UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

MODELAGEM CONJUNTA DE MÉDIA E VARIÂNCIA EM EXPERIMENTOS FRACIONADOS SEM REPETIÇÃO UTILIZANDO GLM

Patrícia Klaser Biasoli

Porto Alegre, 2005.

PATRÍCIA KLASER BIASOLI

MODELAGEM CONJUNTA DE MÉDIA E VARIÂNCIA EM EXPERIMENTOS FRACIONADOS SEM REPETIÇÃO UTILIZANDO GLM

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção como requisito parcial à obtenção do título de MESTRE EM ENGENHARIA DE PRODUÇÃO – Área de Concentração: Sistemas da Qualidade.

Orientador: Flávio Sanson Fogliatto, Ph.D.

Porto Alegre

2005

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção.

Prof. Flávio Sanson Fogliatto, Ph.D.
PPGEP / UFRGS
Orientador

Prof. Luis Antonio Lindau, Ph.D.
Coordenador PPGEP / EE / UFRGS

Banca Examinadora:

João Riboldi, Dr.
Prof. Depto de Estatística / UFRGS

José Luis Duarte Ribeiro, Dr.
Prof. Depto de Engenharia de Produção e Transportes / UFRGS

Liane Werner, Dr.
Prof. Programa de Pós-Graduação em Engenharia de Produção / UFRGS

AGRADECIMENTOS

A realização desse trabalho só foi possível devido a colaboração de diversas pessoas. Estou consciente que não contemplarei todos que fizeram parte dessa jornada, mas devo meus sinceros agradecimentos a todos que participaram direta e indiretamente da realização desse trabalho. Dentre esses gostaria de agradecer especialmente:

ao meu namorado, Andrei, pelo compreensão e paciência nos momentos difíceis; leitura e formatação do dessa dissertação;

aos meus pais e ao meu irmão, por compreenderem a minha irritação;

ao Prof.. Ph.D. Fogliatto, por sua dedicação e pelas suas valiosas orientações;

ao Prof . Dr. Riboldi, pela sua contribuição na parte de modelagem estatística;

ao Prof. Dr. Amaral pelo seu apoio e disposição em me auxiliar;

a tia Clara pela imprescindível correção;

aos colegas Ângelo e Mariana pela ajuda na programação dos pacotes estatísticos SAS e R;

ao LOPP/PPGEP por me conceder espaço físico para realização desse trabalho; e

as minhas colegas de mestrado (Mê, Jú, Lú, Mari, Fabi e Tati), pelos estudos em grupo e pelos momentos de descontração dentro e fora do LOPP.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

RESUMO

A modelagem conjunta de média e variância tem se mostrado particularmente relevante na obtenção de processos e produtos robustos. Nesse contexto, deseja-se minimizar a variabilidade das respostas simultaneamente com o ajuste dos fatores, tal que se obtenha a média da resposta próxima ao valor alvo. Nos últimos anos foram desenvolvidos diversos procedimentos de modelagem conjunta de média e variância, alguns envolvendo a utilização dos Modelos Lineares Generalizados (GLMs) e de projetos fatoriais fracionados. O objetivo dessa dissertação é apresentar uma revisão bibliográfica sobre projetos fatoriais fracionados e GLM, bem como apresentar as propostas de modelagem conjunta encontradas na literatura. Ao final, o trabalho enfatiza a proposta de modelagem conjunta de média e variância utilizando GLM apresentada por Lee e Nelder (1998), ilustrando-a através de um estudo de caso.

Palavras chave: GLM, Fatorial fracionado, Modelagem conjunta de média e variância

ABSTRACT

The joint analysis of responses' mean and dispersion has been particularly relevant to obtain robust processes and products. In this context, it is expected to minimize the variability of the responses simultaneously with the adjustment of the factors, so that it gets close to the target value. In the last years, diverse procedures to joint modeling of mean and dispersion have been developed, some of them proposing the use of Generalized Linear Models (GLMs) and fractional factorial experiments. The objective of this tesis is to present a literature review on fractional factorial experiments and GLM, as well as to introduce proposals of joint modeling available in the literature. Finally, it is emphasized the proposal for the joint modeling of mean and dispersion using GLM presented by Lee and Nelder (1998), which is illustrated in a case study.

Key words: GLM, Fractional factorial, Join analysis of the mean and the dispersion

1 COMENTÁRIOS INICIAIS

1.1 INTRODUÇÃO

A modelagem conjunta de média e variância tem se mostrado relevante em estudos, onde o objetivo é obter processos e produtos robustos. Essa modelagem é utilizada para otimizar a variável resposta. Nela, deseja-se minimizar a variabilidade das respostas simultaneamente com o ajustamento dos fatores, de forma a se obter a média da variável resposta próxima a um valor alvo pré-determinado. Nos últimos anos, foram desenvolvidos diversos procedimentos de modelagem conjunta de média e variância, dentre eles a utilização de GLM (*Generalized Linear Models* – Modelos Lineares Generalizados) e de projetos fatoriais fracionados.

Segundo Bergman e Hynén (1997), tipicamente fatores com efeito de dispersão são ajustados para que se obtenha uma variância mínima da variável resposta em torno do valor alvo (média); já fatores com efeito de localização são utilizados para ajustar produtos e processos no seu valor alvo; por fim, fatores sem efeito sobre a média ou variância são ajustados em seus níveis econômicos.

A modelagem conjunta da média e variância tem se mostrado bastante útil no contexto atual de mercado, em que exigências por otimização de produtos e processos, redução dos custos e melhoria da qualidade e produtividade se fazem crescentes. Níveis ótimos para as variáveis resposta, as quais geralmente correspondem a características dos produtos relevantes em termos mercadológicos, são definidos como aqueles que asseguram a sua variação mínima frente a ruídos e colocam o processo no alvo (CATEN, 1995).

A necessidade da modelagem do efeito de fatores de controle sobre a variabilidade das variáveis resposta foi originalmente proposta por Taguchi, no contexto de planejamentos robustos - *signal-response* (TAGUCHI; ELSAYED; HSIANG, 1990). O objetivo dos planejamentos robustos de Taguchi é identificar uma combinação de níveis dos fatores de controle com efeito sobre a dispersão que minimize a variabilidade. Simultaneamente, deseja-se uma combinação de níveis dos fatores de controle com efeito sobre a localização que assegure uma média próxima a um valor alvo para as variáveis resposta. Apesar da relevância da proposta de Taguchi, elas têm recebido críticas, já que resulta em experimentos com um grande número de rodadas e com matrizes experimentais que desconsideram a importância das interações entre os fatores controláveis (GUNTER, 1987).

Com o intuito de aperfeiçoar a proposta de modelagem conjunta de média e variância, inicialmente desenvolvida por Taguchi, em particular com relação às limitações apontadas, alguns autores sugeriram procedimentos baseados na utilização de projetos fatoriais fracionados, com dados modelados através de GLM. Tais procedimentos são o assunto principal desta dissertação (doravante, o efeito dos fatores de um projeto experimental sobre a média será designado por *efeito de localização* e o efeito dos fatores sobre a variância será designado por *efeito de dispersão*).

1.2 OBJETIVO

O objetivo principal desta dissertação é apresentar um roteiro prático para a modelagem conjunta de média e variância em experimentos fracionados sem repetição, utilizando GLM.

Este trabalho tem como objetivos secundários:

- apresentar diferentes propostas para identificação de efeitos significativos em experimentos e para modelagem conjunta da média e variância e suas respectivas limitações;

- exemplificar a utilização do roteiro prático em um estudo de caso, com dados provenientes de um trabalho já publicado; dessa forma, será possível comparar os resultados da otimização obtidos pelo autor com os resultados encontrados utilizando GLM.

1.3 JUSTIFICATIVA

Inicialmente, a escolha do tema a ser estudado foi motivada por sua importância acadêmica, uma vez que o assunto não se encontra exaustivamente explorado na literatura, especialmente no Brasil. Sendo assim, o estudo visa a ser mais um referencial teórico de modelagem conjunta de média e variância.

As condições para o uso de análise de regressão tradicional são restritivas. Esses modelos são baseados na suposição de que as variáveis resposta são Normalmente distribuídas. Entretanto, existem situações em que se deseja modelar dados discretos (tal como contagem de pacientes doentes), ou dados com respostas binárias (peças com defeito e sem defeito). Além disso, há situações onde se deseja modelar dados contínuos, que não são Normalmente distribuídos.

Será demonstrado nesta dissertação que os Modelos Lineares Generalizados (GLM) foram desenvolvidos para permitir o ajustamento de modelos de regressão a uma variável resposta pertencente à Família Exponencial de distribuições, que contempla, além da distribuição Normal, as distribuições Binomial, Geométrica, Binomial Negativa, Exponencial, Gama e Normal Inversa. Ou seja, os GLM possuem maior flexibilidade de aplicação, uma vez que nem todos os fenômenos podem ser bem modelados supondo distribuição Normal ou através da transformação dos dados (com a finalidade de obter dados Normalmente distribuídos).

Do ponto de vista prático, o assunto se torna relevante devido a seu potencial de aplicação em empresas de manufatura. Três razões são fundamentais para sustentar esta pesquisa: (*i*) os artigos são de difícil compreensão para as empresas; (*ii*) em estudos de aplicação da metodologia Seis Sigma, onde se busca processos centrados, muitas vezes se ignora o efeito da alta variabilidade dos dados sobre os processos, e (*iii*) em experimentos onde a média é função da variância, quando a média aumenta a variância também aumenta, exigindo, assim, a modelagem conjunta destes parâmetros.

O estudo de caso que ilustra o roteiro proposto nesta dissertação utiliza dados de experimentos já realizados e publicados em uma dissertação. A utilização de dados já coletados justifica-se uma vez que estudos em campo são caros e demorados para as empresas. Além disso, exigem um planejamento do experimento e da disponibilidade da empresa em realizar o estudo. Outro empecilho é o fato de não ser possível prever, *a priori*, se

a média será função da variância em um dado experimento; assim, corre-se o risco de utilizar recursos para coletar dados que talvez não viabilizem a modelagem conjunta da média e variância da variável resposta. A partir de dados que já foram analisados em estudos de otimização similares, esse problema deixa de existir, sendo ainda possível comparar os resultados obtidos inicialmente com os resultados utilizando GLM. Dessa forma, será possível ilustrar as vantagens e desvantagens da utilização do GLM para modelagem conjunta de média e variância.

1.4 METODOLOGIA

O método que foi desenvolvido neste trabalho é de natureza aplicada com uma abordagem quantitativa do problema. Segundo Gil *apud* Silva (2000), do ponto de vista dos objetivos, este trabalho é uma pesquisa explicativa e através dela é possível identificar fatores que determinam ou contribuem para a ocorrência dos fenômenos. Ou seja, aprofunda o conhecimento da realidade, uma vez que explica a razão dos fatos (SILVA, 2000).

Utilizou-se a pesquisa bibliográfica e estudo de caso como procedimentos técnicos. Segundo Yin (2001), o estudo de caso visa examinar acontecimentos contemporâneos dentro de um contexto da vida real e lida com uma variedade de evidências, tais como: documentos, entrevistas e observações. Um estudo de caso é composto pelas seguintes etapas (YIN, 2001):

- definição de um projeto de pesquisa, em que se define as questões em estudo, as proposições (se houverem), a unidade de análise, a lógica que une os dados às proposições e os critérios para interpretar os resultados;
- desenvolvimento de proposições teóricas, caso o propósito decorrente do estudo de caso seja determinar ou testar a teoria;
- coleta de dados, que determinará o sucesso do estudo;
- análise dos dados, que consiste em examinar categorizar e classificar as informações coletadas; e
- elaboração de relatório para apresentação dos resultados.

O método de trabalho foi desenvolvido a partir de quatro etapas. A primeira etapa envolveu uma revisão bibliográfica a respeito de experimentos fatoriais fracionados e GLM. Efetuou-se, também, o levantamento de propostas de modelagem necessários à aplicação da metodologia proposta. A revisão bibliográfica foi baseada em artigos científicos e livros.

A segunda etapa consistiu na apresentação de procedimentos de modelagem conjunta relacionados à solução do problema de pesquisa proposto. A partir de informações oriundas dos estudos das propostas de modelagem, propõe-se um roteiro para conduzir pesquisadores com problemas similares de análise, constituindo a terceira etapa deste estudo.

Por fim, após o levantamento bibliográfico e organização escrita das técnicas necessárias para a modelagem conjunta, desenvolveu-se a implementação da mesma em um estudo de caso, a fim de exemplificar a metodologia de modelagem conjunta de média e variância utilizando GLM.

