}0048 & \\ 0{,}0025 & 0{,}0034 & 0{,}0508 \end{bmatrix}.$$

Ainda,

$$\underset{\sim}{z}^{(1)} = \underset{\sim}{\eta}^{(1)} + \left[\underset{\sim}{C}^{(1)}\right]^{-1} \underset{\sim}{L}^{*}\,\underset{\sim}{N}^{-1}\left[\underset{\sim}{y} - \underset{\sim}{N}\,\underset{\sim}{\pi}^{(1)}\right] = \begin{bmatrix} -1{,}0254 \\ 0{,}6912 \\ -0{,}5115 \\ 1{,}3686 \end{bmatrix}$$

e

$$\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(1)}\,\underset{\sim}{z}^{(1)} = \left[-393{,}0048 \quad 436{,}0403 \quad 1{,}7408\right]^{t}.$$

Portanto,

$$\underset{\sim}{\beta}^{*(2)} = \left[\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(1)}\,\underset{\sim}{X}^{*t}\right]^{-1} = \underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(1)}\,\underset{\sim}{z}^{(1)} = \left[-0{,}5086 \quad 1{,}3629 \quad 0{,}6018\right]^{t}.$$

Se o procedimento de estimação fosse interrompido nesse ciclo, as estimativas dos parâmetros seriam $\hat{\underset{\sim}{\beta}}^{*} = \underset{\sim}{\beta}^{*(3)}$ com a matriz assintótica de covariâncias de $\hat{\underset{\sim}{\beta}}^{*}$ dada por $\widehat{\text{cov}}\,\hat{\underset{\sim}{\beta}}^{*} = \left[\underset{\sim}{X}^{*}\,\underset{\sim}{W}^{(2)}\,\underset{\sim}{X}^{*t}\right]^{-1}$. Portanto, os parâmetros são estimados por $\hat{\theta}_1 = 0{,}5086$; $\hat{\theta}_2 = 1{,}3630$ e $\hat{\beta} = 0{,}6028$. Esses resultados estão muito próximos daqueles obtidos através do procedimento LOGISTIC do software estatístico SAS, constantes na Seção B1 do Anexo B. As estimativas dos parâmetros são $\hat{\theta}_1 = 0{,}5085$; $\hat{\theta}_2 = 1{,}3627$ e $\hat{\beta} = 0{,}6026$, obtendo-se naquele caso convergência no quarto ciclo do procedimento de estimação.

É importante salientar que a diferença no sinal da estimativa de $\hat{\beta}$ é consequência da parametrização do modelo usada pelo SAS.

Para os dados do exemplo, foi observado o nível de significância $p = 0{,}5743$ associado à estatística de escores que permite testar a adequacidade do modelo, sugerindo que o de odds proporcionais é apropriado.

Esse modelo também pode ser ajustado através do comando OLOGIT do software estatístico STATA, veja Computing Resource Center (1992, vol.3, p.77). As estimativas dos parâmetros são idênticas àquelas do PROC LOGISTIC, exceto pelo fato da parametrização usada no STATA ser a mesma dessa dissertação, veja os resultados na Seção B2 do Anexo B.

As probabilidades preditas pelo modelo ajustado, de observar uma resposta em cada categoria, para portadores e não portadores de *Streptococcus pyogenes*, determinadas através da relação

$$P\left[Y = j \mid \underset{\sim}{x}\right] = P\left[Y \le j \mid \underset{\sim}{x}\right] - P\left[Y \le j-1 \mid \underset{\sim}{x}\right],$$

onde $P\left[Y \le j \mid \underset{\sim}{x}\right] = \dfrac{\exp\left\{\hat{\theta}_j - \hat{\underset{\sim}{\beta}}^{t}\,\underset{\sim}{x}\right\}}{1 + \exp\left\{\hat{\theta}_j - \hat{\underset{\sim}{\beta}}^{t}\,\underset{\sim}{x}\right\}}$, para $1 \le j < k$ são exibidas na Tabela 4.3.

Tabela 4.3 - Probabilidades observadas e preditas pelo modelo de odds proporcionais para as categorias de resposta do tamanho relativo das amígdalas.

Streptococcus	Probabilidades observadas			Probabilidades preditas		
pyogenes	1	2	3	1	2	3
Portadores	0,26	0,41	0,33	0,25	0,43	0,32
Não portadores	0,38	0,42	0,20	0,38	0,42	0,20

Como pode ser constatado na Tabela 4.3, existe uma similaridade entre as distribuições das probabilidades preditas com as respectivas distribuições observadas. Assim, o modelo de odds proporcionais parece descrever razoavelmente bem as relações entre tamanho relativo de amígdalas e presença de *Streptococcus pyogenes*.

Uma interpretação visual do modelo de odds proporcionais para esse exemplo é mostrada na Figura 4.1, na qual percebe-se claramente que o modelo ajustado é de regressão com linhas paralelas. O segmento não pontilhado representa a estrutura linear do modelo para uma resposta na primeira categoria, que satisfaz

$$\log \frac{\gamma_{i1}}{1-\gamma_{i1}} = \theta_1 - \beta x_i ;$$

isso é, para os indivíduos

$$\text{portadores:} \qquad x_1 = 1 \qquad \Rightarrow \qquad \log \frac{\gamma_{11}}{1-\gamma_{11}} = \theta_1 - \beta$$

$$\text{não portadores:} \qquad x_2 = 0 \qquad \Rightarrow \qquad \log \frac{\gamma_{21}}{1-\gamma_{21}} = \theta_1 .$$

Por sua vez, o segmento pontilhado representa a equação para uma resposta menor ou igual a segunda categoria, satisfazendo:

portadores: $x_1 = 1 \quad \Rightarrow \quad \log \frac{\gamma_{12}}{1-\gamma_{12}} = \theta_2 - \beta$

não portadores: $x_2 = 0 \quad \Rightarrow \quad \log \frac{\gamma_{22}}{1-\gamma_{22}} = \theta_2 .$

Substituindo os parâmetros dessas equações pelos respectivos valores estimados, obtém-se as estimativas dos preditores lineares, os quais podem ser visualizados na Figura 4.1. Os números dentro dos retângulos representam o inverso da transformação logito correspondente; isto é,

$$0{,}6014 = \frac{\hat{\gamma}_{21}}{1-\hat{\gamma}_{21}} = \exp\{\hat{\theta}_1\}$$

e

$$0{,}3292 = \frac{\hat{\gamma}_{11}}{1-\hat{\gamma}_{11}} = \exp\{\hat{\theta}_1 - \hat{\beta}\}.$$

Portanto, o quociente $\frac{0{,}6014}{0{,}3292} = \exp\{\hat{\beta}(x_1 - x_2)\} = \exp\{0{,}6026\} = 1{,}83$ pode ser interpretado como uma medida da chance relativa de um tamanho de amígdalas não aumentada $(j = 1)$ para um indivíduo não portador com respeito a um portador: ela é 1,8 vezes maior para um não portador. Analogamente, a chance relativa de um tamanho de amígdalas não aumentada ou aumentada $(j = 1 \text{ ou } j = 2)$ é $\frac{3{,}9067}{2{,}1385} = 1{,}83$ vezes maior para os não portadores do que para os portadores. Deve ser observado que essas conclusões quantitativas somente podem ser obtidas com base em um modelo paramétrico, veja McCullagh (1980).

A conclusão qualitativa é que o tamanho das amígdalas parece ser maior no grupo de indivíduos portadores da bactéria *Streptococcus pyogenes*.

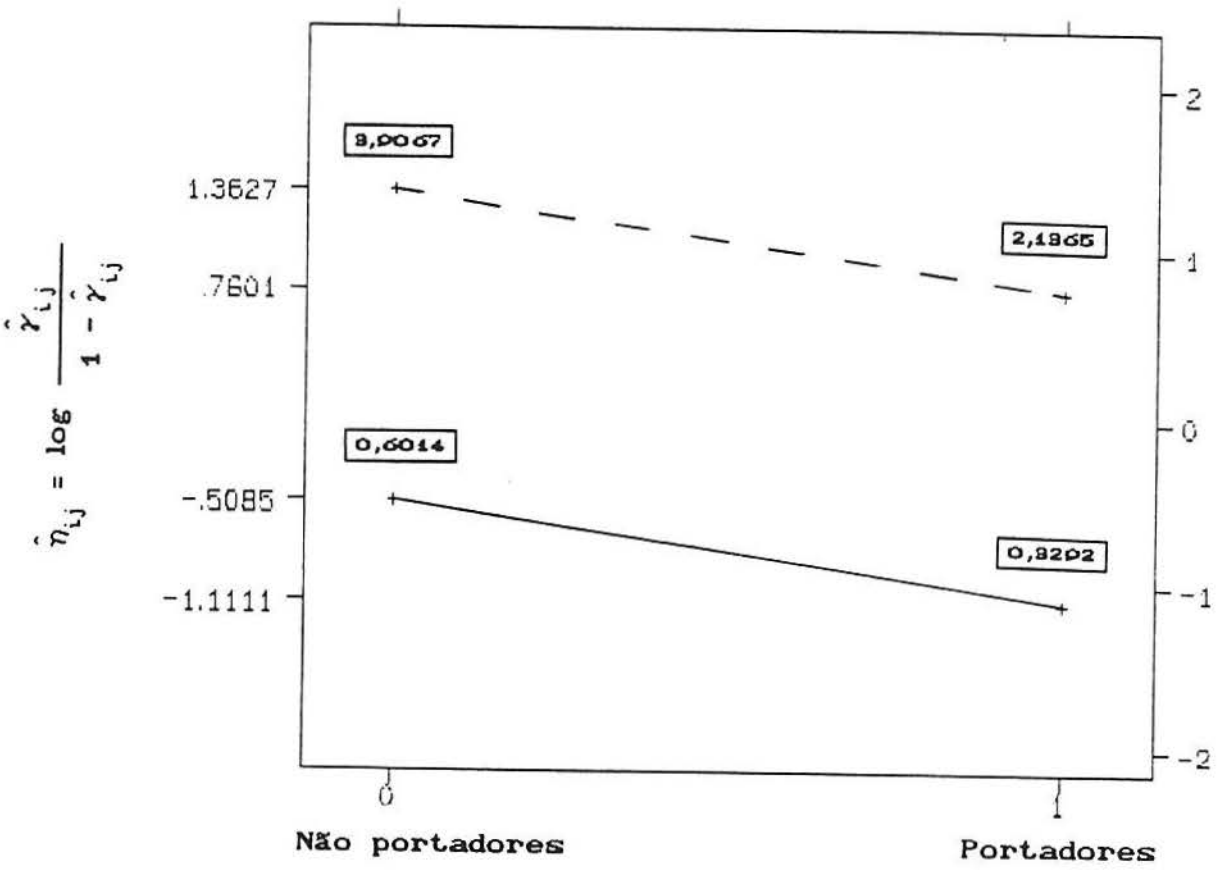

Figura 4.1 - Modelo de odds proporcionais ajustado aos dados do Exemplo A. O segmento não pontilhado representa a estrutura linear do modelo para uma resposta na primeira categoria, enquanto o segmento pontilhado representa a equação para uma resposta menor ou igual a segunda categoria . Os valores das janelas correspondem à exponencial dos respectivos preditores lineares $\hat{\eta}_{ij} = \log \frac{\hat{\gamma}_{ij}}{1-\hat{\gamma}_{ij}}$.

Tabela 4.4 - Resumo dos resultados das técnicas de análise estatística aplicadas aos dados do Exemplo A.

Método de análise	Conclusões
χ^2 DE PEARSON *veja pág. 12*	Há evidências de que as proporções das categorias de tamanho de amígdalas são diferentes para portadores e não portadores de *Streptococcus pyogenes*.
DECOMPOSIÇÃO $\chi^2 = \chi^2_{LIN} + \chi^2_{RES}$ *veja pág. 15*	Há evidências de que as crianças infectadas pelo *Streptococcus pyogenes* apresentam amígdalas maiores.
TESTE DE MANN-WHITNEY *veja pág. 16*	Há evidências de que as crianças portadoras do *Streptococcus pyogenes* possuem amígdalas maiores do que os não portadores.
RIDIT ANALYSIS *veja pág. 23*	Há evidências de que o tamanho relativo das amígdalas é maior para as crianças portadoras de *Streptococcus pyogenes*.
ANÁLISE DE ACUMULAÇÃO *veja pág. 39*	Há evidências de que o *Streptococcus pyogenes* provoca um impacto significativo no tamanho relativo das amígdalas.
MODELO DE ODDS PROPORCIONAIS *veja pág. 97 e McCullagh (1980)*	A chance relativa de um tamanho de amígdala aumentada ou grandemente aumentada é 1,8 vezes maior para os portadores do que para os não portadores de *Streptococcus pyogenes*.