1.5 LIMITAÇÕES DO TRABALHO

Neste trabalho não são abordadas situações nas quais deseja-se modelar mais de uma variável resposta. Também não faz parte do escopo deste trabalho a modelagem conjunta de média e variância para experimentos fatoriais completos.

O estudo parte do princípio que o leitor possui conhecimentos em Planejamentos de Experimentos fatoriais fracionados e fatoriais blocados, além de conhecimentos em fatoriais vinculados (MONTGOMERY, 2001).

A metodologia proposta é aplicada em dados já coletados, sendo assim, podendo ser restrita a este contexto de aplicação. Utilizam-se dados de desempenho em campo, não sendo simulados dados em laboratório. Além disso, a otimização é restrita, pois não foram testadas na prática.

1.6 ESTRUTURA DA DISSERTAÇÃO

O presente trabalho foi organizado em cinco capítulos, cujos conteúdos estão delineados a seguir.

No primeiro capítulo, encontram-se as considerações iniciais, objetivos e metodologia de pesquisa empregada. É feita uma introdução ao tema e explicita-se a sua relevância, tanto para o meio acadêmico como para o meio profissional. São também apresentadas as limitações do trabalho.

No segundo capítulo, apresenta-se uma revisão bibliográfica sobre experimentos fatoriais fracionados e GLM. Neste também se apresenta as propostas de identificação de efeitos de localização e dispersão significativos, além de metodologias para modelagem conjunta encontradas na literatura.

No terceiro capítulo, propõe-se um roteiro de modelagem conjunta de média e variância utilizando GLM.

O quarto capítulo focaliza uma aplicação do modelo em um estudo de caso já publicado.

O quinto capítulo expõe os comentários finais, com apresentação das conclusões do trabalho, juntamente com sugestões para futuros trabalhos.

2 REVISÃO BIBLIOGRÁFICA

2.1 FATORIAIS FRACIONADOS

A utilização de planejamento de experimentos nas indústrias está relacionada à melhoria do processo de manufatura (DAVIES; HAY, 1950; BOX; MEYER, 1986). O objetivo desses estudos é geralmente a busca pelo aumento da produção ou da qualidade do produto, ou uma produção mais econômica (DAVIES; HAY, 1950). Essas questões envolvem a análise de diferentes fatores e o problema é determinar qual a melhor forma de realizar o experimento. Fatores em um experimento são investigados mediante variação de seus níveis, os quais normalmente são definidos *a priori* pelo analista. Em um experimento fatorial completo, por exemplo, todas as combinações de níveis dos fatores são examinadas, o que pode ser inviável na prática, por ser caro e demorado.

Um fatorial 2^k completo requer que todas as combinações dos dois níveis dos k fatores sejam testadas experimentalmente (BOX; HUNTER, 2000). Assim, o número de ensaios aumenta rapidamente à medida que aumenta o número de fatores. Por exemplo, uma repetição completa de um fatorial 2^6 requer 64 ensaios. Neste projeto, apenas 6 dos 63 graus de liberdade disponíveis correspondem aos efeitos principais e 15 graus de liberdade correspondem às interações de primeira ordem. Os 42 graus de liberdade restantes correspondem a interações de maior ordem, as quais, via de regra, não são de interesse do analista, já que são de difícil interpretação física (MONTGOMERY, 2001).

Segundo Box; Hunter e Hunter (1978), os efeitos em um experimento possuem certa hierarquia. Em termos de magnitude absoluta, os efeitos principais tendem a ser maiores que as interações de 2 fatores, as quais tendem a ser maiores que as interações de 3 fatores, e

assim por diante. Logo é razoável pressupor que interações de maior ordem não sejam significativas, o que permitiria obter informações acerca dos efeitos principais e interações de baixa ordem de interesse a partir de uma fração do experimento fatorial completo. Um projeto com essas características é denominado projeto fatorial fracionado.

Projetos fatoriais completos são, geralmente, utilizados para estudar o efeito de muitos fatores (BERGMAN; HYNÉN, 1997), o que demanda muitos ensaios. Entretanto, em projetos fracionados, apenas uma parte dos ensaios é executada. Geralmente, isso não implica em perda significativa de informação, principalmente quando o número de fatores aumenta (MONTGOMERY, 2001). Esse tipo de projeto experimental é o mais utilizado e viável na prática, uma vez que reduz os custos e o tempo de execução do experimento devido ao pequeno número de ensaios demandados.

Segundo Box e Hunter (2000), os fatoriais fracionados são utilizados em diferentes circunstâncias: (*i*) quando se assume, *a priori*, que algumas interações não são significativas; (*ii*) quando se deseja identificar quais variáveis têm influência sob a variável resposta, sem um maior detalhamento sobre a forma do efeito (em experimentos do tipo *screening*); (*iii*) quando o procedimento de experimentação é realizado iterativamente, de tal forma que ambigüidades e erros de estimação possam ser resolvidos em um próximo experimento; e (*iv*) quando o analista é capaz de priorizar os fatores de controle em importância, detalhando o efeito apenas de fatores prioritários (analisando suas interações) e limitando-se a apenas verificar o efeito principal de fatores menos importantes.

Um projeto fatorial 2^k fracionado é usualmente designado por 2^{k-p}, onde k indica o número de fatores e p, o grau de fracionamento. Por exemplo, um fatorial fracionado 2^{3-1} é implementado em quatro rodadas experimentais e corresponde a um fatorial 2^3 (que exige oito combinações) fracionado ao meio, ou seja:

$$\frac{1}{2}2^3 = 2^{-1}2^3 = 2^3 2^{-1} = 2^{3-1}. \qquad (1)$$

De uma forma simplificada, o procedimento para definir projetos fracionados consiste em dividir o projeto completo em dois ou mais blocos, confundindo uma ou mais

interações de ordem superior com fatores principais ou interações de menor ordem. Posteriormente, deve-se executar apenas um dos blocos escolhido aleatoriamente.

Os efeitos confundidos devido ao fracionamento vão gerar efeitos vinculados, ou seja, não será possível distinguir o efeito de dois ou mais fatores na análise estatística dos dados. Assim, recomenda-se que um efeito importante seja vinculado a uma interação de ordem superior (supostamente não significativa).

Davies e Hay (1950) recomendam a determinação dos contrastes de definição e a utilização do método definido por Finney para obter a lista completa dos vínculos. Um vínculo definido por D = ABC significa que tanto o fator D como a interação ABC não poderão ser separados na análise estatística e a comparação correspondente pode ser usada para estimar D apenas quando a interação ABC não é significativa, ou seja, os efeitos D e ABC são vinculados.

Tabela 1: Exemplo de fatorial fracionado 2^{5-1}

ordem	rodada	A	B	C	D	E	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE	Y
1	17	-	-	-	-	+	+	+	+	-	+	+	-	+	-	-	56
2	2	+	-	-	-	-	-	-	-	-	+	+	+	+	+	+	53
3	3	-	+	-	-	-	-	+	+	+	-	-	-	+	+	+	63
4	20	+	+	-	-	+	+	-	-	+	-	-	+	+	-	-	65
5	5	-	-	+	-	-	+	-	+	+	-	+	+	-	-	+	53
6	22	+	-	+	-	+	-	+	-	+	-	+	-	-	+	-	55
7	23	-	+	+	-	+	-	-	+	-	+	-	+	-	+	-	67
8	8	+	+	+	-	-	+	+	-	-	+	-	-	-	-	+	61
9	9	-	-	-	+	-	+	+	-	+	+	-	+	-	+	-	69
10	26	+	-	-	+	+	-	-	+	+	+	-	-	-	-	+	45
11	27	-	+	-	+	+	-	+	-	-	-	+	+	-	-	+	78
12	12	+	+	-	+	-	+	-	+	-	-	+	-	-	+	-	93
13	29	-	-	+	+	+	+	-	-	-	-	-	-	+	+	+	49
14	14	+	-	+	+	-	-	+	+	-	-	-	+	+	-	-	60
15	15	-	+	+	+	-	-	-	-	+	+	+	-	+	-	-	95
16	32	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	82

Adaptado: Box, Hunter e Hunter (1978)

Segundo o exemplo proposto por Box; Hunter e Hunter (1978), um fatorial fracionado 2^{5-1} pode ser construído da seguinte maneira: inicialmente, deve-se escrever o fatorial 2^4 completo para as 4 variáveis A, B, C e D. A coluna de sinais da interação ABCD

deve ser utilizada para definir os níveis da variável E. Os dados desse experimento são apresentados na Tabela 1, onde *Y* designa o valor observado para a variável resposta.

Observa-se que dessa forma é possível estimar 16 efeitos (a média, 5 efeitos principais e 10 interações de 1ª ordem). Entretanto, faltam 16 efeitos que poderiam ser estimados utilizando o fatorial completo (10 interações de 2ª ordem, 5 de 3ª ordem e 1 de 4ª ordem).

Box; Hunter e Hunter (1978) mostram que com o fracionamento não há perda de informações através do exemplo: para estimar o efeito da interação ABC, multiplicam-se as respectivas colunas (A, B e C), obtendo os seguintes resultados (Figura 1):

Efeitos	**Tratamentos**															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ABC	-	+	+	-	+	-	-	+	-	+	+	-	+	-	-	+

Figura 1: Sinais do efeito da interação ABC

Observa-se que esses resultados da Figura 1 são idênticos aos da coluna DE da Tabela 1. Conclui-se, assim, que ABC = DE, ou seja, apresentam os mesmos sinais para todos os fatores e conseqüentemente as interações ABC e DE estão vinculadas. Equivalentemente, em um fatorial fracionado 2^{5-1}, as interações ABC e DE individualmente são ditas vinculadas uma a outra.

Usualmente, utiliza-se a notação l_{DE} para indicar a função linear das observações a qual é utilizada para estimar a interação DE, cujo resultado é:

$$l_{DE} = \frac{1}{8}\left(-56+53+63-65+53-55-67+61-69+45+78-93+49-60-95+82\right) = -9{,}5$$

Ou seja, o contraste l_{DE} indica a diferença de 2 médias dos 8 resultados. Assim, o contraste l_{DE} estima a soma das médias dos valores dos efeitos DE e ABC. Isto é indicado como $l_{\mathrm{DE}} \rightarrow$ DE + ABC. Se as colunas de sinais correspondentes a todos os efeitos de 2ª, 3ª e 4ª ordem forem obtidos pela multiplicação dos sinais, obtém-se os resultados apresentados na Tabela 2, os quais usam as informações da Tabela 1.

A metodologia de confundimento apresentada anteriormente se justifica se os efeitos de 2ª, 3ª e 4ª ordem não forem importantes para o pesquisador.

Tabela 2: Pares confundidos e estimativas dos efeitos em um experimento 2^{5-1}

Relacão entre os pares de colunas			Pares confundidos			Estimativa
A	=	BCDE	l_A	→	A + BCDE	-2
B	=	ACDE	l_B	→	B + ACDE	20,5
C	=	ABDE	lc	→	C + ABDE	0
D	=	ABCE	l_D	→	D + ABCE	12,25
E	=	ABCD	l_E	→	E + ABCD	-6,25
AB	=	CBE	l_{AB}	→	AB + CDE	1,5
AC	=	BDE	l_{AC}	→	AC + BDE	0,5
AD	=	BCE	l_{AD}	→	AD + BCE	-0,75
AE	=	BCD	l_{AE}	→	AE + BCD	1,25
BC	=	ADE	l_{BC}	→	BC + ADE	1,5
BD	=	ACE	l_{BD}	→	BD + ACE	10,75
BE	=	ACD	l_{BE}	→	BE + ACD	1,25
CD	=	ABE	l_{CD}	→	CD + ABE	0,25
CE	=	ABD	l_{CE}	→	CE + ABD	2,25
DE	=	ABC	l_{DE}	→	DE + ABC	-9,5
(I	=	ABCDE	l_I	→	média + 1/2 (ABCDE)	65,25

Adaptado: Box, Hunter e Hunter (1978)

Dados oriundos de projetos fracionados são tipicamente analisados somente quanto há efeito dos fatores de controle sobre a média da variável resposta, já que a ausência de repetições das rodadas experimentais dificulta a modelagem do efeito de fatores de controle sobre a variância. A partir do trabalho seminal de Box e Meyer (1986), diversos autores propuseram procedimentos para a modelagem da média e variância da variável resposta a partir de dados oriundos de projetos fracionados, ver por exemplo, Ribeiro; Fogliatto e Caten (2001) e McGrath e Lin (2001). Uma dessas abordagens propõe a modelagem conjunta da média e variância da variável resposta utilizando GLM. Abordagens de modelagem conjunta

são discutidas na seção 3 desta dissertação. No restante da presente seção, apresenta-se um tutorial sobre GLM que pode auxiliar na compreensão dos conteúdos abordados na próxima seção.

2.2 MODELOS LINEARES GENERALIZADOS - GLM

Um modelo é dito linear quando é uma função linear de seus coeficientes. Modelos lineares e não lineares são baseados na suposição de que as variáveis resposta são Normalmente distribuídas. Entretanto, em certas situações, deseja-se modelar dados que seguem distribuições discretas, assimétricas, binomiais, dados restritos a um intervalo do conjunto dos reais, entre outros. Os modelos lineares generalizados (GLMs) foram desenvolvidos para permitir o ajustamento de modelos de regressão para uma variável resposta pertencente à Família Exponencial de distribuições, que contempla, além da distribuição Normal, as distribuições Binomial, Geométrica, Binomial Negativa, Exponencial, Gama e Normal Inversa. A definição de Família Exponencial será abordada na continuidade dessa seção.

Sendo assim, o GLM permite modelar variáveis discretas, tal como o número de produtos defeituosos em uma amostra aleatória (distribuição Binomial); já o número de defeitos por lote inspecionado segue a distribuição Poisson; o qual também pode ser modelado por GLM. Variáveis essencialmente positivas, com distribuição assimétrica com cauda à direita como o tempo até a falha de determinados dispositivos (distribuição Gama) também podem ser modeladas pelo GLM. Em tais exemplos, a variância não é constante (como na Normal) e sim funções da média.

Segundo Myers e Montgomery (1997), o GLM é utilizado para estimar modelos de regressão quando os erros não seguem uma distribuição Normal e/ou a suposição de homogeneidade da variância é violada (ou seja, a variância é função da média). Os métodos de Mínimos Quadrados Ordinários e da Máxima Verossimilhança, utilizados para estimar modelos de regressão, pressupõem erros com variância constante. Segundo Myers, Montgomery e Vining (2002), o problema da não-homogenidade da variância ocorre freqüentimente na prática, e geralmente em conjunto com a não Normalidade da variável resposta. Uma alternativa para situações, em que tal suposição é violada, é a utilização dos métodos de Mínimos Quadrados Ponderados ou Generalizados, que levam em consideração a

não-homogeneidade da variância (ou seja, a variância pode não ser a mesma para todas as observações). Em casos onde os erros são não-homogêneos e auto-correlacionados, utiliza-se o método dos Mínimos Quadrados Generalizados; no caso de erros com variância não-homogênea, mas não correlacionados, emprega-se o método dos Mínimos Quadrados Ponderados (MONTGOMERY; PECK, 1991). Observa-se que os métodos dos Mínimos Quadrados lidam com variâncias não-homogêneas, mas não com não-Normalidade. Essa limitação pode ser contornada com a utilização do GLM, que permite trabalhar com erros que pertencem à Família Exponencial. O GLM reconhece que a variância das respostas não é constante e, assim, utiliza o método de Mínimos Quadrados Ponderados como base para estimar os seus parâmetros.

Usualmente, quando os dados não apresentam variância homogênea, utilizam-se transformações, tal como o logaritmo natural da variável resposta, conforme metodologia proposta por Box e Cox (1964). Entretanto, segundo Myers e Montgomery (1997), o modelo baseado na transformação dos dados apresenta problemas nos valores estimados e no intervalo de confiança dos parâmetros. Segundo os autores, no GLM os intervalos de confiança são uniformemente menores, sugerindo um modelo com mais eficiência na predição e estimação. O GLM pode fornecer mais informações sobre as variáveis do que a análise tradicional baseada na transformação dos dados, ou seja, é capaz de detectar mais efeitos significativos. Se a suposição de Normalidade for duvidosa, pode-se analisar os dados via GLM e verificar se o modelo fornece mais informações.

Conforme Vieira (2004), o modelo linear clássico exige três suposições: Normalidade, aditividade e variância constante e o GLM pode resolver os três problemas de forma independente. O GLM considera outras distribuições que não a Normal, não exige variância constante (pode ser função da média) e é possível obter linearidade através de uma função que faz a ligação entre a média da variável resposta e o polinômio linear das variáveis independentes (VIEIRA, 2004).

A classe dos GLMs inclui diversos modelos de ampla aplicação prática, tais como: (*i*) casos especiais de modelos de regressão linear e análise de variância; (*ii*) modelos *logit* e *probit* para respostas quantitativas; e (*iii*) modelos log-linear e modelos de respostas múltiplas para respostas na forma de contagem. Todos esses modelos possuem propriedades em comum, o que permite estudá-los de forma conjunta como uma única classe de modelos (MCCULLAGH; NELDER, 1989).

Todas as distribuições pertencentes à Família Exponencial possuem a mesma função de densidade de probabilidade para a resposta observada *y*, definida como:

$$f(y;\theta,\phi)=\exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right\}. \qquad (2)$$

onde *a*(.), *b*(.) e *c*(.) são funções específicas; o parâmetro θ é o parâmetro de localização natural ou canônico e ϕ é freqüentemente designado como parâmetro de dispersão ou escala (usualmente denominado σ). O parâmetro de dispersão ϕ é suposto conhecido para cada observação (CORDEIRO, 1986). A função $a(\phi)$ é a forma generalizada de $a(\phi)=\phi.\text{w}$, onde *w* é uma constante conhecida (ou seja, um peso conhecido *a priori*). Segundo Azzalini *apud* Costa (2003), o parâmetro ϕ isoladamente não é o responsável pela variabilidade das observações, mas sim o produto $\phi.w$ que varia de observação para observação. As características de algumas distribuições pertencentes à Família Exponencial são apresentadas na Figura 2.

	Normal	**Poisson**	**Binomial**	**Gamma**	**Normal Inversa**
Intervalo de *y*	$(-\infty,+\infty)$	$0(1)\infty$	$\frac{0(1)n}{n}$	$(0,\infty)$	$(0,\infty)$
$a(\ )$	$\phi=\sigma^2$	1	$\frac{1}{n}$	$\phi=\nu^{-1}$	$\phi=\sigma^2$
$b(\ )$	$\frac{1}{2}\theta^2$	e^{θ}	$\ln(1+e^{\theta})$	$-\ln(-\theta)$	$-(-2\theta)^{\frac{1}{2}}$
$c(\ )$	$\frac{1}{2}\left(\frac{y^2}{\phi}+\ln(2\pi\phi)\right)$	$-\ln y!$	$\ln\left[\binom{n}{ny}\right]$	$(\nu-1)\ln(y\nu)+\ln\nu-\ln\Gamma(\nu)$	$-\frac{1}{2}\left(\ln(2\pi\phi y^3)+\frac{1}{\phi y}\right)$
$\mu=E(y)$	θ	e^{θ}	$\frac{e^{\theta}}{(1+e^{\theta})}$	$\frac{1}{\theta}$	$(-2\theta)^{-\frac{1}{2}}$
Função de Variância	1	μ	$\mu(1-\mu)$	μ^2	μ^3
Fonte: McCullagh e Nelder (1989)					

Figura 2: Características de algumas distribuições pertencentes à Família Exponencial

Algumas distribuições que pertencem à Família Exponencial possuem variância constante $(\phi=1)$, tais como a Binomial e de Poisson, exceto em situações de dispersão

excessiva, quando é necessário lidar com um parâmetro de escala ϕ (MYERS; MONTGOMERY; VINING, 2002). É relevante destacar que a distribuição de Weibull, de grande utilização prática, não pertence a Família Exponencial devido à estrutura particular do seu modelo.

No GLM, pode-se mostrar que a média e a variância da resposta *y* pode ser definida, respectivamente, como:

$$E(y) = \mu = \frac{db(\theta)}{d\theta} \tag{3}$$

e

$$Var(y) = \frac{d^2 b(\theta)}{d\theta^2} a(\phi). \tag{4}$$

A parte da variância de Y que não depende de $a(\phi)$ é dada por: $Var(\mu) = \frac{Var(y)}{a(\phi)} = \frac{d^2 b(\theta)}{d\theta^2} = \frac{d\mu}{d\theta}$ que representa a parte da variância de *y* que depende da sua média μ. A função $Var(\mu)$ é denominada função de variância. Logo a variância de *y* é o produto de dois fatores, um que depende da média e outro, $a(\phi)$, que não depende (VIEIRA, 2004).

Todo modelo de GLM é definido por três componentes:

- **Distribuição da variável resposta**: às vezes chamada de estrutura dos erros (denominado componente aleatório); membro da Família Exponencial de distribuições de probabilidade;

- **Preditor linear**: que envolve as variáveis regressoras $x_1, x_2, ..., x_k$, que entram no modelo na forma de um modelo linear, denominado componente sistemático $\eta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$; ou seja, conforme Cordeiro (1986), é um conjunto de variáveis independentes que descrevem a estrutura linear do modelo;

- **Função de Ligação**: que une o preditor linear à média natural da variável resposta, ou seja, segundo Demétrio (2001), faz a ligação entre os componentes aleatórios e o sistemático. A função de ligação define a forma como os efeitos sistemáticos de $x_1, x_2, ..., x_k$ são transmitidos para a média: $\eta = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$.

  Tais componentes são detalhados nas seções que se seguem.

2.2.1 Componente aleatório

Os GLMs podem ser utilizados quando se tem uma única variável resposta *Y* e, associado a ela, um conjunto de variáveis regressoras (explicativas) $x_1, x_2, ..., x_k$. Consideram-se $y_1, y_2, ..., y_n$ observações independentes da variável *Y*, com médias $\mu_1, \mu_2, ..., \mu_n$ (MYERS; MONTGOMERY; VINING, 2002), ou seja, $E(Y_i) = \mu_i$, $i = 1, ..., n$.

As observações y_i são aleatórias (componente aleatório do GLM) e seguem uma distribuição pertencente à Família Exponencial (MYERS; MONTGOMERY; VINING, 2002), com um parâmetro desconhecido e com média de uma distribuição de probabilidade pertencente a tal família (CORDEIRO, 1986). Além disso, assume-se que existe apenas um termo de erro no modelo (MCCULLAGH; NELDER, 1989) e que a variância σ_i^2 $(i = 1, 2, ..., n)$ é função da média μ_i (MYERS; MONTGOMERY; VINING, 2002).

Uma característica importante da Família Exponencial é a forma da variância, a qual pode ser definida como $Var(y) = \phi . Var(\mu)$, onde ϕ é o parâmetro de dispersão e $Var(\mu)$ é a função variância. A função variância descreve a possível dependência entre a média μ e a variância (NELDER; LEE, 1991).

Observa-se que uma vez determinada a distribuição de probabilidade dos dados, implicitamente são definidas a função da variância $Var(\mu)$, que é a parte da variância da resposta *y* que depende da média, e o parâmetro de dispersão ϕ, que não depende da média e é constante para os membros da Família Exponencial.

2.2.2 Preditor linear

As variáveis regressoras (variáveis explicativas) x_1, x_2,...., x_k entram no modelo na forma de uma soma linear dos seus efeitos, dando origem ao vetor de preditores lineares (vetor das médias μ_i), que é a porção sistemática do modelo, definida como (MYERS; MONTGOMERY; VINING, 2002):

$$\eta = \mathbf{x}'\boldsymbol{\beta} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \qquad (5)$$

onde η, chamado preditor linear, é um vetor $n \times 1$; $\mathbf{x}' = (x_1, ..., x_k)$ é um vetor de regressores e $\beta = (\beta_1, \beta_2, ..., \beta_k)'$ é um vetor de k parâmetros desconhecidos, a serem estimados, onde $k < n$. Ou seja, a função linear η dos parâmetros desconhecidos β chama-se preditor linear (CORDEIRO, 1986).

Existem muitas situações nas quais a relação aditiva entre o componente sistemático e o aleatório não ocorre. Além disso, nem sempre é possível supor uma distribuição Normal para o componente aleatório ou, homogeneidade de variâncias. Para generalizar tais suposições, Nelder e Wedderburn (1972) propuseram a utilização de GLMs.

A utilização desse preditor linear é responsável pela designação Modelo Linear Generalizado, atribuída a essa família de modelos. Segundo Cordeiro (1986), a palavra "generalizado" implica em uma distribuição de probabilidade mais ampla que a Normal para uma variável resposta e uma função não linear, conectando a média desta variável com a parte determinística do modelo, o preditor linear.

2.2.3 Função de ligação

O modelo de GLM é dado pela relação entre a distribuição da média (componente aleatório) e os preditores lineares. Essa relação é determinada pela função de ligação. A função de ligação descreve como o valor da esperança de Y_i, ou seja μ_i, está relacionado com o preditor linear.