4.3.2 EXEMPLO B

Nessa seção será ajustado um modelo de odds proporcionais aos dados do experimento para melhorar a qualidade de camisas termoplásticas. Para tanto, será utilizado o procedimento LOGISTIC do software estatístico SAS. Os aspectos computacionais encontram-se nas Seções B3 e B4 do Anexo B.

Face à natureza do delineamento fatorial fracionado empregado nesse experimento (do tipo 2_{III}^{15-11}), o modelo completo que poderia ser ajustado é um com efeitos principais para cada um dos 15 fatores. Cabe observar que esses efeitos estão confundidos com as interações de segunda ordem, supostas negligenciáveis.

Como a variável resposta é observada segundo as três categorias do grau de contração das camisas, outra conseqüência do fracionamento é um grande número de caselas vazias: veja a Tabela 2.6, que contém os resultados do experimento. Isso provoca a não convergência do processo iterativo de estimação dos parâmetros do modelo, veja a Seção B3 do Anexo B. Esse problema também foi constatado através da utilização do comando OLOGIT do software STATA.

Assim sendo, o ajuste do modelo de odds proporcionais por máxima verossimilhança requer um procedimento computacional modificado, proposto por Koch et al. (1990). A modificação consiste em atribuir o valor $^{1,05}/_{1,15} = 0,9130$ à resposta observada em cada replicação e o valor $^{0,05}/_{1,15} = 0,0435$ às outras respostas possíveis. Assim, a distribuição observada é misturada com uma distribuição uniforme na razão de 1 para 0,15; (isto é, aproximadamente 6:1). A Tabela 4.5 mostra a nova matriz de observações, resultante desse alisamento dos dados.

O procedimento LOGISTIC foi utilizado para ajustar o modelo com os dados modificados. O teste de escores para a hipótese de validade do modelo odds proporcionais apresentou um nível de significância de 9,22%, indicando que o modelo não se ajusta bem aos dados. Ainda, através da estatística qui-quadrado de Wald pode ser testada a hipótese de que o impacto produzido por um particular fator é nulo. Nesse modelo apenas os coeficientes de regressão dos fatores -E e G são diferentes de zero ao nível de significância 0,05. Assim, é aconselhável ajustar o modelo de odds proporcionais apenas aos fatores -E e G. Os resultados do ajuste do modelo completo para os dados modificados estão na Seção B4 do Anexo B.

Tabela 4.5 - Matriz de observações para os dados da Tabela 2.6 resultante do alisamento proposto por Koch et al. (1990).

Fator														Categoria de resposta		
H	D	-L	B	-J	-F	N	A	-I	-E	M	-C	K	G	-O	1	2	3
-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1	0,1739	0,1739	3,6522
1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1	0,1739	0,1739	3,6522
-1	1	-1	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1	3,6522	0,1739	0,1739
1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	3,6522	0,1739	0,1739
-1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1,0435	2,7826	0,1739
1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	3,6522	0,1739	0,1739
-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1,0435	2,7826	0,1739
1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	3,6522	0,1739	0,1739
-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1	3,6522	0,1739	0,1739
1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	3,6522	0,1739	0,1739
-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	0,1739	0,1739	3,6522
1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	0,1739	0,1739	3,6522
-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	2,7826	1,0435	0,1739
1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	0,1739	3,6522	0,1739
-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	0,1739	2,7826	1,0 435
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0,1739	0,1739	3,6522

Ainda usando o PROC LOGISTIC, foram obtidas as estimativas de máxima verossimilhança do modelo de odds proporcionais para o estudo das relações entre o grau de contração das camisas e os efeitos principais dos fatores -E e G. As estimativas resultantes, seus erros padrões e os níveis de significância p atingidos pela correspondente estatística de Wald são mostrados na Tabela 4.6.

Tabela 4.6 - Estimativas de máxima verossimilhança do modelo de odds proporcionais, para as relações entre o grau de contração das camisas e os efeitos dos fatores -E e G.

	Parâmetro			
	θ_1	θ_2	β_{-E}	β_G
Estimativa	-0,5026	1,2956	1,8937	-1,3897
Erro Padrão	0,3423	0,3905	0,3839	0,3632
p	0,1421	0,0009	0,0001	0,0001

Nota: No procedimento LOGISTIC $\eta_{ij} = \theta_j + \underset{\sim}{\beta}^t \underset{\sim}{x}_i$

O teste baseado em escores para a suposição de odds proporcionais apresentou um nível de significância $p = 0{,}4671$, indicando um bom ajuste do modelo. Para interpretar os parâmetros estimados, é conveniente escrever o modelo ajustado na forma

$$\hat{\eta}_{ij} = \log \frac{\hat{\gamma}_{ij}}{1 - \hat{\gamma}_{ij}} = \hat{\theta}_j - \underset{\sim}{\hat{\beta}}^t \underset{\sim}{x}_i ;$$

para $i = 1, 2$ e $j = 1, 2$; ou seja,

$$\hat{\eta}_{ij} = \hat{\theta}_j - \hat{\beta}_{-E} x_{i1} - \hat{\beta}_G x_{i2} = \hat{\theta}_j - 1{,}8937\, x_{i1} + 1{,}3897\, x_{i2}$$

onde x_{i1} e x_{i2} são os componentes dos vetores de covariáveis, definidos pelas combinações dos níveis dos fatores -E e G. Esses vetores são $\underset{\sim}{x}_1 = (1,-1)$, $\underset{\sim}{x}_2 = (-1, 1)$, $\underset{\sim}{x}_3 = (1, 1)$ e $\underset{\sim}{x}_4 = (-1,-1)$. Assim, quando o vetor de covariáveis assume o valor $\underset{\sim}{x}_1 = (1,-1)$, por exemplo, a parte sistemática do modelo é constituída por

$$\hat{\eta}_{11} = -0{,}5026 - 1{,}8937 - 1{,}3897 = -3{,}7860$$

$$\hat{\eta}_{12} = 1{,}2956 - 1{,}8937 - 1{,}3897 = -1{,}9878 .$$

Para todos os demais valores das covariáveis, os resultados são apresentados na Tabela 4.7.

Tabela 4.7 - Estimativas dos preditores lineares do modelo ajustado aos fatores -E e G.

Valor do vetor de covariáveis	Preditor linear	
	j = 1	j = 2
$x_1 = (1,-1)$	$\hat{\eta}_{11} = -3,7860$	$\hat{\eta}_{12} = -1,9878$
$x_2 = (-1, 1)$	$\hat{\eta}_{21} = 2,7808$	$\hat{\eta}_{22} = 4,5790$
$x_3 = (1, 1)$	$\hat{\eta}_{31} = -1,0066$	$\hat{\eta}_{32} = 0,7916$
$x_4 = (-1,-1)$	$\hat{\eta}_{41} = 0,0014$	$\hat{\eta}_{42} = 1,7996$

Assim, quando o fator G é mantido fixo, em seu nível baixo (-1) por exemplo, enquanto o fator -E passa do nível baixo para o nível alto, a chance relativa de ocorrer um grau de contração leve $(j = 1)$ é determinada por

$$\frac{\dfrac{\hat{\gamma}_{41}}{1-\hat{\gamma}_{41}}}{\dfrac{\hat{\gamma}_{11}}{1-\hat{\gamma}_{11}}} = \exp\left\{\hat{\beta}^t\left(x_1 - x_4\right)\right\} = \exp\{\hat{\eta}_{41} - \hat{\eta}_{11}\} = \exp\{3,7874\} = 44,14 .$$

Isso significa que existe aproximadamente 44 vezes mais chances do grau de contração ser leve quando o fator -E assume o nível baixo do que quando assume o nível alto, mantendo G fixo. Mas o fator E é o tipo de fio; "usual" no nível baixo e "modificado" no nível alto, enquanto o fator G representa o diâmetro de fio (diâmetro "menor" no nível baixo e "usual" no nível alto).

Como $\hat{\beta}_E = -\hat{\beta}_{-E}$, pode-se concluir que o tipo "modificado" de fio possui aproximadamente 44 vezes mais chances de provocar um grau de contração leve do que o tipo de fio "usual", quando o diâmetro do fio permanece inalterado. Por outro lado, mantendo o fator G constante, a chance de provocar um grau de contração forte é 44 vezes maior quando o tipo de fio usual é empregado, em relação ao fio modificado.

Analogamente, mantendo o fator -E fixo, a chance relativa de observar um grau de contração leve é

$$\exp\left\{\underset{\sim}{\hat{\beta}}^{t}\left(\underset{\sim}{x}_2-\underset{\sim}{x}_4\right)\right\}=\exp\left\{\hat{\eta}_{41}-\hat{\eta}_{21}\right\}=\exp\left\{-2{,}7794\right\}=0{,}0621;$$

ou seja, aproximadamente $\frac{1}{0{,}0621}\cong 16$ vezes maior quando o diâmetro do fio é o "usual"do que quando é o "menor".

Portanto, os fatores E e G parecem produzir um impacto decisivo no grau de contração das camisas termoplásticas. Conseqüentemente, para diminuir o grau de contração, deveria ser recomendado o tipo de fio modificado, com diâmetro usual, desde que isso não afete outras características de qualidade do produto.

A partir do modelo ajustado, podem ser determinadas as probabilidades de resposta preditas de cada categoria do grau de contração, através da relação

$$P\left[Y=j \mid \underset{\sim}{x}\right]=P\left[Y\le j \mid \underset{\sim}{x}\right]-P\left[Y\le j-1 \mid \underset{\sim}{x}\right],$$

onde $P\left[Y\le j \mid \underset{\sim}{x}\right]=\dfrac{\exp\left\{\hat{\theta}_j-\underset{\sim}{\hat{\beta}}^{t}\underset{\sim}{x}\right\}}{1+\exp\left\{\hat{\theta}_j-\underset{\sim}{\hat{\beta}}^{t}\underset{\sim}{x}\right\}}$, para $1\le j<k$.

Essas probabilidades são mostradas no lado direito da Tabela 4.8, onde observa-se uma forte similaridade entre as probabilidades preditas e as observadas. Portanto, o modelo de odds proporcionais parece descrever adequadamente as relações entre o grau de contração das camisas termoplásticas e os efeitos principais dos fatores E e G.

Tabela 4.8 - Probabilidades preditas pelo modelo de odds proporcionais, para avaliar as relações entre o grau de contração das camisas termoplásticas e os efeitos dos fatores E e G.

Fator		Grau de contração			Probabilidades preditas		
-E	G	Leve	Médio	Forte	Leve	Médio	Forte
+	-	0	0	16	0,02	0,10	0,88
-	+	16	0	0	0,94	0,05	0,01
+	+	5	6	5	0,27	0,42	0,31
-	-	8	8	0	0,50	0,36	0,14

Deve ser observado que as modificações impostas aos dados originais para facilitar o procedimento computacional, usualmente são menos necessárias na medida em que o tamanho da amostra aumenta, mas indispensáveis quando o número de fatores cresce. Nesse sentido, o modelo de odds proporcionais deveria ser ajustado somente com aqueles fatores identificados como mais importantes em uma etapa de triagem, veja Koch et al. (1990).

Os dados do Exemplo B já foram analisados através de diversos métodos. A Tabela 4.9 contém um resumo dos respectivos resultados obtidos, no que diz respeito aos fatores que produzem um impacto significativo no grau de contração das camisas termoplásticas. As estimativas dos parâmetros do modelo de resposta média reduzido são mostradas na Tabela 2.7 da Seção 2.5. Entre essas técnicas, o modelo ajustado de odds proporcionais parece possuir uma estrutura mais simples, possibilitando uma interpretação clara de magnitude e direção dos efeitos dos fatores.

Tabela 4.9 - Resumo dos resultados de diversas análises dos dados do Exemplo B, quanto ao impacto produzido pelos fatores no grau de contração das camisas termoplásticas.