No caso particular da regressão linear, a função de ligação é a identidade, ou seja, $\eta_i = \mu_i$. Observa-se que a média da resposta *i* é: $\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} ...)$.

O GLM é encontrado através da função de ligação: $\eta_i = g(\mu_i)$, *i* = 1, 2, ..., *n*., onde $g(.)$ é a função de ligação utilizada. Essa função faz a ligação entre a média (componente aleatório) e o preditor linear (porção sistemática do modelo), por meio de uma função conhecida $g(.)$, ou seja, $g(\mu_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ (MYERS; MONTGOMERY; VINING, 2002), onde x_i é o vetor das variáveis regressoras para a *i*-éssima observação e $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos ou coeficientes de regressão.

A função de ligação é responsável pela transformação da média da população (e não dos dados), com o objetivo principal de encontrar uma escala sobre a qual o modelo linear aditivo ocorra (COSTA, 2003). A escolha inadequada da função de ligação pode resultar em uma modelagem imprópria dos dados. A escolha da função de ligação determina a natureza do modelo de GLM a ser utilizado. Nelder e Lee (1998) demonstram que a escolha correta da função de ligação simplifica o modelo ajustado.

Existem diversas possibilidades de escolha da função de ligação. Entretanto, essa escolha depende do problema de modelagem em particular e, pelo menos em teoria, cada observação pode apresentar uma função de ligação diferente (CORDEIRO, 1986).

Se a função de ligação selecionada for igual ao parâmetro de localização da distribuição ($\eta_i = \theta_i$), o preditor linear modela diretamente o parâmetro canônico θ e a função de ligação η_i é denominada de ligação canônica (MCCULLAGH; NELDER, 1989). Segundo Cordeiro (1986), o parâmetro canônico caracteriza a distribuição de probabilidade membro da Família Exponencial. As ligações canônicas para as distribuições de probabilidade mais comuns são apresentadas na Figura 3. Algumas de suas propriedades teóricas mais interessantes estão descritas a seguir.

A utilização de uma ligação canônica freqüentemente resulta em uma escala adequada para a modelagem, com interpretação prática para os parâmetros de regressão, além de vantagens estatísticas teóricas em termos de existência de um conjunto de estatísticas suficientes para os parâmetros $\beta's$ e algumas simplificações no algoritmo de estimação dos parâmetros do modelo (DEMÉTRIO, 2001). Uma estatística suficiente corresponde à maior

redução que os dados podem alcançar, sem qualquer perda de informação relevante para a inferência sobre o parâmetro desconhecido (CORDEIRO, 1986).

Distribuição	Ligação Canônica	
Normal	$\eta_i = \mu_i$	Ligação identidade
Binomial	$\eta_i = \ln\left(\frac{\mu}{1-\mu}\right)$	Ligação logística
Poisson	$\eta_i = \ln(\mu_i)$	Ligação logarítmica
Exponencial	$\eta_i = \frac{1}{\mu_i}$	Ligação recíproca
Gama	$\eta_i = \frac{1}{\mu_i}$	Ligação recíproca
Normal Inversa	$\eta_i = \frac{1}{\mu^2}$	Ligação recíproca ao quadrado
Fontes: Myers, Montgomery e Vining (2002) e McCullagh e Nelder (1989)		

Figura 3: Funções de ligação canônica para algumas distribuições de probabilidade

É importante destacar que sendo a ligação canônica a mais natural a ser considerada, dada a distribuição que caracteriza a variável resposta, isso não implica em descartar funções não-canônicas do menu de opções. A sua escolha é conveniente não apenas por simplificar as estimativas de máxima verossimilhança dos parâmetros do modelo, mas, também, o cálculo do intervalo de confiança da estimativa da média da resposta. Entretanto, a conveniência não implica necessariamente na qualidade do ajustamento do modelo, o que é mais importante.

Escolher a função de ligação é equivalente a determinar o modelo em uma regressão múltipla tradicional. Embora as funções canônicas levem a propriedades estatísticas desejáveis, principalmente no caso de pequenas amostras, não existe nenhuma razão *a priori* para que os efeitos sistemáticos do modelo devam ser aditivos na escala dada por tais funções (MCCULLAGH; NELDER, 1989).

Existem outras funções de ligação que também podem ser utilizadas em GLM (adaptado MCCULLAGH; NELDER, 1989; MYERS; MONTGOMERY; VINING, 2002):

- **Ligação *Probit***: $\eta_i = \phi^{-1}\left[\mu_i\right]$, onde ϕ representa a função de distribuição Normal padronizada acumulada;
- **Ligação *Logit***: $\eta_i = \log\left[\mu_i / (1-\mu_i)\right]$;
- **Ligação complementar *log-log***: $\eta_i = \log\left\{\log\left[1-\mu_i\right]\right\}$; e
- **Ligação da família de potências**: $\eta_i = \begin{cases} \mu_i^{\lambda}, & \lambda \neq 0 \\ \ln\left[\mu_i\right], & \lambda = 0 \end{cases}$.

Conforme Cordeiro (1986), as funções de ligação *Probit*, *Logit* e complementar *log-log* são apropriadas para o modelo Binomial, pois transformam o intervalo (0,1) em (-∞, +∞).

Para aplicar o GLM, é necessário determinar a distribuição da variável resposta e a função de ligação. Essas duas informações não são independentes, pois alguns modelos de GLM são mais apropriados para algumas distribuições do que para outras. Por exemplo, o modelo Binomial exige que o parâmetro *p* obedeça à restrição $0 < p < 1$; logo um modelo que permite *p* negativo ou maior que um não é adequado. No caso da distribuição de Poisson, o modelo não pode permitir valores negativos para o parâmetro μ, logo a função de ligação deve satisfazer essa condição em todo o seu domínio (MCCULLAGH; NELDER, 1989).

Observa-se que a resposta média é definida como: $E\left(y_i\right) = g^{-1}\left(\eta_i\right) = g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$. No caso da regressão linear múltipla, o modelo $\mu_i = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ representa um caso especial onde $g\left(\mu_i\right) = \mu_i$ e a função de ligação usada é denominada ligação identidade (MYERS; MONTGOMERY; VINING, 2002). Segundo Cordeiro (1986), a ligação é denominada identidade no sentido de que os valores esperados dos dados e preditores lineares podem ser qualquer valor real. Assim, um modelo de regressão linear tradicional também pode ser definido com um GLM, no qual se utiliza uma função de ligação linear.

Dependendo da escolha da função de ligação, o GLM pode também incluir um modelo não linear. Por exemplo, se no caso apresentado anteriormente for escolhida a função de ligação logarítmica $g\left(a\right) = \ln\left(a\right)$, ao invés da função identidade, obtém-se

$E(y) = \mu = \exp\{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\}$. A ligação logarítmica tem uma certa relação com o uso da transformação logarítmica da variável resposta, no caso dos modelos de regressão linear tradicionais. Naqueles casos, realiza-se a transformação dos dados; já no GLM, é realizada a transformação da média. Segundo Myers; Montgomery e Vining, (2002), também é importante destacar que a transformação da média não altera a distribuição dos erros, como ocorre na transformação dos dados.

Assim, o GLM pode ser visto como uma unificação dos modelos de regressão linear e não linear, que incorpora distribuições normais e não normais como variável resposta, desde que pertença à Família Exponencial de distribuições de probabilidade. Logo no modelo ajustado por GLM, suas inferências podem ser realizadas como nos modelos tradicionais de regressão (MYERS; MONTGOMERY; VINING, 2002).

Observa-se, assim, que uma decisão importante na modelagem de GLM é a identificação da distribuição de probabilidade, que caracteriza a variável resposta, e da matriz de realizações das variáveis independentes. Deve-se também escolher adequadamente a função de ligação. Essa determinação pode resultar de uma análise dos dados ou pode ser baseada na experiência dos pesquisadores.

Observa-se que, para a especificação do GLM, os parâmetros θ_i (parâmetro canônico ou natural) da Família Exponencial não são de interesse direto (pois existe um para cada observação), mas sim um conjunto menor de parâmetros $\beta_1, ..., \beta_k$ tal que uma combinação linear dos β's seja igual a uma função do valor esperado de Y_i (DEMÉTRIO, 2001). A função de ligação é uma função diferenciável monotônica, ou seja, a função $f(x)$ é monotônica se para quaisquer dois pontos x_1 e x_2 tem-se: $f(x_1) \leq f(x_2)$ e $f(x_1) \geq f(x_2)$. Uma função não precisa ser contínua para ser monotônica. Além disso, a função é dita diferenciável em x_0 se $f(x_0)$ existe, onde $f(x_0)$ é a declividade da reta tangente à função no ponto (SIMMONS, 1987).

O processo de trabalho com GLMs pode ser dividido em três etapas (CORDEIRO, 1986):

- formulação dos modelos, que consiste na identificação da distribuição de probabilidade dos dados e determinação do preditor linear e da função de ligação;

- ajustamento dos modelos; e
- inferência.

As duas últimas etapas serão definidas nas próximas seções.

2.2.4 Estimação do vetor de parâmetros $\boldsymbol{\beta}$

O ajustamento do GLM é determinado pelo vetor de parâmetros $\hat{\boldsymbol{\beta}}$, sendo o método de máxima verossimilhança a base teórica para estimação desses parâmetros. Diferentemente dos modelos de regressão tradicionais, que utilizam os Mínimos Quadrados Ordinários para estimação dos parâmetros do modelo, no GLM a solução das equações normais do sistema formado utiliza Mínimos Quadrados Ponderados (DEMÉTRIO, 2001), pois as equações de máxima verossimilhança são não-lineares, exceto para distribuição Normal, e, portanto, não podem ser resolvidas explicitamente (CORDEIRO, 1986).

Em algumas situações, os estimadores de máxima verossimilhança para os parâmetros $\boldsymbol{\beta}$ no preditor linear η podem ser obtidos por Mínimos Quadrados Ponderados Iterativo. Na solução das equações de máxima verossimilhança, a variável dependente não é *y*, mas *z*, uma forma linearizada da função de ligação aplicada em *y*, e os pesos *W* são funções dos valores ajustados de $\hat{\mu}$. O processo é iterativo, pois tanto a variável dependente ajustada *z* como o peso *W* dependem dos valores ajustados, para os quais apenas as estimativas correntes estão disponíveis (MCCULLAGH; NELDER, 1989,). Ou seja, a solução das equações consiste em calcular repetidamente uma regressão linear ponderada de uma variável modificada *z* sobre *y*, usando uma função peso *W* que se modifica no processo iterativo. O inverso da função peso é igual à covariância de *z* (CORDEIRO, 1986). Segundo Costa (2003). esse processo converge rapidamente (de 3 a 4 interações), exceto em casos de amostras pequenas. O procedimento iterativo exige valores iniciais de $\boldsymbol{\beta}$ que podem ser obtidos das estimativas de μ_i, baseadas nos valores observados y_i (COSTA, 2003).

2.2.5 *Quase*-verossimilhança

A estimação dos efeitos fixos do GLM é baseada na função de verossimilhança; uma extensão para ajustar efeitos aleatórios é a *quase*-verossimilhança (COSTA, 2003). Sendo

assim, a estimação por máxima-verossimilhança é utilizada em GLM quando se supõe independência entre observações pertencentes à Família Exponencial. Entretanto, há situações em que, apesar de as respostas serem independentes, elas não pertencem à Família Exponencial, além de situações onde a variância das respostas é função da média, mas as observações são correlacionadas (MYERS; MONTGOMERY; VINING, 2002).

O uso da *quase*-verossimilhança se aplica no GLM em situações onde as variáveis explicativas são correlacionadas. Essa técnica de estimação também pode ser aplicada quando é necessário realizar inferências de experimentos em que a função de verossimilhança não pode ser construída. Por exemplo, pode existir um modelo razoável com variância conhecida, mas não existem informações que indiquem a distribuição de probabilidade da variável resposta (MYERS; MONTGOMERY; VINING, 2002) e, assim, apenas se define uma função entre a média e a variância da variável resposta (VIEIRA, 2004).

Wedderburn (1974) foi o primeiro a introduzir a noção de *quase*-verossimilhança. A *quase*-verossimilhança se baseia na idéia nos Mínimos Quadrados Ponderados; mais genericamente, nos Mínimos Quadrados Generalizados para o caso das respostas correlacionadas. Para utilização da *quase*-verossimilhança, não é necessário especificar completamente a distribuição de probabilidade da variável resposta, pois a *quase*-verossimilhança é baseada apenas na suposição de forma dos dois primeiros momentos. Wedderburn (1974) demonstra que o uso de Mínimos Quadrados Generalizados produz propriedades assintóticas similares aos estimadores de máxima verossimilhança. Logo, pode-se obter boa eficiência dos estimadores mesmo quando a verossimilhança não é conhecida.

Suponha que $y_i \left(i = 1, 2, ..., n\right)$ seja um conjunto de observações com $E\left(\mathbf{Y}_i\right) = \boldsymbol{\mu}_i$ e $V\left(\mathbf{Y}_i\right) \propto V\left(\mu_i\right)$, em que $V\left(\mu\right)$ seja uma função conhecida. Além disso, suponha que $\boldsymbol{\mu}_i$ seja uma função de um conjunto de parâmetros $\beta_1, \beta_2, ..., \beta_p$. Assim, a função de *quase*-verossimilhança $Q\left(\mu_i, y_i\right)$ é definida conforme (COSTA, 2003):

$$\frac{\partial Q\left(\mu_i, y_i\right)}{\partial \mu_i} = \frac{y_i - \mu_i}{V\left(\mu_i\right)} \tag{6}$$

ou, de forma análoga:

$$Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{y_i - \mu_i'}{V(\mu_i')} d\mu_i'. \qquad (7)$$

O logaritmo da função de verossimilhança é um caso especial da *quase*-verossimilhança. Wedderburn (1974) demonstra que se pode utilizar qualquer função que satisfaça a equação (7) como base para definir um modelo linear generalizado e obter estimativas de β_i pelo procedimento conhecido.

A relação entre a média e a variância de Y_i permite a definição da *quase*-verossimilhança, que é maximizada em relação aos parâmetros β pelo uso iterativo das equações de mínimos quadrados ponderados: $X'W\Delta X\hat{\beta} = X'W\Delta Y$, em que $W = diag(W_i)$, ou seja o algoritmo de solução dessa equação equivale ao cálculo repetido de uma regressão linear ponderada por *W*.

Cordeiro (1986) recomenda que o algoritmo de ajustamento não deve ser aplicado a um GLM isolado, mas a vários modelos de um conjunto bem amplo. Esse conjunto deve ser realmente relevante para o conjunto de dados analisados. Deve-se definir uma família de ligações, considerando diferentes opções para a escala de medição e adicionando (ou retirando) variáveis independentes. O conjunto de modelos propostos deve ser estabelecido pela facilidade de interpretação, boas previsões e conhecimento aprofundado da estrutura de dados.

Myers; Montgomery e Vining (2002) recomendam o trabalho de Carrol e Ruppert (1988) para maiores detalhamentos sobre a técnica de *quase*-verossimilhança para estimação de parâmetros.

2.2.6 *Quase*-verossimilhança extendida

Ao modelar-se a média pela técnica de *quase*-verossimilhança ou utilizando GLM, a variância é modelada como uma função da média, multiplicada por um parâmetro de dispersão constante. Entretanto, quando os fatores afetam a dispersão (ou seja, o parâmetro de dispersão ϕ varia conforme o tratamento), essa relação entre a variância e a média não é suficiente para explicar a variância da resposta. Nesses casos casos existe uma "sobre-

dispersão", explicada pelo parâmetro de dispersão, que não é constante, mas sim função dos fatores. Assim, a técnica de *quase*-verossimilhança ou o GLM não são indicados para modelar a média nos casos em que há efeitos na dispersão. A solução é maximizar a função de *quase*-verossimilhança extendida (QVE), definida por Nelder e Pregibon (1987):

$$-2Q^{+} = \sum_{i=1}^{n} \left\{ \frac{d_i}{\phi_i} + \ln\left[2\pi\phi_i V(y_i)\right] \right\}, \tag{8}$$

onde d_i é o componente da *deviance* do GLM (*quase-deviance*), dado por:

$$d_i = 2\int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)}\, dt, \tag{9}$$

sendo $V(t)$ a função de variância para um valor t da média de y_i.

Conforme Vieira (2004), a QVE pode ser vista como um artifício para os casos onde a função de variância não explica completamente a variabilidade da resposta; quando, então, o parâmetro de dispersão não é constante para cada tratamento (como nos GLMs), mas depende dos fatores. Assim, um modelo de dispersão é então construído para estabelecer esta relação de dependência.

2.2.7 Inferência

Segundo Cordeiro (1986), a etapa de inferência tem como objetivo principal verificar a adequação de um modelo de regressão como um todo, além de realizar um estudo detalhado sobre a presença de valores atípicos. Essas discrepâncias, quando significativas, podem implicar na escolha de um modelo alternativo, isto é, na escolha de outro componente sistemático e/ou distribuição de probabilidade dos dados. Nesta etapa, é necessário verificar a precisão e a interdependência das estimativas, construir intervalos de confiança, testar os parâmetros de interesse, analisar estatisticamente os resíduos e realizar previsões, utilizando o modelo ajustado.

Todas as inferências aplicadas em regressão logística podem ser utilizadas com os mesmos objetivos e aplicações no GLM: o *model deviance* pode ser usado para testar o modelo como um todo; a diferença de *deviance* entre o modelo saturado e o modelo nulo pode ser usada para testar o conjunto de parâmetros do modelo. Por fim, a inferência de Wald pode ser aplicada em testes de hipóteses e na construção de intervalos de confiança para o modelo individual dos parâmetros.

Os métodos de inferência dos GLMs baseiam-se, fundamentalmente, na teoria assintótica de máxima verossimilhança, pois, em geral, não é possível obter distribuições exatas para estimativas e estatísticas de teste. As condições de regularidade que garantem esses resultados são satisfeitas nos GLMs (CORDEIRO, 1986).

Para permitir o teste de hipótese de cada um dos coeficientes do modelo, a estatística de Wald se baseia na Normalidade assintótica dos estimadores de máxima verossimilhança. Se a $H_0 : \beta_j = 0$ for verdadeira, a estatística $\left(\frac{b_j}{se(b_j)} \right)^2$ segue uma distribuição Qui-Quadrado com 1 grau de liberdade para grandes amostras. Outra maneira de realizar inferência, utilizando a estatística de Wald, é através dos intervalos de confiança para a resposta média e para as observações individuais (MYERS; MONTGOMERY; VINING, 2002). Esse teste é usual na seleção de covariáveis do modelo (CORDEIRO, 1986).

No procedimento GENMOD do aplicatvo estatístico SAS, o intervalo de confiança de Wald para a média é calculado da seguinte maneira:

$$\hat{\beta} \pm z_{1-\alpha/2}.\hat{\sigma}\,, \tag{10}$$

onde z é o percentil $(1-\alpha/2)$ da distribuição Normal padrão, $\hat{\beta}$ é o parâmetro estimado e $\hat{\sigma}$ é a estimativa do erro padrão. Quando o teste de hipótese não rejeitar H_0, ou equivalentemente, quando o intervalo de confiança do parâmetro contém o valor zero, pode-se afirmar que o parâmetro testado não é significativo para o modelo.

2.2.7.1 Teste de significância dos coeficientes

Seja $\hat{\boldsymbol{\beta}}$ o vetor dos parâmetros estimados. Para a Família Exponencial, a matriz de covariância de $\hat{\boldsymbol{\beta}}$ é: $\text{cov}(\hat{\boldsymbol{\beta}}) = [\mathbf{X}`\mathbf{W}\mathbf{X}]^{-1}\sigma^2$, onde $\mathbf{W}$ é a matriz diagonal com os seguintes elementos:

$$w_{ii}^{(m)} = \frac{1}{\text{var}(y_i)}\left[\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^{(m)}\right]^2, \qquad (11)$$

onde *m* indica o número da iteração e $i = 1,2,...,n$.

Observa-se que no caso do modelo linear com Mínimos Quadrados Ordinários, a matriz de variância-covariância de $\hat{\boldsymbol{\beta}}$ é $\text{cov}(\hat{\boldsymbol{\beta}}) = [\mathbf{X}`\mathbf{X}]^{-1}\sigma^2$.

Para a seleção de parâmetros em distribuição Normal, é comum utilizar a estatística $t_o = \frac{\hat{\beta}_j}{\sqrt{\text{var}(\hat{\beta}_j)}}$, que segue distribuição *t*-Student com $n - p$ graus de liberdade, onde *n* é o número de observações e *p* é número de parâmetros do modelo. Alguns programas computacionais fornecem esta estatística juntamente com o valor de *p-value* para testar parâmetros em modelos não Normais, auxiliando na decisão de incluir ou não um determinado parâmetro no modelo. Mas para modelos não Normais, a distribuição *t* é uma aproximação da verdadeira distribuição do parâmetro. Segundo Lindey *apud* Vieira (2004), esta aproximação pode não ser boa mesmo com amostras grandes, trazendo, conseqüentemente, resultados enganosos.

Entretanto, Vieira (2004) afirma que a estatística t_0 pode ser útil para identificar coeficientes significativos ou não significativos. Assim, um valor elevado de $|t_0|$, digamos maior que três, é uma indicação de significância, em geral, para qualquer distribuição de probabilidade. Por outro lado, valores pequenos de $|t_0|$, menores que um, indicam não significância em geral para qualquer distribuição.

2.2.8 Medidas de ajustamento

Um GLM é considerado uma boa representação dos dados se explicar a relação variância/média satisfatoriamente e se produzir efeitos aditivos na escala definida pela função de ligação. O modelo também deve ser parcimonioso, ou seja, o número de parâmetros deve ser o menor possível. Segundo Demétrio (2001), deseja-se uma combinação satisfatória da distribuição da variável resposta e da função de ligação que descreva a estrutura linear dos dados.

Segundo McCullagh e Nelder (1989), o ajustamento de um modelo a um conjunto de dados observados y_i pode ser encarado como uma maneira de substituir $\mathbf{y}$ por um conjunto de valores estimados (μ_i) para um modelo com poucos parâmetros. Esses valores não serão exatos, logo é necessário definir um limite para essa discrepância.

2.2.8.1 Tipos de modelos

O modelo saturado possui um parâmetro para cada observação e atribui toda a variabilidade dos y_i à porção sistemática. Assim, o modelo se ajusta perfeitamente aos dados, uma vez que os reproduz, sendo, portanto, não-informativo, já que não resume a informação disponível nos dados. Entretanto, o modelo saturado serve como base para medir a discrepância de um modelo intermediário com *p* parâmetros (MCCULLAGH; NELDER, 1989). Existem dois outros tipos de modelos limitantes, porém, menos extremos, descritos a seguir.

Certos termos devem estar contidos no modelo, dependendo do delineamento experimental utilizado. O modelo contendo apenas esse tipo de parâmetro é denominado modelo minimal, ou seja, é aquele que contém o menor número de termos necessários para o ajustamento. Assim, o modelo que contém o maior número de termos possíveis é denominado modelo maximal. Os termos desses dois tipos de modelos são definidos a partir de conhecimentos a *priori* da estrutura dos dados (DEMÉTRIO, 2001).

Por exemplo, em um experimento em blocos com 2 fatores, têm-se os seguintes modelos:

- **Nulo**: $\eta_i = \mu$
- **Minimal**: $\eta_i = \mu + \beta_l$
- **Maximal**: $\eta_i = \mu + \beta_l + \alpha_j + \gamma_k + (\alpha\gamma)_{jk}$
- **Saturado**: $\eta_i = \mu + \beta_l + \alpha_j + \gamma_k + (\alpha\gamma)_{jk} + (\beta\alpha)_{lj}(\beta\gamma)_{lk} + (\beta\alpha\gamma)_{ljk}$,

onde μ é o efeito associado à média geral; β_l é o efeito associado ao bloco *l*; α_j é o efeito associado ao *j*-ésimo nível do fator A; γ_k é o efeito associado ao *k*-ésimo nível do fator B e $(\alpha\gamma)_{jk}, (\beta\alpha)_{lj}, (\beta\gamma)_{lk}$ *e* $(\beta\alpha\gamma)_{ljk}$ são os efeitos associados às interações.

O modelo saturado inclui todas as interações com os blocos, as quais nem sempre são de interesse prático. Neste contexto, qualquer modelo com *p* parâmetros linearmente independentes, situado entre os modelos minimal e maximal, é chamado modelo corrente ou sob pesquisa. Segundo Demétrio (2001), o problema é determinar a utilidade da inclusão de um parâmetro e a falta de ajustamento induzida pela omissão dele. A fim de discriminar esses modelos, medidas de discrepância devem ser utilizadas para medir o ajustamento de um modelo.