Método de análise	Ordem de importância dos fatores
QUOCIENTE DE SINAL-RUÍDO *Quinlan (1985)*	E, G, K, A, C, F, D, H
GRÁFICO PROBABILÍSTICO NORMAL DOS EFEITOS DOS FATORES *Box (1988) e Box, Bisgaard and Fung (1988)*	E, G
ANÁLISE DE ACUMULAÇÃO *veja pág. 50*	Todos os 15 fatores
GRÁFICO PROBABILÍSTICO NORMAL DOS EFEITOS DOS FATORES, ESTIMADOS SEGUNDO OS PRINCÍPIOS DA AA *veja pág. 51 e Figura 3.1*	E, G
MODELO DE RESPOSTA MÉDIA *veja pág. 24 e Tabela 2.7*	E, G, B, F, D, A, C, K, I
MODELO DE ODDS PROPORCIONAIS *veja pág. 103-104 e Tabela 4.8*	E, G

CAPÍTULO 5: CONSIDERAÇÕES FINAIS

Nesse trabalho foram discutidos aspectos da análise estatística de dados observacionais ou experimentais com resposta ordinal. Eles surgem freqüentemente tanto na área biomédica quanto na industrial, muitas vezes como conseqüência da descrição de regiões para variáveis latentes.

Na área industrial, em especial, os métodos estudados estão ligados aos experimentos para melhorar a qualidade de produtos e processos. Na maioria das situações, o experimentador tem como objetivo não só identificar os fatores que produzem maior impacto na característica de qualidade, mas também estimar a magnitude e direção dos efeitos. Assim, ele pode determinar as combinações dos níveis dos fatores que otimizam o processo, no sentido de atingir um nível pré-especificado de qualidade.

As técnicas de análise tradicionais, tais como a estatística χ^2, métodos não paramétricos e ridits não são adequados, pois não extraem informação suficiente dos dados. Isso é, para estimar a magnitude dos efeitos é necessário um modelo paramétrico, tal como o modelo de resposta média ou de odds proporcionais, abordados nessa dissertação. Outros modelos poderiam ser utilizados, como, por exemplo, o modelo de odds adjacentes ou de regressão logística; veja Koch et al. (1990).

Com o objetivo de analisar dados categóricos produzidos em situações experimentais, G. Taguchi introduziu a técnica da análise de acumulação. Ela foi formalmente desenvolvida para a situação descrita por apenas um fator explanatório e, posteriormente, estendida para o caso multifatorial. Ao longo do desenvolvimento do Capítulo 3 foram abordados os principais problemas metodológicos e suas conseqüências para a análise. Contudo, tendo em vista suas deficiências e também sua complexidade, a AA não é recomendada.

Seguindo os mesmos princípios da AA, foi sugerido um método para estimar os efeitos principais dos fatores. O gráfico probabilístico normal desses efeitos pode ser usado para

fazer uma triagem dos fatores, determinando aqueles que possivelmente possuem maior impacto na característica de qualidade.

Também foram mencionadas algumas modificações da técnica da AA, propostas por Nair (1986) e Hamada & Wu (1990). Basicamente, tratam dos testes de posição e de dispersão definidos a partir da decomposição da estatística modificada da AA. A eficiência desses testes, porém, deve ser investigada.

Ainda com respeito à AA, seria interessante realizar um estudo de simulação para investigar o comportamento da estatística de teste $F_A = \frac{QM_A}{QM_{ERRO}}$, verificando em que condições a distribuição nula dessa estatística se aproxima da distribuição de referência de Snedecor-Fisher.

Um método mais sofisticado para analisar dados com resposta ordinal é o modelo de odds proporcionais proposto por McCullagh (1980). Esse modelo foi detalhadamente desenvolvido no Capítulo 4 e posteriormente aplicado a dois conjuntos de dados reais apresentados na literatura. O investimento matemático feito nesse capítulo - necessário devido ao fato de que as derivações não estão disponíveis na literatura - parece rentável, pois a construção passo a passo do modelo incrementa sua compreensão e permite comparar resultados computacionais.

Os parâmetros são estimados por máxima verossimilhança. O procedimento LOGISTIC do software estatístico SAS utiliza o método de IRLS, enquanto que no STATA o método disponível é o método de Newton-Raphson. Não foram observadas discrepâncias entre os resultados produzidos por ambos os métodos.

O ajuste do modelo de odds proporcionais aos dados do Exemplo A tem como objetivo principal facilitar a compreensão do modelo, por tratar-se de um conjunto de dados simples. Por isso o processo de ajuste foi realizado passo a passo, incluindo o procedimento iterativo de estimação. Os resultados são consistentes com aqueles obtidos no SAS, no STATA e também por McCullagh (1980).

O Exemplo B, por sua vez, ilustra a potência do modelo de odds proporcionais na análise de dados com resposta ordinal. Apesar do problema de estimabilidade dos parâmetros, provocado pelo grande número de celas vazias da tabela de dados, foi possível obter um modelo simples que descreve muito bem as relações entre o grau de contração das camisas e os efeitos principais dos fatores E (tipo de fio) e G (diâmetro do fio). A modificação dos dados através do alisamento proposto por Koch et al. (1990) mostrou ser um método eficiente para resolver o

problema de convergência. Contudo, na referência citada não consta a base teórica desse método, nem uma discussão sobre as situações nas quais ela é recomendada.

Com respeito ao modelo de odds proporcionais, uma análise mais aprofundada, baseada em instrumentos gráficos de diagnóstico, pode permitir uma melhor avaliação da qualidade do ajustamento. No entanto, para respostas politômicas, esses métodos não estão implementados nos procedimentos LOGISTIC e OLOGIT. Esse é um aspecto importante, que deve ser ainda estudado. Nesse sentido, algumas referências úteis são McCullagh (1980), McCullagh & Nelder (1989, p.391) e Andersen (1992).

O modelo de odds proporcionais mostrou-se uma ferramenta útil e poderosa para analisar experimentos industriais com resposta categórica ordenada. Sua utilização parece ser mais fácil do que comentado em Hamada & Wu (1990). Como objeto de pesquisas futuras, cabe investigar ainda os demais modelos pertencentes à classe de modelos de regressão proposta por McCullagh (1980). Eles surgem naturalmente se ao invés da função $\text{logit}\left\{\gamma_j\left(\underset{\sim}{x}\right)\right\}$, que conduz ao modelo de odds proporcionais, forem usadas as funções de ligação $-\log\left\{1-\gamma_j\left(\underset{\sim}{x}\right)\right\}$ ou $\log\left\{-\log\left[1-\gamma_j\left(\underset{\sim}{x}\right)\right]\right\}$. Esses modelos são potencialmente úteis quando há evidências de que a distribuição subjacente é assimétrica. Ainda, uma generalização natural do modelo de odds proporcionais pode ser feita se for relaxada a suposição de variância constante, considerando um modelo multiplicativo definido por $\text{logit}\left\{\gamma_{ij}\left(\underset{\sim}{x}_i\right)\right\} = \dfrac{\left\{\theta_j - \underset{\sim}{\beta}^t \underset{\sim}{x}_i\right\}}{\sigma_i}$. As quantidades $\underset{\sim}{\beta}^t \underset{\sim}{x}_i$ e σ_i são chamadas de posição e escala da i-ésima linha, respectivamente. Segundo McCullagh (1986), o objetivo desse modelo é tentar descrever a variabilidade dos dados em termos de efeitos de posição e de dispersão.

ANEXOS

O Anexo A é composto por quatro seções. Na Seção A1 são obtidas as somas de quadrados da Tabela da AA, para o caso de um delineamento fatorial do tipo 2^{3-1}. Na Seção A2 são desenvolvidas detalhadamente as derivadas parciais da função log-verossimilhança da i-ésima subpopulação, para o modelo de odds proporcionais. A matriz de informação para esse modelo é desenvolvida na Seção A3. Finalmente, a Seção A4 contém uma demonstração de que, condicional ao tamanho da amostra n do i-ésimo tratamento, a variável aleatória C_{ij} definida na Seção 3.1.1 possui uma distribuição binomial com parâmetros n e γ_{ij}.

O Anexo B apresenta as rotinas necessárias para ajustar os modelos de odds proporcionais aos dados do Exemplo A e do Exemplo B, através do procedimento LOGISTIC do SAS e do comando OLOGIT do STATA. São apresentados também os resultados produzidos por essas rotinas para cada modelo ajustado, conforme descrição abaixo.

Para os dados do Exemplo A, a Seção B1 ilustra o modelo ajustado pelo PROC LOGISTIC do SAS, enquanto que o ajuste pelo OLOGIT do STATA é apresentado na Seção B2.

As Seções B3 e B4 correspondem aos dados do Exemplo B, tendo sido utilizado apenas o procedimento LOGISTIC para ajustar os modelos. Na primeira, pode ser observado que não houve convergência das estimativas dos parâmetros para os modelos ajustados. A Seção B4, por sua vez, trata do ajuste de modelos de odds proporcionais a esses dados, modificados segundo a proposta de Koch et al. (1990), veja a Seção 4.3.2. Dois modelos foram ajustados; o primeiro com todos os fatores e, posteriormente, um modelo reduzido com apenas dois fatores.

ANEXO A: CÁLCULOS E DEMONSTRAÇÕES

A1. SOMAS DE QUADRADOS

Na página 44 foi visto que para o delineamento fatorial do tipo 2^{3-1}, as somas de quadrados da AA são aquelas apresentadas na Tabela 3.13. A demonstração desse fato é como segue.

Como $\quad \overline{p}_j = \dfrac{c_{.j}}{n} \quad \Rightarrow \quad \overline{p}_j\left(1-\overline{p}_j\right) = \dfrac{c_{.j}\left(n-c_{.j}\right)}{n^2}$.

A soma total de quadrados é obtida fazendo

$$SQ_{TOTAL} = \sum_{j=1}^{K-1}\left[\overline{p}_j\left(1-\overline{p}_j\right)\right]^{-1} SQ_{TOTAL,j}$$

$$SQ_{TOTAL} = \sum_{j=1}^{K-1} \frac{n^2}{c_{.j}\left(n-c_{.j}\right)}\frac{4}{n}c_{.j}\left(n-c_{.j}\right) = 4n(K-1).$$

Analogamente,

$$SQ_A = \sum_{j=1}^{K-1}\left[\overline{p}_j\left(1-\overline{p}_j\right)\right]^{-1} SQ_{A,j}$$

$$= \sum_{j=1}^{K-1} \frac{n^2}{c_{.j}\left(n-c_{.j}\right)}\frac{\left(c_{1j}+c_{2j}-c_{3j}-c_{4j}\right)^2}{4n} = \frac{n}{4}\sum_{j=1}^{K-1}\frac{\left(c_{1j}+c_{2j}-c_{3j}-c_{4j}\right)^2}{c_{.j}\left(n-c_{.j}\right)}.$$

$$SQ_B = \sum_{j=1}^{K-1}\left[\overline{p}_j\left(1-\overline{p}_j\right)\right]^{-1} SQ_{B,j}$$

$$= \sum_{j=1}^{K-1} \frac{n^2}{c_{.j}\left(n-c_{.j}\right)}\frac{\left(c_{1j}-c_{2j}+c_{3j}-c_{4j}\right)^2}{4n} = \frac{n}{4}\sum_{j=1}^{K-1}\frac{\left(c_{1j}-c_{2j}+c_{3j}-c_{4j}\right)^2}{c_{.j}\left(n-c_{.j}\right)}$$

e

$$SQ_C = \sum_{j=1}^{K-1} \left[\bar{p}_j\left(1-\bar{p}_j\right)\right]^{-1} SQ_{C,j}$$

$$= \sum_{j=1}^{K-1} \frac{n^2}{c_{.j}\left(n-c_{.j}\right)} \frac{\left(c_{1j}-c_{2j}-c_{3j}+c_{4j}\right)^2}{4n} = \frac{n}{4} \sum_{j=1}^{K-1} \frac{\left(c_{1j}-c_{2j}-c_{3j}+c_{4j}\right)^2}{c_{.j}\left(n-c_{.j}\right)}.$$