2.2.8.2 Deviance

Para testar a significância dos coeficientes do modelo, Athinson e Riani *apud* Vieira (2004) e McCullagh e Nelder (1989) recomendam a estatística *deviance*. A *deviance* está para o GLM assim como a soma de quadrados está para o método dos mínimos quadrados.

Segundo Nelder e Lee (1991), a *deviance* pode ser utilizada para comparar modelos com diferentes preditores lineares e/ou funções de ligação. Entretanto, a *deviance* não pode ser utilizada na comparação de modelos com diferentes funções de ligação ou estrututa de dispersão.

A *deviance* pode ser definida como o logaritmo das razões de verossimilhança, isto é:

$$D(\boldsymbol{\beta}) = -2\ln\left[\frac{\zeta(\boldsymbol{\beta})}{\zeta(\boldsymbol{\mu})}\right] \tag{12}$$

onde $\zeta(\boldsymbol{\beta})$ é a verossimilhança do modelo observado (também denominado modelo nulo) e é função apenas dos dados. O modelo nulo possui apenas um parâmetro, representado pela média comum para todas as observações y_i. Esse modelo atribui toda a variação entre as observações y_i ao componente aleatório. Neste contexto, $\zeta(\boldsymbol{\mu})$, por sua vez, denota a verossimilhança do modelo saturado ou completo, que é o modelo para o qual os valores ajustados $\hat{\mu}_i$ são iguais às respostas y_i, assim, o modelo saturado contém *n* parâmetros, um para cada observação. Neste modelo, todo a variação dos y_i é alocada para o componente sistemático. A Figura 4 apresenta a fórmula para o cálculo da *deviance* das principais distribuições de probabilidade.

Distribuição	**Deviance**
Normal	$\sum_{i=1}^{n}(y_i - \hat{\mu})^2$
Poisson	$2\sum_{i=1}^{n}[y_i \ln(y_i/\hat{\mu}) - (y_i - \hat{\mu})]$
Binomial	$2\sum_{i=1}^{n}\{y_i \ln(y_i/\hat{\mu}) + (n - y_i)\ln[(n - y_i)/(n - \hat{\mu})]\}$
Gama	$2\sum_{i=1}^{n}[-\ln(y_i/\hat{\mu}) + (y_i - \hat{\mu})/\hat{\mu}]$
Normal Inversa	$\sum_{i=1}^{n}(y_i - \hat{\mu})^2/(\hat{\mu}^2 y_i)$
Fonte: McCullagh e Nelder (1989)	

Figura 4: Fórmulas para o cálculo da *deviance*

Sendo assim, o valor de *deviance* é sempre maior ou igual a zero. A medida em que se inclui variáveis explicativas no componente sistemático, o valor da *deviance* decresce até zero (onde zero corresponde ao valor de *deviance* para o modelo saturado). Quanto melhor for o ajustamento do modelo aos dados, menor será o seu valor correspondente de *deviance*. Logo, um modelo bem ajustado tem *deviance* pequena. Uma forma de diminuir o valor da

deviance é incluir mais parâmetros no modelo, aumentando a sua complexidade. Na prática, procuram-se modelos simples com *deviance* moderada (DEMETRIO, 2001).

Assintoticamente, a estatística *deviance* $D(\boldsymbol{\beta})$ segue uma distribuição Qui-Quadrado com $(n-p)$ graus de liberdade, onde *n* é o número de observações e *p* o número de parâmetros do modelo sob investigação. Se a *deviance* for maior que $\chi^2_{(n-p);\alpha}$, o modelo é rejeitado, não sendo significativamente diferente do desempenho obtido para o modelo saturado (MYERS e MONTGOMERY, 1997). Isso implica que as estimações de parâmetros extras no modelo saturado não são necessárias. Esse teste não é adequado para pequenas amostras.

Segundo Myers; Montgomery e Vining (2002), usualmente o ajustamento do modelo aos dados é problemático quando a razão $deviance/(n-p)$ exceder 1 de forma substancial.

2.2.8.3 *Quase*-deviance

Para testar a significância dos coeficientes do modelo ajustado pela função de *quase*-verossimilhança utiliza-se a *quase-deviance*, da mesma forma que a *deviance* é utilizada para a modelagem dos GLMs.

A *quase-deviance* de um modelo é definida como sendo o desvio deste modelo em relação ao modelo saturado, conforme definição abaixo:

$$D_i(y_i,\hat{\mu}_i) = -2\phi[Q_i(y_i,\hat{\mu}_i) - Q_i(y_i,y_i)] = 2\phi[Q_i(y_i,\hat{\mu}_i)] = 2\int_{\mu_i}^{y_i} \frac{y_i-\hat{\mu}_i}{\mathrm{var}(\hat{\mu}_i)}, \qquad (13)$$

onde $Q_i(y_i,\hat{\mu}_i)$ é a função de máxima verossimilhança do modelo a ser analisado e $Q_i(y_i,y_i)$ é a função de máxima verossimilhança do modelo saturado. A análise da *quase-deviance* é equivalente à da *deviance*.

2.2.8.4 Análise de deviance *(ANODEV)*

A análise de variância (ANOVA), particularmente quando aplicada em dados ortogonais com erros Normalmente distribuídos, é uma poderosa técnica para a identificação de efeitos significativos de um modelo. Segundo McCullagh e Nelder (1989), existem dois problemas em aplicar a ANOVA em GLM: (*i*) geralmente os termos do GLM não serão independentes e (*ii*) a soma de quadrados, para dados não Normais, não será uma medida adequada da contribuição do termo para a discrepância total.

Assim, sugere-se a utilização da *deviance* como medida de discrepância para os GLMs elaborando uma tabela similar a tabela da ANOVA para avaliar a contribuição de cada termo no modelo final. Entretanto, a interpretação da tabela da análise de *deviance* (ANODEV) é complicada pela não ortogonalidade dos termos. Cada medida da *deviance* representa a variação explicada por esse termo, que elimina o efeito dos termos incluídos antes dele, mas ignora qualquer efeito dos termos incluídos posteriormente. Assim, é necessário considerar diferentes seqüências para o modelo, pois cada uma produzirá uma tabela ANODEV diferente.

Suponha o Modelo 1 como sendo saturado, o Modelo 2 com p+1 parâmetros $\beta_o, \beta_1, ...,\beta_{p-1}, \beta_p$ e o Modelo 3, aninhado ao Modelo 2, com p parâmetros $\beta_o, \beta_1, ...,\beta_{p-1}$. Definindo a *deviance* destes dois modelos, respectivamente, tem-se:

$$D_2\left(y_i, \hat{\mu}_i^{(2)}\right) = -2\left[\ln L_2\left(y_i, \hat{\mu}_i^{(2)}\right) - \ln L_1\left(y_i, y_i\right)\right] \text{ e} \qquad (14)$$

$$D_3\left(y_i, \hat{\mu}_i^{(3)}\right) = -2\left[\ln L_3\left(y_i, \hat{\mu}_i^{(3)}\right) - \ln L_1\left(y_i, y_i\right)\right]. \qquad (15)$$

Então, a diferença de *deviance* dos modelos 3 e 2 é

$$D_3\left(y_i, \hat{\mu}_i^{(3)}\right) - D_2\left(y_i, \hat{\mu}_i^{(2)}\right) = -2\left[\ln L_3\left(y_i, \hat{\mu}_i^{(3)}\right)\right] + 2\left[\ln L_2\left(y_i, \hat{\mu}_i^{(2)}\right)\right] = 2 - \ln\left[\frac{L_3\left(y_i, \hat{\mu}_i^{(3)}\right)}{L_2\left(y_i, \hat{\mu}_i^{(2)}\right)}\right]. \qquad (16)$$

McCullagh e Nelder (1989) afirmam que a diferença de *deviance* segue assintoticamente uma distribuição Qui-Quadrado com *n-p* graus de liberdade, quando o modelo com *p* parâmetros é considerado correto (neste caso, o Modelo 3). Portanto, conforme Lindsey *apud* Vieira (2004), se o Modelo 3 for correto, o quociente abaixo segue a distribuição $F_{1,\, n-p}$.

$$F_0 = \left| \frac{\frac{\left(D_3\left(y_i, \hat{\mu}_i\right) - D_2\left(y_i, \hat{\mu}_i\right)\right)}{1}}{\frac{D_3\left(y_i, \hat{\mu}_i\right)}{n-p}} \right| \sim F_{1.n-p} \qquad (17)$$

Assim, para uma seqüência de *k* modelos aninhados, pode-se calcular a *deviance* e proceder aos testes de significância, construindo a tabela ANODEV (VIEIRA, 2004). No entanto, McCullagh e Nelder (1989) afirmam que a ANODEV deve ser considerada com uma maneira de evitar a remoção do modelo de termos importantes e não uma maneira de atribuir significância aos termos.

2.2.8.5 Estatística Qui-Quadrado Generalizada de Pearson

A estatística Qui-Quadrado Generalizada de Pearson também pode ser aplicada no GLM como medida de discrepância do modelo ajustado em relação aos dados. Essa medida é definida como:

$$\chi^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\mu}_i\right)^2}{V\left(\hat{\mu}\right)}, \qquad (18)$$

onde $V\left(\hat{\mu}\right)$ é a estimativa da função de variância para a distribuição considerada (MCCULLAGH; NELDER, 1989).

Para respostas com distribuição Normal, a estatística Qui-Quadrado é igual a soma dos quadrados dos resíduos $\left(\chi^2 = SQ\,\mathrm{Re}\,s\right)$ e, portanto, a estatística $\frac{\chi^2}{\sigma^2} \sim \chi^2_{(n-p)}$ é exata. Para

distribuições não Normais, tem-se resultados assintóticos, ou seja, utiliza-se a estatística Qui-Quadrado com $(n-p)$ graus de liberdade $\left(\chi^2_{(n-p)}\right)$ como aproximação.

2.2.8.6 Comparativo entre deviance e estatística Qui-Quadrado Generalizada de Pearson

A *deviance* possui uma vantagem como medida de discrepância quando a máxima verossimilhança é utilizada. Entretanto, a estatística χ^2 é preferida em algumas situações devido a sua interpretação mais direta (MCCULLAGH; NELDER, 1989). A vantagem da *deviance* é que ela e aditiva e acrescentando-se variáveis explicativas ao modelo a *deviance* deve decrescer, diferentemente do χ^2.

Tanto a *deviance* como a estatística χ^2 Generalizada de Pearson possuem distribuição assintoticamente Normal. Por fim, cabe ressaltar que toda a inferência feita para os GLMs é baseada em resultados assintóticos. Segundo Cordeiro (1986), pouco se sabe sobre a validade desses resultados no caso de amostras muito pequenas. A literatura mais atual consultada não menciona a aplicabilidade dessas medidas de ajustamento em amostras pequenas.

2.2.8.7 Análise dos resíduos

Nos modelos padrão de regressão, os resíduos, definidos como a diferença entre os valores observados e os estimados, são freqüentemente utilizados para detectar: (*i*) violação na suposição de não homogeneidade das variâncias ou de Normalidade dos resíduos; (*ii*) presença de valores atípicos e (*iii*) influência de observações individuais no ajustamento global do modelo. Esse tipo de análise só é adequado quando a variância das observações é constante, suposição esta que não é necessária em aplicações de GLM (MCCULLAGH; NELDER, 1989). No contexto dos GLMs, os resíduos são utilizados para verificar a adequação do modelo ajustado em relação à escolha da função de variância, da função de ligação e dos termos do preditor linear. Além disso, os resíduos são também úteis para identificar a presença de pontos atípicos, que podem ser influentes ou não no modelo final (CORDEIRO; NETO, 2004). Sendo assim, os resíduos apropriados pra medir as discrepâncias de um GLM são (MCCULLAGH; NELDER, 1989):

- **Resíduo de Pearson:** $r_p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{Var}(y_i)}} = \frac{y_i - \mu}{\sqrt{Var}(\mu)}$: é preferido ao invés da *deviance* em algumas situações, pois tem interpretação mais direta; entretanto, a sua distribuição é, geralmente, assimétrica para modelos não-normais. A estatística Qui-Quadrado Generalizada de Pearson $\left(\chi_p^2\right)$ pode ser definida em ternos do resíduo de Pearson $\left(r_p\right)$, ou seja, $\chi_p^2 = \sum r_p^2$;

- **Resíduo *Deviance***: para cada resposta y_i, pode-se definir a *deviance* como $d_{i,r} = \left[\text{sgn}\left(y_i - \hat{\mu}_i\right)\right].\sqrt{d_i}$, $i = 1, 2, ..., n$, e então $\sum_{i=1}^{n} d_{i,r}^2 = D(\boldsymbol{\beta})$, onde d_i mede a contribuição da *i*-ésima observação para a *deviance*, ou seja, testa se um nova covariável pode ser incoporada ao modelo sob investigação. A *deviance* também pode ser definida em termos de seus resíduos, ou seja, $D = \sum r_d^2$ (MCCULLAGH; NELDER, 1989); e

- **Resíduo *Deviance* Studentizado:** Os resíduos *deviance* studentizados são definidos como $r_i = \frac{r_{Di}}{\sqrt{\hat{\phi}^2(1 - h_{ii})}}$, onde $\hat{\phi}^2 = D(y_i, \mu_i)/(n - p)$ é a estimativa do parâmetro de dispersão e h_{ii} é o *i*-ésimo elemento da diagonal da matriz **H**. A matriz **H** (matriz chapéu) nos GLMs é dada por $\mathbf{H}=\mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}`\mathbf{W}\mathbf{X})^{-1}\mathbf{X}`\mathbf{W}^{1/2}$, onde a matriz **W** é a matriz diagonal, com os elementos da diagonal principal dados por $w_i = \frac{1}{\text{var}(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$ (MCCULLAGH; NELDER, 1989).

Se as suposições do modelo forem corretas, a variância do resíduo studentizado será igual a 1 e a correlação entre os resíduos será pequena. Os resíduos studentizados são preferidos em análises gráficas, pois foram padronizados quanto à variação. Observa-se que o resíduo studentizado pode somente corrigir a variação natural não constante nos resíduos

quando os erros tiverem a variação constante. Se houver alguma heterodasticidade subjacente nos erros, o resíduo studentizado não poderá corrigí-la.

Para a distribuição de probabilidade Normal, o resíduo *deviance* (d) e o resíduo de Pearson elevado ao quadrado (r_p^2) são idênticos. Ambos seguem uma distribuição Gama com dispersão $\phi = \sigma^2$, $Var(\mu) = 1$ e fator de escala igual a 2, pois a $Var(d) = Var(y - \mu)^2 = 2\sigma^4 = 2\phi^2$ (NELDER e LEE, 1991) e, portanto, $E(d) = \phi Var(\mu)$.

A análise dos resíduos pode ser utilizada para avaliar a adequação do modelo, tanto na escolha da função de variância como em termos do preditor linear. O problema da análise dos resíduos no GLM está na identificação do tipo mais apropriado de resíduo na aplicação de determinado modelo.

Conforme Cordeiro (1986), para modelos bem ajustados a diferença entre os resíduos de Pearson e o d*eviance* é pequena, o que pode não ocorrer no caso de modelos mal ajustados e/ou dados com observações discrepantes. Pierce e Schafer (1986) sugerem que o resíduo *deviance* seja usado para construir gráficos de diagnóstico.

McCullagh e Nelder (1989) recomendam plotar os resíduos *deviance* contra os valores ajustados para uma escala constante da variância, e contra os regressores. Os autores também sugerem plotar os valores absolutos dos resíduos *deviance* contra os valores ajustados. Uma má escolha da função da variância pode gerar tendência no gráfico.

McCullagh e Nelder (1989) também recomendam a utilização do gráfico de probabilidade Normal dos resíduos *deviance*. Cordeiro (1986) afirma que um gráfico de resíduos padronizados *versus* os valores ajustados que não apresenta tendência pode ser um indicativo de que a relação funcional variância/média proposta para os dados é satisfatória. Todos esses gráficos são interpretados como no caso da regressão múltipla tradicional.

Lee e Nelder (1998) afirmam que os resíduos *deviance* podem ser considerados como aproximadamente Normais, com média zero e variância constante. Sendo assim, recomendam o uso de dois tipos de gráficos: (*i*) resíduos studentizado padronizado *versus* valores ajustados e (*ii*) resíduos absolutos *versus* resíduos ajustados. Os autores também ressaltam que se uma função de ligação e de variância forem encontradas, preditores lineares

parcimoniosos podem ser determinados através da eliminação retroativa ou de outros métodos.

A adequação do modelo e a existência de observações atípicas podem ser observadas com o gráfico de probabilidade Normal dos resíduos *deviance* studentizados (DEMÉTRIO, 2001).

2.2.8.8 Adequação da função de ligação e da função de variância

A adequação da função de ligação pode ser verificada pelo gráfico dos resíduos studentizados *versus* valores ajustados. A adequação da função de variância pode ser verificada através do gráfico do valor absoluto dos resíduos studentizados *versus* os valores ajustados. Caso essses gráficos não apresentem nenhum padrão óbvio (comportamento não aleatório) e a linha resultante do amortecimento (*lowess*) seja aproximadamente horizontal e próxima à linha reta horizontal de ordenada zero, há indicativo de que a função de ligação é adequada (VIEIRA, 2004).

Através do gráfico dos resíduos *versus* os valores ajustados é possível identificar pontos isolados com grandes valores de resíduos, o que indica uma função de ligação inadequada. Já a tendência à propagação do aumento de valores ajustados aponta uma função de variância insatisfatória.

A função de variância geralmente é definida como uma função de potência da média $\left(\mu^{\lambda}\right)$. Assim, quando a linha do amortecimento cresce sistematicamente, da esquerda para direita, com o aumento da média, há indícios de que se deve usar um valor maior para λ do que o valor correspondente à distribuição que foi usada no modelo. Caso contrário, ou seja, quando há um decréscimo sistemático, indica a adequação de um valor menor para λ.

Conforme Vieira (2004), a tentativa de realizar transformações nos dados para atender as suposições de um modelo de regressão linear pode auxiliar na definição dessas funções. Por exemplo, se a transformação logarítmica da resposta for considerada a mais adequada para alcançar a aditividade nos efeitos, sugere-se a uso da função de ligação logarítmica para o GLM.

Outra maneira de identificar a função de ligação e de variância que descreve o conjunto de dados é através do comportamento da variância. Para verificar esse comportamento pode-se plotar o gráfico dos resíduos contra as variáveis explicativas, para verificar se há indícios de variância constante (ou seja, gráfico com comportamento aleatório), ou se a variância cresce com o aumento da média. Vieira (2004) apresenta um teste formal proposto por Cook e Weisberg (1999) para verificar se a variância é constante. Se não há indícios de que a variância cresce com o aumento da média, sugere-se utilizar as distribuições Gama, Normal Inversa e a função de *quase*-verossimilhança com função de variância igual à média.

Caso as funções de variância ou de ligação sejam totalmente desconhecidas, Vieira (2004) sugere a utilização do valor da função de log-verossimilhança e a minimização do critério de informação de Akaike (AIC) como critério de escolha dessas funções.

2.2.8.9 Crítério de informação de Akaike (AIC)

O critério de informação de Akaike é uma estatística que leva em conta a verossimilhança e o número de parâmetros do modelo, sendo construída de tal forma que quanto menor o seu valor, melhor o modelo ajustado. O critério de informação de Akaike (AIC) é definido como:

$$AIC = -2\sum_{i=1}^{n} \ln L(\hat{\mu}_i, y_i) + 2p\,, \tag{19}$$

onde:

- y_i é o *i*-ésimo valor da resposta ($i = 1,2,...,n$);
- $\hat{\mu}_i = E[y_i] = g^{-1}\left(\mathbf{x}_i \hat{\boldsymbol{\beta}}\right)$ é a estimativa de y_i ao ajustar-se um modelo com *p* parâmetros pela maximização da função de log-verossimilhança (FLV);
- $\mathbf{x}_i = [1, x_{i1}, x_{i2}, ..., x_{ik}]$ é a *i*-ésima linha da matriz $(\mathbf{X})$ das variáveis de regressão; e

- $L(\hat{\mu}_i, y_i)$ é o valor da FLV maximizada para a resposta y_i, ao ajustar-se um modelo com *p* parâmetros.

De um modo geral, pode-se afirmar que o AIC é uma medida da distância entre o modelo verdadeiro, que usualmente é uma abstração, e um modelo candidato. Entre diversos modelos ajustados, deve-se optar por aquele que minimize o valor de AIC, pois assim será escolhido o modelo com maior valor do somatório da FLV, penalizando os modelos mais complexos.

Burnham e Anderson *apud* Vieira (2004) recomendam o AIC para a seleção de modelos, quando o número de observações da variável resposta é maior do que pelo menos 40 vezes o número de parâmetros. Conforme Cordeiro (1986), esse critério foi desenvolvido para estender o método de máxima verossimilhança para a situação de ajustamento de modelos com diferentes números de parâmetros e para decidir quando parar o ajustamento. Essa estatística, segundo o autor, pode auxiliar na seleção de modelos complexos e tem demonstrado produzir soluções razoáveis para muitos problemas de seleção de modelos que não podem ser tratados pela teoria convencional da máxima verossimilhança.

2.2.8.10 Distância de Cook

Um pequeno conjunto de dados pode exercer grande influência no ajustamento de um modelo. Ou seja, a estimação dos parâmetros pode depender mais de um conjunto influente de valores do que da maioria dos dados. Assim, é desejável identificar tais valores e avaliar seu impacto no modelo.

Pode-se utilizar a Distância de Cook para verificar a existência de observações influentes no ajustamento do modelo. A Distância de Cook para os GLM's é definida por Atkinson e Riani *apud* Vieira (2004) como $D_i = \frac{r_{Pi}^2 h_{ii}}{p\hat{\phi}(1-h_{ii})^2}$, onde *p* é o número de parâmetros e r_{Pi}^2 é o resíduo de Pearson studentizado, definido como:

$r_{Pi}^2 = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\,\mathrm{var}(\mu_i)(1-h_{ii})}}$. Valores superiores a 0,5 devem ser investigados, pois apresentam grande influência no modelo.

2.2.9 Exemplo de aplicação de GLM

O exemplo apresentado a seguir, retirado de Myers; Montgomery e Vining (2002), ilustra a aplicação da metodologia de modelos lineares generalizados (GLM), utilizando o aplicativo SAS-PROC GENMOD. Os dados do experimento apresentado na Tabela 3 foram obtidos de um processo de manufatura de semicondutores. Para a coleta de dados, rodou-se um experimento fatorial sem repetições, onde a variável resposta é a resistência. Sabe-se que a resistência segue uma distribuição assimétrica e que a distribuição Gama é adequada para modelar estes dados.

Tabela 3: Dados de resistência do exemplo de aplicação de GLM

Ordem	x_1	x_2	x_3	x_4	**Resistência (*Y*)**
1	-	-	-	-	193,4
2	+	-	-	-	247,6
3	-	+	-	-	168,2
4	+	+	-	-	205,0
5	-	-	+	-	303,4
6	+	-	+	-	339,9
7	-	+	+	-	226,3
8	+	+	+	-	208,3
9	-	-	-	+	220,0
10	+	-	-	+	256,4
11	-	+	-	+	165,7
12	+	+	-	+	203,5
13	-	-	+	+	285,0
14	+	-	+	+	268,0
15	-	+	+	+	169,1
16	+	+	+	+	208,5

Inicialmente, foi realizada a transformação logarítmica dos dados para estabilizar a sua variância. Posteriormente, foi ajustado um modelo de regressão linear supondo variâncias

homogêneas. O modelo ajustado é o seguinte: $\ln(\hat{y}) = 2{,}351 + 0{,}027x_1 - 0{,}065x_2 + 0{,}039x_3$, sendo que apenas três efeitos principais foram significativos.

Como análise alternativa, os autores sugerem a utilização do GLM, utilizando distribuição Gama e função de ligação logarítmica. Esse modelo foi obtido através do procedimento GENMOD do pacote computacional SAS. O procedimento ajusta o GLM conforme definido por Nelder e Wedderburn (1972).

Rodou-se o experimento no GENMOD com todos os efeitos principais e todas as interações de 1º ordem. Todos efeitos principais resultaram significativos, bem como interações de 1° ordem envolvendo o fator de controle x_3. Para verificar a adequação do modelo, também foi obtido o gráfico de dispersão dos valores preditos contra os resíduos *deviance*. Os resultados gerados no SAS através do procedimento GENMOD são apresentados a seguir, juntamente com algumas interpretações.