Já a soma de quadrados do erro pode ser obtida pela diferença

$$SQ_{ERRO} = SQ_{TOTAL} - SQ_A - SQ_B - SQ_C$$

$$= 4n(K-1) - \frac{n}{4}\left[\sum_{j=1}^{K-1} \frac{\left(c_{1j}+c_{2j}-c_{3j}-c_{4j}\right)^2}{c_{.j}\left(n-c_{.j}\right)} + \sum_{j=1}^{K-1} \frac{\left(c_{1j}-c_{2j}+c_{3j}-c_{4j}\right)^2}{c_{.j}\left(n-c_{.j}\right)} + \right.$$

$$\left. + \sum_{j=1}^{K-1} \frac{\left(c_{1j}-c_{2j}-c_{3j}+c_{4j}\right)^2}{c_{.j}\left(n-c_{.j}\right)}\right].$$

■

A2. DERIVADAS DA FUNÇÃO LOG-VEROSSIMILHANÇA

Para derivar o núcleo do logaritmo da função de verossimilhança $l\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)$, sob as restrições $\sum_{j=1}^{k}\pi_{ij}=1$, para todo $i=1, 2,\ldots,n$, é necessário introduzir n multiplicadores de Lagrange $\lambda_1, \lambda_2,\ldots, \lambda_n$, derivando em relação a π_{ij} a função

$$L=l\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)-\sum_{i=1}^{n}\lambda_i\left(\sum_{j=1}^{k}\pi_{ij}-1\right);$$

ou seja,

$$\frac{\partial L}{\partial \pi_{ij}}=\frac{\partial}{\partial \pi_{ij}}\sum_{i'=1}^{n}\sum_{j'=1}^{k}y_{i'j'}\log\pi_{i'j'}-\lambda_i=\frac{y_{ij}}{\pi_{ij}}-\lambda_i=\frac{y_{ij}-\lambda_i\pi_{ij}}{\pi_{ij}}\ .$$

Anulando todas as derivadas parciais e somando com respeito ao índice j, resulta

$$\frac{\partial L}{\partial \pi_{ij}}=\frac{y_{ij}-\lambda_i\pi_{ij}}{\pi_{ij}}=0 \quad\Rightarrow\quad y_{ij}-\lambda_i\pi_{ij}=0$$

e

$$0=\sum_{j=1}^{k}\left(y_{ij}-\lambda_i\pi_{ij}\right)=\sum_{j=1}^{k}y_{ij}-\lambda_i\sum_{j=1}^{k}\pi_{ij}=n_i-\lambda_i \quad\Rightarrow\quad \lambda_i=n_i.$$

Portanto, sob as restrições $\sum_{j=1}^{k}y_{ij}=n_i$ e $\sum_{j=1}^{k}\pi_{ij}=1$,

$$\frac{\partial L\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)}{\partial \pi_{ij}}=\frac{y_{ij}-n_i\pi_{ij}}{\pi_{ij}}, \qquad \forall\ i=1, 2,\ldots,n.$$

Introduzindo a notação matricial, $\forall\ i = 1, 2, \ldots, n$,

$$\frac{\partial L\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)}{\partial\ \underset{\sim}{\pi_i}} = n_i \Sigma_i^{-}\left[\underset{\sim}{y_i} - n_i\ \underset{\sim}{\pi_i}\right].$$

É importante notar que $\dfrac{\partial L\left(\underset{\sim}{\pi_i}\ ;\ \underset{\sim}{y_i}\right)}{\partial\ \underset{\sim}{\pi_i}} = \dfrac{\partial L\left(\underset{\sim}{\pi}\ ;\ \underset{\sim}{y}\right)}{\partial\ \underset{\sim}{\pi_i}}$, uma vez que existe independência entre os n vetores de observações $\underset{\sim}{y_1}$, $\underset{\sim}{y_2}$, ... , $\underset{\sim}{y_n}$. Para maiores detalhes, veja McCullagh & Nelder (1989, p.171).

Para $i = 1, 2, \ldots, n$, seja $\underset{\sim}{\gamma_i} = \underset{\sim}{L}\,\underset{\sim}{\pi_i} = \left[\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{i\,k-1}\right]^t$ o vetor de probabilidades acumuladas, onde $\underset{\sim}{L}$ é uma matriz triangular inferior de ordem $(k-1) \times k$, cujos elementos subdiagonais são iguais a 1. Uma inversa generalizada de $\underset{\sim}{L}$ é definida por

$$\underset{\sim}{L}^{-} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Usando a definição do vetor $\underset{\sim}{\gamma_i}$, o vetor de probabilidades $\underset{\sim}{\pi_i}$ pode ser escrito como

$$\underset{\sim}{\pi_i} = \begin{bmatrix} \gamma_{i1} \\ \gamma_{i2} - \gamma_{i1} \\ \vdots \\ \gamma_{i\,k-1} - \gamma_{i\,k-2} \\ 1 - \gamma_{i\,k-1} \end{bmatrix} \text{ e, assim,}$$

$$\underset{(k-1)\times k}{\frac{\partial \underset{\sim}{\pi_i}}{\partial \underset{\sim}{\gamma_i}}} = \begin{bmatrix} \frac{\partial \pi_{i1}}{\partial \gamma_{i1}} & \frac{\partial \pi_{i2}}{\partial \gamma_{i1}} & \cdots & \frac{\partial \pi_{ik}}{\partial \gamma_{i1}} \\ \frac{\partial \pi_{i1}}{\partial \gamma_{i2}} & \frac{\partial \pi_{i2}}{\partial \gamma_{i2}} & \cdots & \frac{\partial \pi_{ik}}{\partial \gamma_{i2}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \pi_{i1}}{\partial \gamma_{i\,k-1}} & \frac{\partial \pi_{i2}}{\partial \gamma_{ik-1}} & \cdots & \frac{\partial \pi_{ik}}{\partial \gamma_{i\,k-1}} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} = \left[\underset{\sim}{L^-}\right]^t .$$

Portanto, a derivada do núcleo da função log-verossimilhança da i-ésima subpopulação, com respeito ao vetor de probabilidades acumuladas $\underset{\sim}{\gamma_i}$, definido por

$$\frac{\partial l\left(\underset{\sim}{\pi_i}\ ;\ \underset{\sim}{y_i}\right)}{\partial \underset{\sim}{\gamma_i}} = \frac{\partial \underset{\sim}{\pi_i}}{\partial \underset{\sim}{\gamma_i}} \frac{\partial l\left(\underset{\sim}{\pi_i}\ ;\ \underset{\sim}{y_i}\right)}{\partial \underset{\sim}{\pi_i}}$$

é dado pelo vetor de dimensão $(k-1)\times k$,

$$\underset{\sim}{u_i} = n_i \underset{\sim}{L^{-t}}\, \underset{\sim}{\Sigma_i^-} \left[\underset{\sim}{y_i} - n_i \underset{\sim}{\pi_i}\right], \qquad \forall\ i = 1, 2, \ldots, n. \tag{A2.1}$$

Seja $\underset{\sim}{\eta_i} = \underset{\sim}{\theta} - \left[\underset{\sim}{\beta^t}\, \underset{\sim}{x_i}\right] \underset{\sim}{1_{k-1}}$ o vetor de dimensão $(k-1)\times 1$, o qual contém os valores ajustados para a i-ésima subpopulação e $\underset{\sim}{\beta^*} = \left[\theta_1, \theta_2, \ldots, \theta_{k-1}, \beta_1, \beta_2, \ldots, \beta_p\right]^t$ o vetor dos $p^* = p + k - 1$ parâmetros do modelo de odds proporcionais. Então,

$$\underset{(k-1)\times 1}{\underset{\sim}{\eta_i}} = \begin{bmatrix} \theta_1 - \sum_{r=1}^{p}\beta_r x_{ir} \\ \theta_2 - \sum_{r=1}^{p}\beta_r x_{ir} \\ \vdots \\ \theta_{k-1} - \sum_{r=1}^{p}\beta_r x_{ir} \end{bmatrix}$$

e

$$\underset{p^*\times(k-1)}{\frac{\partial \underset{\sim}{\eta_i}}{\partial \underset{\sim}{\beta^*}}} = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\left[\theta_1 - \sum_{r=1}^{p}\beta_r x_{ir}\right] & \cdots & \frac{\partial}{\partial\theta_1}\left[\theta_{k-1} - \sum_{r=1}^{p}\beta_r x_{ir}\right] \\ \frac{\partial}{\partial\theta_2}\left[\theta_1 - \sum_{r=1}^{p}\beta_r x_{ir}\right] & \cdots & \frac{\partial}{\partial\theta_2}\left[\theta_{k-1} - \sum_{r=1}^{p}\beta_r x_{ir}\right] \\ \vdots & & \vdots \\ \frac{\partial}{\partial\beta_p}\left[\theta_1 - \sum_{r=1}^{p}\beta_r x_{ir}\right] & \cdots & \frac{\partial}{\partial\beta_p}\left[\theta_{k-1} - \sum_{r=1}^{p}\beta_r x_{ir}\right] \end{bmatrix}.$$

Denominando $\underset{\sim}{X_i}$ a matriz de derivadas parciais de $\frac{\partial \underset{\sim}{\eta_i}}{\partial \underset{\sim}{\beta^*}}$, segue que

$$\underset{\sim}{X_i} = \frac{\partial \underset{\sim}{\eta_i}}{\partial \underset{\sim}{\beta^*}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ & & & \\ -x_{i1} & -x_{i1} & \cdots & -x_{i1} \\ -x_{i2} & -x_{i2} & \cdots & -x_{i2} \\ \vdots & \vdots & & \vdots \\ -x_{ip} & -x_{ip} & \cdots & -x_{ip} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{I_{k-1}} \\ \\ -\underset{\sim}{x_i}\,\underset{\sim}{1_{k-1}^t} \end{bmatrix}$$

$$\underset{\sim}{X}_i = \left[\underset{\sim}{I}_{k-1} \;\vdots\; -1_{k-1}\, \underset{\sim}{x}_i^t \right]^t . \qquad \text{(A2.2)}$$

Por sua vez, a derivada $\dfrac{\partial \underset{\sim}{\gamma}_i}{\partial \underset{\sim}{\eta}_i}$ é dada por

$$\underset{(k-1)\times(k-1)}{\dfrac{\partial \underset{\sim}{\gamma}_i}{\partial \underset{\sim}{\eta}_i}} = \begin{bmatrix} \dfrac{\partial \gamma_{i1}}{\partial \eta_{i1}} & \dfrac{\partial \gamma_{i2}}{\partial \eta_{i1}} & \cdots & \dfrac{\partial \gamma_{i\,k-1}}{\partial \eta_{i1}} \\ \dfrac{\partial \gamma_{i1}}{\partial \eta_{i2}} & \dfrac{\partial \gamma_{i2}}{\partial \eta_{i2}} & \cdots & \dfrac{\partial \gamma_{i\,k-1}}{\partial \eta_{i2}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial \gamma_{i1}}{\partial \eta_{i\,k-1}} & \dfrac{\partial \gamma_{i2}}{\partial \eta_{i\,k-1}} & \cdots & \dfrac{\partial \gamma_{i\,k-1}}{\partial \eta_{i\,k-1}} \end{bmatrix}$$

onde $\dfrac{\partial \gamma_{ij}}{\partial \eta_{ij}} = \gamma_{ij}\left(1-\gamma_{ij}\right)$, como é mostrado a seguir: o modelo de odds proporcional especifica que

$$\log \frac{\gamma_{ij}}{1-\gamma_{ij}} = \theta_j - \sum_{r=1}^{p} \beta_r x_{ir} = \eta_{ij} .$$

Assim,

$$\frac{\gamma_{ij}}{1-\gamma_{ij}} = \exp\left\{\eta_{ij}\right\} \quad \Rightarrow \quad \gamma_{ij} = \frac{\exp\left\{\eta_{ij}\right\}}{1+\exp\left\{\eta_{ij}\right\}} .$$

Procedendo a derivação,

$$\frac{\partial \gamma_{ij}}{\partial \eta_{ij}} = \frac{\left[1+\exp\{\eta_{ij}\}\right]\exp\{\eta_{ij}\} - \left[\exp\{\eta_{ij}\}\right]^2}{\left[1+\exp\{\eta_{ij}\}\right]^2}$$

$$= \frac{\exp\{\eta_{ij}\}}{1+\exp\{\eta_{ij}\}} - \left[\frac{\exp\{\eta_{ij}\}}{1+\exp\{\eta_{ij}\}}\right]^2$$

$$= \gamma_{ij} - \gamma_{ij}^2 = \gamma_{ij}(1-\gamma_{ij}) \ .$$

Para todo $j' \neq j$, $\frac{\partial \gamma_{ij}}{\partial \eta_{ij'}} = 0$ e, assim, $\frac{\partial \underset{\sim}{\gamma_i}}{\partial \underset{\sim}{\eta_i}}$ é definida pela matriz de dimensão $(k-1)\times(k-1)$,

$$\underset{\sim}{C_i} = \text{diag}\{\gamma_{i1}(1-\gamma_{i1}),\ \gamma_{i2}(1-\gamma_{i2}),\ldots,\ \gamma_{i\,k-1}(1-\gamma_{i\,k-1})\}.$$

Portanto, para a i-ésima subpopulação, a derivada do núcleo da função log-verossimilhança, com respeito aos parâmetros do modelo, é definida por

$$\underset{p^*\times 1}{\frac{\partial l\left(\underset{\sim}{\pi_i}\ ;\ \underset{\sim}{y_i}\right)}{\partial\ \underset{\sim}{\beta^*}}} = \frac{\partial \underset{\sim}{\gamma_i}}{\partial \underset{\sim}{\beta^*}} \frac{\partial l\left(\underset{\sim}{\pi_i}\ ;\ \underset{\sim}{y_i}\right)}{\partial \underset{\sim}{\gamma_i}} = \frac{\partial \underset{\sim}{\eta_i}}{\partial \underset{\sim}{\beta^*}} \frac{\partial \underset{\sim}{\gamma_i}}{\partial \underset{\sim}{\eta_i}} \frac{\partial l\left(\underset{\sim}{\pi_i}\ ;\ \underset{\sim}{y_i}\right)}{\partial \underset{\sim}{\gamma_i}} = \underset{\sim}{D_i^t}\ \underset{\sim}{u_i}$$

onde $\underset{\sim}{D_i}$ é a matriz de ordem $(k-1)\times p^*$ definida por

$$\underset{\sim}{D_i} = \underset{\sim}{C_i}\,\underset{\sim}{X_i^t} = \frac{\partial\,\underset{\sim}{\gamma_i}}{\partial\,\underset{\sim}{\eta_i}}\left[\frac{\partial\,\underset{\sim}{\eta_i}}{\partial\,\underset{\sim}{\beta^*}}\right]^t. \qquad \text{(A2.3)}$$

Assim,

$$\underset{\sim}{u_i} = n_i\,\underset{\sim}{L^{-t}}\,\Sigma_i^{-}\left[\underset{\sim}{y_i} - n_i\,\underset{\sim}{\pi_i}\right] = \begin{bmatrix} \dfrac{y_{i1}-n_i\pi_{i1}}{\pi_{i1}} - \dfrac{y_{i2}-n_i\pi_{i2}}{\pi_{i2}} \\ \dfrac{y_{i2}-n_i\pi_{i2}}{\pi_{i2}} - \dfrac{y_{i3}-n_i\pi_{i3}}{\pi_{i3}} \\ \vdots \\ \dfrac{y_{i\,k-2}-n_i\pi_{i\,k-2}}{\pi_{i\,k-2}} - \dfrac{y_{i\,k-1}-n_i\pi_{i\,k-1}}{\pi_{i\,k-1}} \\ \dfrac{y_{i\,k-1}-n_i\pi_{i\,k-1}}{\pi_{i\,k-1}} - \dfrac{y_{ik}-n_i\pi_{ik}}{\pi_{ik}} \end{bmatrix}$$

$$\underset{\sim}{D_i^t} = \begin{bmatrix} \gamma_{i1}(1-\gamma_{i1}) & 0 & \cdots & 0 \\ 0 & \gamma_{i2}(1-\gamma_{i2}) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \gamma_{i\,k-1}(1-\gamma_{i\,k-1}) \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ -x_{i1}\gamma_{i1}(1-\gamma_{i1}) & -x_{i1}\gamma_{i2}(1-\gamma_{i2}) & \cdots & -x_{i1}\gamma_{i\,k-1}(1-\gamma_{i\,k-1}) \\ -x_{i2}\gamma_{i1}(1-\gamma_{i1}) & -x_{i2}\gamma_{i2}(1-\gamma_{i2}) & \cdots & -x_{i2}\gamma_{i\,k-1}(1-\gamma_{i\,k-1}) \\ \vdots & \vdots & & \vdots \\ -x_{ip}\gamma_{i1}(1-\gamma_{i1}) & -x_{ip}\gamma_{i2}(1-\gamma_{i2}) & \cdots & -x_{ip}\gamma_{i\,k-1}(1-\gamma_{i\,k-1}) \end{bmatrix}.$$

Então,

$$\frac{\partial l\left(\underset{\sim}{\pi_i}\ ;\ \underset{\sim}{y_i}\right)}{\partial \underset{\sim}{\beta^*}} = \underset{\sim}{D_i^!}\ \underset{\sim}{u_i} = \begin{bmatrix} -\gamma_{i1}\left(1-\gamma_{i1}\right)\left[\dfrac{y_{i2}-n_i\pi_{i2}}{\pi_{i2}} - \dfrac{y_{i1}-n_i\pi_{i1}}{\pi_{i1}}\right] \\ -\gamma_{i2}\left(1-\gamma_{i2}\right)\left[\dfrac{y_{i3}-n_i\pi_{i3}}{\pi_{i3}} - \dfrac{y_{i2}-n_i\pi_{i2}}{\pi_{i2}}\right] \\ \vdots \\ -\gamma_{ik-1}\left(1-\gamma_{i\,k-1}\right)\left[\dfrac{y_{ik}-n_i\pi_{ik}}{\pi_{ik}} - \dfrac{y_{i\,k-1}-n_i\pi_{i\,k-1}}{\pi_{i\,k-1}}\right] \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{i1}\sum_{j=1}^{k-1}\gamma_{ij}\left(1-\gamma_{ij}\right)\left[\dfrac{y_{ij+1}-n_i\pi_{i\,j+1}}{\pi_{i\,j+1}} - \dfrac{y_{ij}-n_i\pi_{ij}}{\pi_{ij}}\right] \\ x_{i2}\sum_{j=1}^{k-1}\gamma_{ij}\left(1-\gamma_{ij}\right)\left[\dfrac{y_{ij+1}-n_i\pi_{i\,j+1}}{\pi_{i\,j+1}} - \dfrac{y_{ij}-n_i\pi_{ij}}{\pi_{ij}}\right] \\ \vdots \\ x_{ip}\sum_{j=1}^{k-1}\gamma_{ij}\left(1-\gamma_{ij}\right)\left[\dfrac{y_{ij+1}-n_i\pi_{i\,j+1}}{\pi_{i\,j+1}} - \dfrac{y_{ij}-n_i\pi_{ij}}{\pi_{ij}}\right] \end{bmatrix}.$$

■

A3. MATRIZ DE INFORMAÇÃO

Para o modelo de odds proporcionais, usando o resultado da equação (4.9), a matriz de informação pode ser escrita como

$$\underset{\sim}{I} = E\left[\underset{\sim}{D}^t \underset{\sim}{u}\,\underset{\sim}{u}^t \underset{\sim}{D}\right] = \underset{\sim}{D}^t E\left[\underset{\sim}{u}\,\underset{\sim}{u}^t\right]\underset{\sim}{D} = \underset{\sim}{D}^t \underset{\sim}{A}\,\underset{\sim}{D} \qquad \text{(A3.1)}$$

onde

$$\underset{\sim}{A} = E\left[\underset{\sim}{u}\,\underset{\sim}{u}^t\right] = E\left[\underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\Sigma^{-}\left(\underset{\sim}{y}-\underset{\sim}{N}\,\pi\right)\left(\underset{\sim}{y}-\underset{\sim}{N}\,\pi\right)^t \Sigma^{-t}\,\underset{\sim}{N}^t\,\underset{\sim}{L}^{*-}\right]$$

$$= \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\Sigma^{-} E\left[\left(\underset{\sim}{y}-\underset{\sim}{N}\,\pi\right)\left(\underset{\sim}{y}-\underset{\sim}{N}\,\pi\right)^t\right]\Sigma^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-}$$

$$= \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\Sigma^{-}\Sigma\Sigma^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-} = \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\left[\Sigma^{-} - \underset{\sim}{N}^{-1}\,\underset{\sim}{1}_{nk}\,\underset{\sim}{1}_{nk}^t\right]\underset{\sim}{N}\,\underset{\sim}{L}^{*-}$$

$$= \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\Sigma^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-} - \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\underset{\sim}{N}^{-1}\,\underset{\sim}{1}_{nk}\,\underset{\sim}{1}_{nk}^t\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-}$$

e, como $\underset{\sim}{L}^{*-t}\,\underset{\sim}{1}_{nk}\,\underset{\sim}{1}_{nk}^t = 0$, segue que $\underset{\sim}{A} = \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\Sigma^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-}$, podendo ser escrita como uma matriz de blocos diagonais, pois

$$\underset{nk\times nk}{\Sigma^{-}} = \begin{bmatrix} \Sigma_1^{-} & 0 & \cdots & 0 \\ 0 & \Sigma_2^{-} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_n^{-} \end{bmatrix}, \qquad \underset{\sim}{L} = \underset{nk\times n(k-1)}{\begin{bmatrix} \underset{\sim}{L^{-}} & \underset{\sim}{0} & \cdots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{L^{-}} & \cdots & \underset{\sim}{0} \\ \vdots & \vdots & & \vdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \cdots & \underset{\sim}{L^{-}} \end{bmatrix}}$$

$$\underset{\sim}{\underset{k\times k}{N_i}} = \begin{bmatrix} n_i & 0 & \cdots & 0 \\ 0 & n_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_i \end{bmatrix}, \qquad \underset{\sim}{\underset{nk\times nk}{N}} = \begin{bmatrix} N_1 & 0 & \cdots & 0 \\ 0 & N_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & N_n \end{bmatrix}$$

e

$$\underset{\sim}{\underset{k\times(k-1)}{L^{-}}} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{bmatrix}.$$

Então,

$$\underset{\sim}{A} = \underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\Sigma^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-} = \left[\begin{array}{c:c:c:c} \underset{\sim}{L}^{-t}\,\underset{\sim}{N}_1\,\Sigma_1^{-}\,\underset{\sim}{N}_1\,\underset{\sim}{L}^{-} & 0 & \cdots & 0 \\ \hdashline 0 & \underset{\sim}{L}^{-t}\,\underset{\sim}{N}_2\,\Sigma_2^{-}\,\underset{\sim}{N}_2\,\underset{\sim}{L}^{-} & \cdots & 0 \\ \hdashline \vdots & \vdots & \ddots & \vdots \\ \hdashline 0 & 0 & \cdots & \underset{\sim}{L}^{-t}\,\underset{\sim}{N}_n\,\Sigma_n^{-}\,\underset{\sim}{N}_n\,\underset{\sim}{L}^{-} \end{array}\right]$$

(A3.2)

onde $\underset{\sim}{A}_i = \underset{\sim}{L}^{-t}\,\underset{\sim}{N}_i\,\underset{\sim}{\Sigma}_i^{-}\,\underset{\sim}{N}_i\,\underset{\sim}{L}^{-} = \underset{\sim}{L}^{-t}\,n_i\Sigma_i^{-}n_i\,\underset{\sim}{L}^{-}$, como segue,

$$\underset{\sim}{A}_i = \begin{bmatrix} n_i\dfrac{\pi_{i1}+\pi_{i2}}{\pi_{i1}\,\pi_{i2}} & -n_i\dfrac{1}{\pi_{i2}} & 0 & \cdots & 0 & 0 \\ -n_i\dfrac{1}{\pi_{i2}} & n_i\dfrac{\pi_{i2}+\pi_{i3}}{\pi_{i2}\,\pi_{i3}} & -n_i\dfrac{1}{\pi_{i3}} & \cdots & 0 & 0 \\ 0 & -n_i\dfrac{1}{\pi_{i3}} & n_i\dfrac{\pi_{i3}+\pi_{i4}}{\pi_{i3}\,\pi_{i4}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & n_i\dfrac{\pi_{i\,k-2}+\pi_{i\,k-1}}{\pi_{i\,k-2}\,\pi_{i\,k-1}} & -n_i\dfrac{1}{\pi_{i\,k-1}} \\ 0 & 0 & 0 & \cdots & -n_i\dfrac{1}{\pi_{i\,k-1}} & n_i\dfrac{\pi_{i\,k-1}+\pi_{ik}}{\pi_{i\,k-1}\,\pi_{ik}} \end{bmatrix}.$$

Dessa forma, a matriz $\underset{\sim}{A}$ pode ser reapresentada como

$$\underset{\sim}{A} = \text{bdiag}\left\{\underset{\sim}{A}_1,\ \underset{\sim}{A}_2,\ldots,\ \underset{\sim}{A}_n\right\} = \begin{bmatrix} \underset{\sim}{A}_1 & 0 & \cdots & 0 & 0 \\ 0 & \underset{\sim}{A}_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \underset{\sim}{A}_n \end{bmatrix}.$$

Por sua vez, a matriz $\underset{\sim}{D}$, definida por $\underset{\sim}{D} = \underset{\sim}{C}\,\underset{\sim}{X}^{*t}$ assume a forma

$$\underset{\sim}{D} = \underset{\sim}{C}\,\underset{\sim}{X}^{*t} = \begin{bmatrix} \underset{\sim}{C}_1 & 0 & \cdots & 0 \\ 0 & \underset{\sim}{C}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \underset{\sim}{C}_n \end{bmatrix} \begin{bmatrix} \underset{\sim}{X}_1^t \\ \underset{\sim}{X}_2^t \\ \\ \underset{\sim}{X}_n^t \end{bmatrix} = \begin{bmatrix} \underset{\sim}{C}_1\,\underset{\sim}{X}_1^t \\ \underset{\sim}{C}_2\,\underset{\sim}{X}_2^t \\ \\ \underset{\sim}{C}_n\,\underset{\sim}{X}_n^t \end{bmatrix} = \begin{bmatrix} \underset{\sim}{D}_1 \\ \underset{\sim}{D}_2 \\ \\ \underset{\sim}{D}_n \end{bmatrix}.$$

Portanto, a matriz de informação definida em (A3.1) fica

$$\underset{\sim}{I} = \underset{\sim}{D}^t\,\underset{\sim}{A}\,\underset{\sim}{D} = \underset{\sim}{D}^t\,\underset{\sim}{L}^{*-t}\,\underset{\sim}{N}\,\underset{\sim}{\Sigma}^{-}\,\underset{\sim}{N}\,\underset{\sim}{L}^{*-}\,\underset{\sim}{D} = \sum_{i=1}^{n}\underset{\sim}{D}_i^t\,\underset{\sim}{A}_i\,\underset{\sim}{D}_i\,. \quad \blacksquare$$

A4. DISTRIBUIÇÃO DA VARIÁVEL $C_{ij} = \sum_{s=1}^{j} Y_{is}$

A variável aleatória definida por $C_{ij} = \sum_{s=1}^{j} Y_{is}$ representa a freqüência de respostas em uma categoria menor ou igual a j, para o i-ésimo tratamento. Na página 33 foi visto que, condicional ao número de observações ou replicações fixo do tratamento i, a distribuição de probabilidade da variável aleatória C_{ij} é uma binomial com parâmetros $\left(n; \gamma_{ij}\right)$. A demonstração desse fato é apresentada abaixo.

Seja $\underset{\sim}{y_i} = \left(y_{i1}, y_{i2}, \ldots, y_{iK}\right)^t$ o vetor de freqüências observadas no i-ésimo tratamento, tal que $\sum_{j=1}^{K} y_{ij} = n$. Se o vetor aleatório $\underset{\sim}{Y_i} = \left(Y_{i1}, Y_{i2}, \ldots, Y_{iK}\right)$ segue uma distribuição multinomial com parâmetros n e $\underset{\sim}{\pi_i} = \left(\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK}\right)$; isto é,

$$\underset{\sim}{Y_i} = \left(Y_{i1}, Y_{i2}, \ldots, Y_{iK}\right)^t \sim M\left(n; \underset{\sim}{\pi_i}\right),$$

então a distribuição de probabilidades condicional de $C_{ij} = \sum_{s=1}^{j} Y_{is} \mid C_{iK} = \sum_{s=1}^{K} Y_{is} = n$ pode ser obtida como segue:

$$P\left[C_{ij} = y \mid C_{iK} = n\right] = P\Big[Y_{i1} = y_{i1},\ Y_{i2} = y_{i2},\ldots,\ Y_{ij} = y_{ij} \cap$$
$$\cap \left(Y_{i\,j+1} = y_{i\,j+1},\ Y_{i\,j+2} = y_{i\,j+2},\ldots,\ Y_{iK} = y_{iK}\right);$$
$$\sum_{s=1}^{j} y_{is} = y \quad \text{e} \quad \sum_{s=j+1}^{K} y_{is} = n - y\Big]$$

$$P\left[C_{ij} = y \mid C_{iK} = n\right] = P\left[\left(\sum_{s=1}^{j} Y_{is} = y\right) \cap \left(\sum_{s=j+1}^{K} Y_{is} = n - y\right)\right]$$

$$= \frac{n}{y!\,(n-y)!}\left(\sum_{s=1}^{j} \pi_{is}\right)^{y}\left(\sum_{s=j+1}^{K} \pi_{is}\right)^{n-y}$$

$$= \binom{n}{y} \gamma_{ij}^{y}\left[1 - \gamma_{ij}\right]^{n-y}$$

onde $\gamma_{ij} = \sum_{s=1}^{j} \pi_{is}$.

Portanto, a distribuição condicional de $C_{ij} = y$, dado o número de replicações do i-ésimo tratamento $C_{iK} = n$, é uma binomial com parâmetros n e γ_{ij}; isto é,

$$C_{ij} = y \mid C_{iK} = n \quad \sim \quad B\left(n\,;\gamma_{ij}\right). \blacksquare$$

ANEXO B: COMPUTACIONAL

B1. AJUSTE DOS DADOS DO EXEMPLO A - PROC LOGISTIC

Os resultados que figuram a seguir foram obtidos no procedimento LOGISTIC do software estatístico SAS, para ajustar um modelo de odds proporcionais aos dados do Exemplo A, através da seqüência de comandos:

```
data pyogenes;
infile 'a:pyogenes.dat';
input x y;
proc logistic data=pyogenes order=data;
model y=x / itprint covb corrb;
run;
```

EXEMPLO A - MODELO DE ODDS PROPORCIONAIS

The LOGISTIC Procedure

Data Set: WORK.PYOGENES
Response Variable: Y
Response Levels: 3
Number of Observations: 1398
Link Function: Logit

Response Profile

Ordered Value	Y	Count
1	1	516
2	2	589
3	3	293

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
X	0.051502	0.221099	0	1.00000

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCP1	INTERCP2	X
0	INITIAL	2962.513617	-0.536085	1.327428	0
1	IRLS	2955.510894	-0.505748	1.357766	-0.589058
2	IRLS	2955.494878	-0.508507	1.362710	-0.602363
3	IRLS	2955.494876	-0.508509	1.362720	-0.602643
4	IRLS	2955.494876	-0.508509	1.362720	-0.602649

Last Change in -2 Log L: 1.653772E-6

Last Evaluation of Gradient

INTERCP1	INTERCP2	X
-0.000055445	-0.00006959	-0.000124149

Score Test for the Proportional Odds Assumption
Chi-Square = 0.3155 with 1 DF (p=0.5743)

EXEMPLO A - MODELO DE ODDS PROPORCIONAIS

The LOGISTIC Procedure

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	2966.514	2961.495	.
SC	2976.999	2977.223	.
-2 LOG L	2962.514	2955.495	7.019 with 1 DF (p=0.0081)
Score	.	.	6.838 with 1 DF (p=0.0089)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCP1	-0.5085	0.0564	81.2265	0.0001	.
INTERCP2	1.3627	0.0674	409.2395	0.0001	.
X	-0.6026	0.2251	7.1683	0.0074	-0.073462

Association of Predicted Probabilities and Observed Responses

Concordant = 6.3%	Somers' D = 0.026
Discordant = 3.8%	Gamma = 0.256
Tied = 89.9%	Tau-a = 0.017
(627689 pairs)	c = 0.513

Estimated Covariance Matrix

Variable	INTERCP1	INTERCP2	X
INTERCP1	0.003183461	0.001576472	-0.002279263
INTERCP2	0.001576472	0.0045376974	-0.003242652
X	-0.002279263	-0.003242652	0.0506658523

Estimated Correlation Matrix

Variable	INTERCP1	INTERCP2	X
INTERCP1	1.00000	0.41478	-0.17947
INTERCP2	0.41478	1.00000	-0.21386
X	-0.17947	-0.21386	1.00000

B2. AJUSTE DOS DADOS DO EXEMPLO A - OLOGIT

Paralelamente ao procedimento LOGISTIC do software estatístico SAS (veja o Anexo B1), o modelo de odds proporcionais também pode ser ajustado aos dados do Exemplo A através do pacote STATA. As rotinas seguintes ilustram os passos necessários para realizar o ajuste.

. use pyogene.dta

. ologit y x,table

(Resultando) ...

Iteration 0: Log Likelihood = -1481.2568,

Iteration 1: Log Likelihood = -1477.7498,

Iteration 2: Log Likelihood = -1477.7474,

Ordered Logit Estimates

Number of obs = 1398
chi2(1) = 7.02
Prob > chi2 = 0.0081

Log Likelihood = -1477.7474

Pseudo R2 = 0.0024

y	Coef.	Std. Err.	t	P>| t |	[95% Conf. Interval]	
x	0.6026492	.2274157	2.650	0.008	.1565362	1.048762
_cut1	-0.5085091	.0563953			(Ancillary parameters)	
_cut2	1.36272	0.0673404				

y	Probability	Observed
1	Pr(xb+u<_cut1)	0.3691
2	Pr(_cut1<xb+u<_cut2)	0.4213
3	Pr(_cut2<xb+u)	0.2096

AS PROBABILIDADES DE RESPOSTA PREDITAS PODEM SER OBTIDAS POR:

. *ologitp cat1 cat2 cat3*

. *list x cat1 cat2 cat3*

(Resultando) ...

x	cat1	cat2	cat3
1	.247655	.433714	.318631
0	.3755431	.4206583	.2037986

B3. AJUSTE DOS DADOS DO EXEMPLO B - PROC LOGISTIC

A seqüência de comandos a seguir foi utilizada no software estatístico SAS, para ajustar o modelo de odds proporcionais aos dados do Exemplo B, mostrados na Tabela 2.6.

```
data quinlan;
infile 'a:quinlan.dat';
input h d l_ b j_ f_ n a i_ e_ m c_ k g o_ y;

proc logistic data=quinlan order=data;
      model y=a b c_ d e_ f_ g h i_ j_ k l_ m n o_ /
            maxiter=100 covb corrb;
      title1 'EXEMPLO B - MODELO DE ODDS';
      title2 'PROPORCIONAIS COM TODOS OS FATORES';
run;

proc logistic data=quinlan order=data;
      model y= e_ g / maxiter=100 covb corrb;
      title1 'EXEMPLO B - MODELO DE ODDS';
      title2 'PROPORCIONAIS COM OS FATORES -E e G';
run;
```

Os resultados foram os seguintes.

EXEMPLO B - MODELO DE ODDS PROPORCIONAIS COM TODOS OS FATORES

The LOGISTIC Procedure

Data Set: WORK.QUINLAN
Response Variable: Y
Response Levels: 3
Number of Observations: 64
Link Function: Logit

Response Profile

Ordered Value	Y	Count
1	1	29
2	2	14
3	3	21

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
A	0	1.007905	-1.00000	1.00000
B	0	1.007905	-1.00000	1.00000
C	0	1.007905	-1.00000	1.00000
D	0	1.007905	-1.00000	1.00000
E	0	1.007905	-1.00000	1.00000
F	0	1.007905	-1.00000	1.00000
G	0	1.007905	-1.00000	1.00000
H	0	1.007905	-1.00000	1.00000
I	0	1.007905	-1.00000	1.00000
J	0	1.007905	-1.00000	1.00000
K	0	1.007905	-1.00000	1.00000
L	0	1.007905	-1.00000	1.00000
M	0	1.007905	-1.00000	1.00000
N	0	1.007905	-1.00000	1.00000
O	0	1.007905	-1.00000	1.00000

WARNING: Convergence was not attained in 100 iterations. Iteration control is available with the MAXITER and the CONVERGE options on the MODEL statement.

EXEMPLO B - MODELO DE ODDS COM OS FATORES -E e G

The LOGISTIC Procedure

Data Set: WORK.QUINLAN
Response Variable: Y
Response Levels: 3
Number of Observations: 64
Link Function: Logit

Response Profile

Ordered Value	Y	Count
1	1	29
2	2	14
3	3	21

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
E_	0	1.007905	-1.00000	1.00000
G	0	1.007905	-1.00000	1.00000

Score Test for the Proportional Odds Assumption

Chi-Square = 3.0343 with 1 DF (p=0.0815)

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	139.270	69.648	.
SC	143.588	78.284	.
-2 LOG L	135.270	61.648	73.622 with 2 DF (p=0.0001)
Score	.	.	43.852 with 2 DF (p=0.0001)

EXEMPLO B - MODELO DE ODDS COM OS FATORES -E e G

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCP1	-0.4804	0.3798	1.6001	0.2059	.
INTERCP2	1.7702	0.5005	12.5063	0.0004	.
E_	-16.4000	0.3539	2147.4068	0.0	-9.113262
G	15.7551	.	.	.	.

Association of Predicted Probabilities and Observed Responses

Concordant = 85.6%		Somers' D = 0.825	
Discordant = 3.1%		Gamma = 0.931	
Tied = 11.4%		Tau-a = 0.536	
(1309 pairs)		c = 0.913	

Estimated Covariance Matrix

Variable	INTERCP1	INTERCP2	E_	G
INTERCP1	0.1442479791	0.0530981198	0.0132038402	.
INTERCP2	0.0530981198	0.2505497218	-0.039947031	.
E_	0.0132038402	-0.039947031	0.1252484851	.
G	.	.	.	.

Estimated Correlation Matrix

Variable	INTERCP1	INTERCP2	E_	G
INTERCP1	1.00000	0.27930	0.09823	.
INTERCP2	0.27930	1.00000	-0.22550	.
E_	0.09823	-0.22550	1.00000	.
G	.	.	.	.

B4. AJUSTE DOS DADOS MODIFICADOS DO EXEMPLO B - PROC LOGISTIC

A seqüência de comandos a seguir foi utilizada no software estatístico SAS. Para ajustar o modelo de odds proporcionais aos dados do Exemplo B, modificados segundo a proposta de Kock et al. (1990), definida na Seção 4.3.2.

```
data quinlan;,
    drop z1-z3;
    array z z1-z3;
    input h d l_ b j_ f_ n a i_ e_ m c_ k g o_ z1-z3;
   do over z;
      y = _i_ ;
      freq=z;
      output;
   end;
   label y='Grau de contracao';
```

(continua ...

... continuação)

```
cards;
-1 -1  1 -1  1  1 -1 -1  1  1 -1  1 -1 -1  1 0.1739 0.1739 3.6522
 1 -1 -1 -1 -1  1  1 -1 -1  1  1  1  1 -1 -1 0.1739 0.1739 3.6522
-1  1 -1 -1  1 -1  1 -1  1 -1  1  1 -1  1 -1 3.6522 0.1739 0.1739
 1  1  1 -1 -1 -1 -1 -1 -1 -1 -1  1  1  1  1 3.6522 0.1739 0.1739
-1 -1  1  1 -1 -1  1 -1  1  1 -1 -1  1  1 -1 1.0435 2.7826 0.1739
 1 -1 -1  1  1 -1 -1 -1 -1  1  1 -1 -1  1  1 3.6522 0.1739 0.1739
-1  1 -1  1 -1  1 -1 -1  1 -1  1 -1  1 -1  1 1.0435 2.7826 0.1739
 1  1  1  1  1  1  1 -1 -1 -1 -1 -1 -1 -1 -1 3.6522 0.1739 0.1739
-1 -1  1 -1  1  1 -1  1 -1 -1  1 -1  1  1 -1 3.6522 0.1739 0.1739
 1 -1 -1 -1 -1  1  1  1  1 -1 -1 -1 -1  1  1 3.6522 0.1739 0.1739
-1  1 -1 -1  1 -1  1  1 -1  1 -1 -1  1 -1  1 0.1739 0.1739 3.6522
 1  1  1 -1 -1 -1 -1  1  1  1  1 -1 -1 -1 -1 0.1739 0.1739 3.6522
-1 -1  1  1 -1 -1  1  1 -1 -1  1  1 -1 -1  1 2.7826 1.0435 0.1739
 1 -1 -1  1  1 -1 -1  1  1 -1 -1  1  1 -1 -1 0.1739 3.6522 0.1739
-1  1 -1  1 -1  1 -1  1 -1  1 -1  1 -1  1 -1 0.1739 2.7826 1.0435
 1 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1 0.1739 0.1739 3.6522
;
proc logistic data=quinlan;
      model y=a b c_ d e_ f_ g h i_ j_ k l_ m n o_ / itprint;
      weight freq;
      title1 'EXEMPLO B - DADOS MODIFICADOS';
      title2 'MODELO DE ODDS PROPORCIONAIS: TODOS FATORES';
run;
proc logistic data=quinlan;
      model y=e_ g / itprint corrb covb;
      weight freq;
      title1 'EXEMPLO B - DADOS MODIFICADOS';
      title2 'MODELO DE ODDS PROPORCIONAIS: FATORES -E e G';
run;
```

Os resultados foram os seguintes.

EXEMPLO B - DADOS MODIFICADOS
MODELO DE ODDS PROPORCIONAIS: TODOS FATORES

The LOGISTIC Procedure

Data Set: WORK.QUINLAN
Response Variable: Y Grau de contracao
Response Levels: 3
Number of Observations: 48
Weight Variable: FREQ
Sum of Weights: 64
Link Function: Logit

Response Profile

Ordered Value	Y	Count	Total Weight
1	1	16	28.000100
2	2	16	14.956400
3	3	16	21.043500

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
A	0	1.166920	-1.00000	1.00000
B	0	1.166920	-1.00000	1.00000
C_	0	1.166920	-1.00000	1.00000
D_	0	1.166920	-1.00000	1.00000
E_	0	1.166920	-1.00000	1.00000
F_	0	1.166920	-1.00000	1.00000
G_	0	1.166920	-1.00000	1.00000
H	0	1.166920	-1.00000	1.00000
I_	0	1.166920	-1.00000	1.00000
J_	0	1.166920	-1.00000	1.00000
K_	0	1.166920	-1.00000	1.00000
L_	0	1.166920	-1.00000	1.00000
M_	0	1.166920	-1.00000	1.00000
N	0	1.166920	-1.00000	1.00000
O_	0	1.166920	-1.00000	1.00000

EXEMPLO B - DADOS MODIFICADOS
MODELO DE ODDS PROPORCIONAIS: TODOS FATORES

The LOGISTIC Procedure

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCP1 C_ G K O_	INTERCP2 D H L_	A E_ I_ M	B F_ J_ N
0	INITIAL	136.592519	-0.251308 0 0 0 0	0.713596 0 0 0	0 0 0 0	0 0 0 0
1	IRLS	84.082915	-0.251308 -0.352735 0.700358 -0.362937 0.015314	0.713596 -0.226646 0.015314 0.015314	-0.352735 -1.153651 -0.362937 0.015314	0.196018 -0.226646 0.015314 0.015314
2	IRLS	68.873219	-0.622101 -0.518171 1.147283 -0.580835 -0.048277	1.297175 -0.320346 0.050052 -0.048277	-0.518171 -1.787975 -0.580835 -0.048277	0.318571 -0.320346 0.050052 -0.048277
3	IRLS	66.392348	-0.924612 -0.645484 1.394733 -0.694284 -0.054965	1.713419 -0.390431 0.056093 -0.054965	-0.645484 -2.175595 -0.694284 -0.054965	0.389302 -0.390431 0.056093 -0.054965
4	IRLS	66.313200	-0.997930 -0.677343 1.446972 -0.715220 -0.054409	1.818068 -0.409757 0.054522 -0.054409	-0.677343 -2.266485 -0.715220 -0.054409	0.409644 -0.409757 0.054522 -0.054409
5	IRLS	66.313130	-0.996460 -0.677688 1.447858 -0.715581 -0.054589	1.816131 -0.409822 0.054593 -0.054589	-0.677688 -2.267502 -0.715581 -0.054589	0.409818 -0.409822 0.054593 -0.054589
6	IRLS	66.313130	-0.996580 -0.677730 1.447964 -0.715633 -0.054601	1.816292 -0.409856 0.054601 -0.054601	-0.677730 -2.267676 -0.715633 -0.054601	0.409855 -0.409856 0.054601 -0.054601

EXEMPLO B - DADOS MODIFICADOS
MODELO DE ODDS PROPORCIONAIS: TODOS FATORES

The LOGISTIC Procedure

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCP1	INTERCP2	A	B
			C_	D	E_	F_
			G	H	I_	J_
			K	L_	M	N
			O_			
7	IRLS	66.313130	-0.996570	1.816278	-0.677730	0.409854
			-0.677730	-0.409854	-2.267671	-0.409854
			1.447963	0.054600	-0.715633	0.054600
			-0.715633	-0.054600	-0.054600	-0.054600
			-0.054600			

Last Change in -2 Log L: 2.2082942E-7

Last Evaluation of Gradient

INTERCP1	INTERCP2	A	B	C_	D
0.000050959	-0.00006004	-0.00002884	-5.75533E-6	-0.00002884	7.314565E-6

E_	F_	G	H	I_	J_
-6.72007E-6	7.314565E-6	0.000021349	-0.00001629	-7.24152E-6	-0.00001629

K	L_	M	N	O_
-7.24152E-6	0.000014733	0.000014733	0.000014733	0.000014733

Score Test for the Proportional Odds Assumption

Chi-Square = 22.6347 with 15 DF (p=0.0922)

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	140.593	100.313	.
SC	144.335	132.124	.
-2 LOG L	136.593	66.313	70.279 with 15 DF (p=0.0001)
Score	.	.	45.851 with 15 DF (p=0.0001)

EXEMPLO B - DADOS MODIFICADOS
MODELO DE ODDS PROPORCIONAIS: TODOS FATORES

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCP1	-0.9966	0.4966	4.0265	0.0448	.
INTERCP2	1.8163	0.5708	10.1233	0.0015	.
A	-0.6777	0.3954	2.9379	0.0865	-0.436022
B	0.4099	0.3913	1.0968	0.2950	0.263682
C_	-0.6777	0.3954	2.9379	0.0865	-0.436022
D	-0.4099	0.3913	1.0968	0.2950	-0.263682
E_	-2.2677	0.4529	25.0697	0.0001	-1.458921
F_	-0.4099	0.3913	1.0968	0.2950	-0.263682
G	1.4480	0.4075	12.6280	0.0004	0.931557
H	0.0546	0.3877	0.0198	0.8880	0.035127
I_	-0.7156	0.3917	3.3380	0.0677	-0.460407
J_	0.0546	0.3877	0.0198	0.8880	0.035127
K	-0.7156	0.3917	3.3380	0.0677	-0.460407
L_	-0.0546	0.3877	0.0198	0.8880	-0.035127
M	-0.0546	0.3877	0.0198	0.8880	-0.035127
N	-0.0546	0.3877	0.0198	0.8880	-0.035127
O_	-0.0546	0.3877	0.0198	0.8880	-0.035127

Association of Predicted Probabilities and Observed Responses

Concordant = 36.7%	Somers' D = 0.000
Discordant = 36.7%	Gamma = 0.000
Tied = 26.6%	Tau-a = 0.000
(768 pairs)	c = 0.500

EXEMPLO B - DADOS MODIFICADOS
MODELO DE ODDS PROPORCIONAIS: FATORES -E e G

The LOGISTIC Procedure

Data Set: WORK.QUINLAN
Response Variable: Y Grau de contracao
Response Levels: 3
Number of Observations: 48
Weight Variable: FREQ
Sum of Weights: 64
Link Function: Logit

Response Profile

Ordered Value	Y	Count	Total Weight
1	1	16	28.000100
2	2	16	14.956400
3	3	16	21.043500

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
E_	0	1.166920	-1.00000	1.00000
G	0	1.166920	-1.00000	1.00000

Maximum Likelihood Iterative Phase

Iter	Step	-2 Log L	INTERCP1	INTERCP2	E_	G
0	INITIAL	136.592519	-0.251308	0.713596	0	0
1	IRLS	97.874590	-0.251308	0.713596	-1.153651	0.700358
2	IRLS	89.668038	-0.446651	1.106017	-1.638369	1.174803
3	IRLS	88.987277	-0.500879	1.274030	-1.859614	1.357711
4	IRLS	88.976791	-0.502786	1.295108	-1.892444	1.388079
5	IRLS	88.976770	-0.502611	1.295609	-1.893637	1.389675
6	IRLS	88.976770	-0.502591	1.295636	-1.893676	1.389721

Last Change in -2 Log L: 0.0000207226

Last Evaluation of Gradient

INTERCP1	INTERCP2	E_	G
0.0001672096	0.0000316368	-0.000123835	0.0002598611

EXEMPLO B - DADOS MODIFICADOS
MODELO DE ODDS PROPORCIONAIS: FATORES -E e G

The LOGISTIC Procedure

Score Test for the Proportional Odds Assumption

Chi-Square = 1.5225 with 2 DF (p=0.4671)

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	140.593	96.977	.
SC	144.335	104.462	.
-2 LOG L	136.593	88.977	47.616 with 2 DF (p=0.0001)
Score	.	.	33.726 with 2 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCP1	-0.5026	0.3423	2.1552	0.1421	.
INTERCP2	1.2956	0.3905	11.0061	0.0009	.
E_	-1.8937	0.3839	24.3311	0.0001	-1.218309
G	1.3897	0.3632	14.6395	0.0001	0.894086

Association of Predicted Probabilities and Observed Responses

Concordant	= 37.5%	Somers' D	= 0.000
Discordant	= 37.5%	Gamma	= 0.000
Tied	= 25.0%	Tau-a	= 0.000
(768 pairs)		c	= 0.500

Estimated Covariance Matrix

Variable	INTERCP1	INTERCP2	E_	G
INTERCP1	0.1172029979	0.0441687398	0.0091872277	-0.000236045
INTERCP2	0.0441687398	0.1525223375	-0.051486003	0.0317962009
E_	0.0091872277	-0.051486003	0.1473838976	-0.082280999
G	-0.000236045	0.0317962009	-0.082280999	0.1319255233

EXEMPLO B - DADOS MODIFICADOS
MODELO DE ODDS PROPORCIONAIS: FATORES -E e G

The LOGISTIC Procedure

Estimated Correlation Matrix

Variable	INTERCP1	INTERCP2	E_	G
INTERCP1	1.00000	0.33035	0.06990	-0.00190
INTERCP2	0.33035	1.00000	-0.34340	0.22415
E_	0.06990	-0.34340	1.00000	-0.59008
G	-0.00190	0.22415	-0.59008	1.00000

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. (1984). **Analysis of Ordinal Categorical Data**. New York, Wiley.

AGRESTI, A. (1986). Discussion of "Testing in industrial experiments with ordered categorical data", by V. N. Nair. **Technometrics. 28**(4): 292-294.

AGRESTI, A. (1990). **Categorical Data Analysis**. New York, Wiley.

AITKIN, M. et al. (1989). **Statistical Modelling in GLIM**. Oxford, Clarendon Press.

ANDERSEN, E.B. (1992). Diagnostics in categorical data analysis. **J. R. Statist. Soc. B. 54**(3): 781-791.

ANDERSON, J.A. (1984). Regression and ordered categorical variables. **J. R. Statist. Soc. B. 46**(1): 1-30.

ANDERSON, J.A. & PHILIPS, P.R. (1981). Regression, discrimination and measurement models for ordered categorical variables. **Applied Statistics. 30**(1): 22-31.

ANÔNIMO (1990). Collor lança programa de produtividade. **Folha de São Paulo, B-16**. São Paulo, 08/11/90.

ARMITAGE, P. (1955). Tests for linear trends in proportions and frequencies. **Biometrics. 11**: 375-385.

ARMITAGE, P. (1974). **Statistical Methods in Medical Research**. New York, Wiley.

BASU, A.P. (1983). Identifiability. Em: JOHNSON, N.E. & KOTZ, S. (Editores**). Encyclopedia of Statistical Sciences**, vol. 4. p.2-6. New York, John Wiley & Sons.

BISHOP, Y.V.V.; FIENBERG, S.E. and HOLLAND, P.W. (1975). **Discrete Multivariate Analysis**. Cambridge, MIT Press.

BOX, G. (1988). Signal-to-noise ratios, performance, criteria, and transformations. **Technometrics. 30**(1): 1-40.

BOX, G. & BISGAARD, S. (1987). The scientific context of quality improvement. Report. n° 25, **Center for Quality and Productivity Improvement**, University of Wisconsin-Madison.

BOX, G.; BISGAARD, S. and FUNG, C. (1988). An explanation and critique of Taguchi's contributions to quality engineering. Report n° 28, **Center for Quality and Productivity Improvement**, University of Wisconsin-Madison.

BOX, G.E.P.; HUNTER, W.G. and HUNTER, J.S. (1978). **Statistics for Experimenters**. New York, John Wiley.

BOX, G. & JONES, S. (1986a). Discussion of "Testing in industrial experiments with ordered categorical data", by V.N. Nair. **Technometrics. 28**(4): 295-301.

BOX, G. & JONES, S. (1986b). An investigation of the method of accumulation analysis. Report n° 19, **Center for Quality and Productivity Improvement**, University of Wisconsin-Madison.

BROSS, I.D.J. (1958). How to use ridit analysis. **Biometrics. 14**: 18-38.

BUNKE, H. & BUNKE, O. (1986). (Editores). **Statistical Inference in Linear Models**. Berlin, John Wiley.

CAMPOS, V.F. (1992). **TQC: Controle da Qualidade Total (No Estilo Japonês)**. Belo Horizonte, Fundação Christiano Ottoni, Escola de Engenharia da UFMG.

COMPUTING RESOURCE CENTER (1992). **Stata Reference Manual**: Release 3. 5th edition. Santa Monica, CA. Vol. 1,2 e 3.

CONOVER, W.J. (1980). **Practical Nonparametric Statistics**. Second Edition. New York, Wiley.

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

CURETON, E.E. (1978). Psychometrics. Em: KRUSKAL, W.H. & TANUR, J.M. (Editores). **International Encyclopedia of Statistics**, p.764-782. New York, The Free Press.

DANIEL, W.W. (1978). **Applied Nonparametric Statistics**. Boston, Houghton Mifflin.

DOBSON, A.J. (1983). **An Introduction to Statistical Modelling**. London, Chapman and Hall.

EVERITT, B.S. (1992). **The Analysis of Contingency Tables**. Second Edition. London, Chapman and Hall.

FLEISS, J.L. (1973). **Statistical Methods for Rates and Proportions**. New York, Wiley.

FLORA, J.D. Jr. (1988). Ridit analysis. Em: JOHNSON, N.E. & KOTZ, S. (Editores). **Encyclopedia of Statistical Sciences**, vol. 8. p.136-139. New York, John Wiley.

GANDOUR, R. (1990). Qualidade é desafio à indústria nacional. **Folha de São Paulo, F4**. São Paulo, 07/11/90.

GRAYBILL, F.A. (1976). **Theory and Application of Linear Model**. North Scituate, Duxbury Press.

GREEN, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives. **J. R. Statist. Soc. B. 46**(2): 149-192.

GRIZZLE, J.E.; STARMER, C.F. and KOCH, G.G. (1969). Analysis of categorical data by linear models. **Biometrics. 43**: 471-476.

HAMADA, M. & WU, C.F.J. (1990). A critical look at accumulation analysis and related methods. **Technometrics. 32**(2): 119-162.

HAMADA, M. & WU, C.F.J. (1986). Should accumulation analysis and related methods be used for industrial experiments? **Technometrics. 28**(4):302-306.

HARREL, F.E. (1986). **"The LOGIST Procedure," SUGI Supplemental Library Guide**. Version 5 Edition, Cary, NC:SAS Institute Inc.

HASTIE, T. & TIBSHIRANI, R. (1987). Non-parametric logistic and proportional odds regression. **Appl. Statist. 36**(2): 260-276.

HASTIE, T.; BOTHA, J.L. and SCHNITZLER, C.M. (1989). Regression with an ordered categorical response. **Statistics in Medicine. 8**: 785-794.

HOLMES, M.C. & WILLIAMS, R.E.O. (1954). The distribution of carriers of *Streptococcus pyogenes* among 2413 healthy children. **J. Hyg. Camb. 52**: 165-179.

ISHIKAWA, K. (1986). **TQC - Total Quality Control: Estratégia e Administração da Qualidade**. São Paulo, IMC Internacional Sistemas Educativos.

KIRKPATRICK, C.H. & ALLING, D.W. (1978). Treatment of chronic oral candidiasis with clotrimazole troches: a controlled clinical trial. **The New England Journal of Medicine. 299**: 1201-1203.

KOCH, G.G. et al. (1990). Strategies and issues for the analysis of ordered categorical data from multifactor studies in industry. **Technometrics. 32**(2): 137-149.

LEAR, C. & STANTON, J. (1985). Contact stain requeriment. **Third Symposium on Taguchi Methods**. Romulus, MI: American Supplier Institute, p.117-134.

LEHMANN, E.L. (1975). **Nonparametrics: Statistical Methods Based on Ranks**. San Francisco, Holden-Day.

LOBOS, J. (1991). **Qualidade! Através das Pessoas**. São Paulo.

McCULLAGH, P. (1979). PLUM: An Interactive Computer Package for Analysing Ordinal Data. Mimeo. Dept. Statistics, University of Chicago, Illinois, 60637, USA.

McCULLAGH, P. (1980). Regression models for ordinal data. **J. R. Statist. Soc. B. 42**(2): 109-142.

McCULLAGH, P. (1986). Discussion of "Testing in industrial experiments with ordered categorical data", by V. N. Nair. **Technometrics. 28**(4): 307.

McCULLAGH, P. & NELDER, J.A. (1989). **Generalized Linear Models**. Second Edition. London, Chapman and Hall.

MONTGOMERY, D.C. (1991). **Design and Analysis of Experiments**. Third Edition. New York, John Wiley.

MOSES, L.E. et al. (1984). Analyzing data from ordered categories. **The New England Journal of Medicine. 111**: 442-448.

NAIR, V.N. (1992). Editor. Taguchi's parameter design: a panel discussion. **Technometrics. 34**(2): 127-161.

NAIR, V.N. (1986). Testing in industrial experiments with ordered categorical data. **Technometrics. 28**(4): 283-311.

PHADKE, M.S. et al. (1983). Off-line quality control in integrated circuit fabrication using experimental design optimization. **The Bell System Technical Journal. 11**: 1273-1309.

PILAGALLO, O. (1993a). Indústrias ainda desprezam o treinamento. **Folha de São Paulo**. São Paulo, 29/06/93.

PILAGALLO, O. (1993b). Crise torna indústria mais eficiente. **Folha de São Paulo**. São Paulo, 29/06/93.

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

QUINLAN, J. (1985). Product improvement by application of Taguchi methods. **Third Supplier Symposium on Taguchi Methods**. American Supplier Institute, Inc., Dearborn, MI.

SAS Institute Inc. (1989). **SAS/STAT User's Guide**. Version 6. Fourth Edition, Vol. 1,2. Cary, NC: SAS Institute, Inc.

SCHWARTZ, G. (1991). Pesquisa mostra que indústria brasileira está atrasada vinte anos. **Folha de São Paulo - Caderno 3**. São Paulo, 23/02/91.

SNEDECOR, G.W. & COCHRAN, W.G. (1967). **Statistical Methods**. Sixth Edition. Ames, The Iowa State University Press.

SNEDECOR, G.W. & COCHRAN, W.G. (1980). **Statistical Methods**. Eigth Edition. Ames, The Iowa State University Press.

TAGUCHI, G. (1987). **System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs**. New York, UNIPUB/Klaus International, White Plains, Vol. 1 e Vol. 2.

TAGUCHI, G. & WU, Y. (1980). **Introduction to Off-Line Quality Control**. Nagoya, Central Japan Quality Control Association.

WHEELWRIGHT, S.C. (1984). Strategic management of manufacturing. **Advances in Applied Business Strategy. 1**:1-15. (Published by JAI Press, Inc.).

YATES, F. (1937). The Design and Analysis of Factorial Experiments. Technical Communication n° 35. Harpenden, Hertfordshire: Imperial Bureau of Soil Science.