Informações do modelo	
Banco de dados	WORK.RESIST
Distribuição	Gama
Função de ligação	Log
Variável dependente	Res
N º de observações	16

Critérios para avaliar o ajustamento do modelo			
Critério	GL	Valor	Valor/GL
Deviance	8	0,0363	0,0045
Scaled Deviance	8	160.060	20.008
Qui-Quadrado de Pearson	8	0,0362	0,0045
Qui-Quadrado Generalizado de Pearson	8	159,769	19,971
Verossimilhança		-605,996	

Figura 5: Resultados obtidos no SAS

A Figura 5 apresenta informações do modelo ajustado e a avaliação de seu ajustamento através das estatísticas *deviance* e Qui-Quadrado Generalizado de Pearson, apresentadas na seção 2.2.8. Sabe-se que a estatística *deviance* e de Pearson seguem, assintoticamente, uma distribuição Qui-Quadrado. Por isso e devido ao fato do procedimento PROC GENMOD poder ajustar dados com diversas distribuições, não é possível calcular a significância $(p-value)$ destes testes. Sendo assim, pode-se utilizar a razão da estatística

calculada pelos graus de liberdade *(Valor/GL)* como um indicador de ajustamento. Deseja-se que o valor de *Valor/GL* seja aproximadamente igual a 1 (ORELIEN, 2000).

Além disso, os resultados apresentados na Figura 5 também apresentam duas estatísticas padronizadas, descritas a seguir (ORELIEN, 2000):

- **"Scaled" Deviance**: para um valor fixo do parâmetro de dispersão ϕ, a estatística *Scaled Deviance* é definida como sendo duas vezes a máxima verossimilhança para o modelo corrente. A *Scaled Deviance* pode ser definida da seguinte forma:

$$S_p = \frac{1}{\phi}\sum_{i=1}^{n} d_i^2 \,, \tag{20}$$

sendo que d_i^2 mede a diferença dos logaritmos das funções de verossimilhanças observadas e ajustadas, para as observações correspondentes e são chamadas componentes da *deviance*. A soma deles mede a discrepância total entre as duas funções de verossimilhança. É, portanto, uma medida de discrepância dos valores ajustados em relação aos dados observados, ou de forma equivalente, do modelo corrente em relação ao modelo saturado. Quanto melhor for o ajustamento do modelo aos dados, menor será o valor de S_p.

- **Qui-Quadrado Generalizado de Pearson**: corresponde ao valor da estatística generalizada de Pearson dividido pela estimativa do parâmetro de dispersão ϕ.

A Tabela 4 apresenta a estimativa dos parâmetros do modelo, seus erros padrões, os intervalos de confiança de Wald e a Estatística Qui-Quadrado com a sua respectiva significância.

Ao final da Tabela 4, é calculado o parâmetro de escala pelo método da máxima verossimilhança. Esse parâmetro está relacionado com o parâmetro de dispersão (ϕ) definido na Família Exponencial.

Tabela 4: Análise da estimativa dos parâmetros

Parâmetro	GL	Estimativa	Erro Padrão	Limites de Confiança de Wald de 95%		Qui-Quadrado (χ^2)	Prob. > χ^2
Intercepto	1	5,4142	0,0119	5,3909	5,4375	207005	<0,0001
x1	1	0,0613	0,0119	0,0379	0,0846	26,50	<0,0001
x2	1	-0,1496	0,0119	-0,1729	-0,1262	157,93	<0,0001
x3	1	0,0899	0,0119	0,0666	0,1133	57,13	<0,0001
x4	1	-0,0278	0,0119	-0,0511	-0,0045	5,46	0,0195
x1*x3	1	-0,0389	0,0119	-0,0622	-0,0155	10,67	0,0011
x2*x3	1	-0,0441	0,0119	-0,0674	-0,0207	13,71	0,0002
x3*x4	1	-0,0455	0,0119	-0,0688	-0,0222	14,60	0,0001
Parâmetro de Escala	1	441,3557	155,9839	220,7787	882,3083		

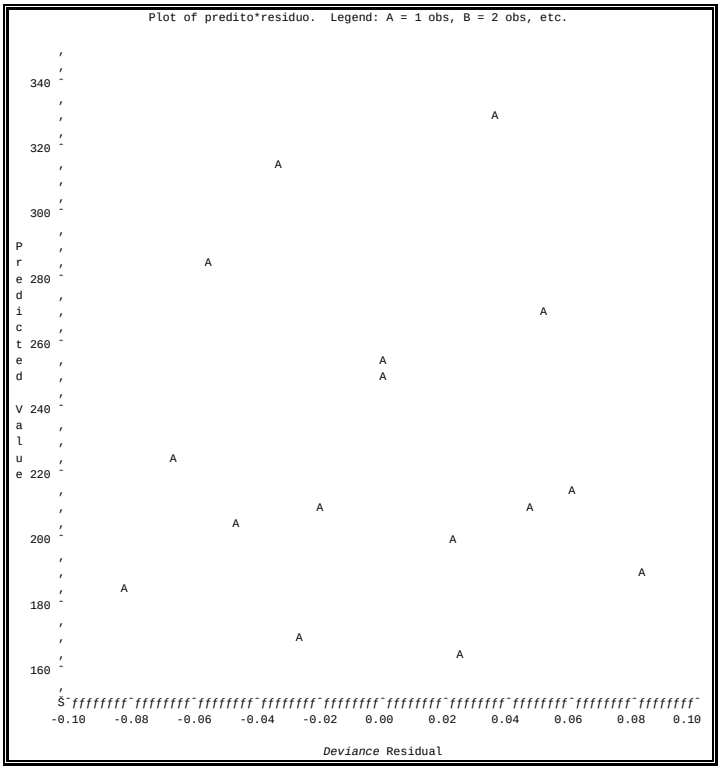


Figura 6: Gráfico dos valores preditos x resíduos *deviance*

A Figura 6 apresenta o gráfico de dispersão dos valores preditos contra os resíduos *deviance*, para análise da adequação do modelo ajustado. Como o gráfico não apresenta nenhum padrão, pode-se afirmar que o modelo é adequado.

Enfim, analisando os resultados gerados pelo programa SAS, o modelo ajustado utilizando GLM, considerando a distribuição Gama para os dados e função de ligação logarítmica, é o seguinte: $\hat{y} = e^{a}$, onde:

$$a = 5{,}41 + 0{,}06x_1 - 0{,}15x_2 + 0{,}09x_3 - 0{,}028x_4 - 0{,}039x_1x_3 - 0{,}04x_2x_3 - 0{,}045x_3x_4 .$$

Ao comparar o modelo utilizando a transformação logarítmica dos dados (modelo de regressão linear tradicional) com o modelo obtido utilizando o procedimento de GLM, observa-se que este último fornece mais informações. O modelo tradicional de regressão identificou apenas os efeitos principais como significativos; já o modelo obtido através do GLM, além dos efeitos principais, identificou como significativas todas as interações com o efeito x_3, sugerindo que esse efeito possui forte influência na variável resposta.

No exemplo apresentado, os autores optaram por utilizar uma função logarítmica de ligação. Entretanto, sabe-se que é possível utilizar a função de ligação canônica. No caso da distribuição Gama, a função canônica é definida como a recíproca da média $(\eta_i = 1/\mu_i)$. Entretanto, a função recíproca não está disponível como função de ligação do procedimento GENMOD do SAS, sendo necessário elaborar uma rotina para o seu cálculo, o que foge do escopo deste exemplo.

2.3 PROPOSTAS DE MODELAGEM DE MÉDIA E VARIÂNCIA

As características de processos industriais são afetadas por muitos fatores. Alguns afetam a média das características, tendo efeito de localização; outros afetam a variação do processo, tendo, portanto, efeito de dispersão (ou fatores de ruído, segundo Nair (1986); Box; Meyer (1986)). Em um processo de otimização, deseja-se minimizar a variabilidade e, simultaneamente, ajustar a média da resposta em seu valor alvo. Assim, é necessário conhecer sob quais condições é possível encontrar fatores de ajustamento que permitam trazer o processo para um valor-alvo de média sem alterar a variabilidade da variável analisada.

O Planejamento de Experimentos tem sido utilizado para reunir informações acerca do desempenho de processos (WANG; LIN, 2001). Através dessa ferramenta estatística é possível identificar os fatores preponderantes (ativos) em um processo. Identificados tais

fatores, pode-se modelar seus efeitos de localização e de dispersão, ou seja, a sua média e variância, respectivamente. De posse dos modelos é possível estabelecer a combinação de níveis da variável independente que resulta num valor de resposta desejado.

Os dados de um experimento podem ser obtidos de quatro fontes principais. Podem ser completos ou fracionados e, ainda, subdivididos em com e sem repetição de tratamentos. Em experimentos que geram dados completos, todos os tratamentos possuem o mesmo número de observações e todos são analisados. Se houver repetição de um tratamento, os dados são ditos completos com repetição. Quando somente uma fração dos tratamentos é analisada, o projeto é chamado de fracionado, podendo ou não haver repetição dos tratamentos estudados.

Para determinar as condições de variação mínima de um processo, a maioria das técnicas de análise existentes requer dados oriundos de experimentos, envolvendo muitas repetições (BERGMAN; HYNEN, 1997). Dados completos são muito eficientes na prática, mas podem ser de difícil e cara obtenção (LIAO, 2000). Para contornar esses problemas, experimentos fracionados têm sido utilizados. Diversos autores, tais como Box e Meyer (1986) e Wang (1989), sugeriram métodos para determinar fatores e/ou interações com efeito de dispersão em experimentos sem repetições.

Freqüentemente, em experimentos industriais, uma grande parte da variação do processo está associada a um pequeno número de variáveis. Nessas situações, o uso de fatoriais fracionados sem repetição tem sido geralmente efetivo para identificar isoladamente os fatores ativos no processo (BOX; MEYER, 1986; BERGMAN; HYNEN, 1997). Em experimentos sem repetição, os contrastes associados com as interações de alta ordem (supostamente não significativas) são geralmente usados para estimar a variância do erro. Entretanto, esse método nem sempre é satisfatório, pois contrastes inertes podem ser de difícil ou mesmo impossível de identificação (BOX; MEYER, 1986).

Definido o tipo de dados a ser analisado, existem diversas estratégias de modelagem de média e variância descritas na literatura. Algumas propostas modelam parâmetros de forma individual e outras, de forma conjunta. Dentre as estratégias individuais, tem-se os métodos de ANOVA (Análise de Variância) combinado com a análise de regressão e os métodos derivados da proposta de Box e Meyer (1986). Já as propostas de modelagem conjunta podem ser divididas entre aquelas que utilizam como ferramenta o GLM (Modelos Lineares

Generalizados) e as que não o utilizam e, são, portanto, mais restritivas quanto ao comportamento dos dados.

As relações entre os tipos de dados e as estratégias de modelagem são apresentadas na Figura 7 e descritas na próxima seção.

<table>
<tr><th colspan="2" rowspan="2">Tipos de dados</th><th colspan="2">Modelagem não conjunta</th><th colspan="2">Modelagem conjunta</th></tr>
<tr><th>Anova +Regressão</th><th>Tipo Box e Meyer (1986)</th><th>GLM</th><th>Outras</th></tr>
<tr><td rowspan="2">Fatorial completo</td><td>com repetição</td><td>Possível, utilizando os resíduos do modelo da média para estimar o modelo da dispersão.</td><td>Não é possível</td><td colspan="2" rowspan="2">O estudo desta possibilidade não fazia parte do escopo dessa dissertação.</td></tr>
<tr><td>sem repetição</td><td>Modela somente a média</td><td>Não foi encontrado na literatura.</td></tr>
<tr><td rowspan="2">Fatorial fracionado</td><td>com repetição</td><td>Não foi encontrado na literatura.</td><td>* Pan (1999)</td><td>* Lee e Nelder (1998)</td><td>* McGrath e Lin (2001)
* Wang (1989)
* Wolfinger e Tobias (1998) – simulação</td></tr>
<tr><td>sem repetição</td><td>Modela somente a média</td><td>* Box e Meyer (1986)
* Ribeiro, Fogliatto e Caten (2001)
* Bergman e Hynen (1997)
* Wang (1989)
* Liao (2000)</td><td>ESCOPO DESSA DISSERTAÇÃO</td><td>* Wolfinger e Tobias (1998)</td></tr>
</table>

Figura 7: Tipos de dados x tipos de modelagens

2.3.1 Modelagem individual de média e variância

2.3.1.1 Análise de variância e análise de regressão

Em procedimentos de modelagem, a Análise de Variância (ANOVA) pode ser utilizada em conjunto com a análise de regressão. A ANOVA é utilizada para identificar fatores com efeito significativo sobre a resposta. O efeito de um fator é definido como a

mudança que se verifica na variável resposta quando o nível desse fator é alterado. Após a identificação dos fatores significativos, pode-se utilizar a análise de regressão para gerar modelos baseados em tais fatores que otimizem a variável resposta.

A modelagem individual que combina as técnicas de ANOVA e regressão permite modelar apenas a média quando não há repetição dos tratamentos, podendo ser aplicada quando o conjunto de dados é completo ou fracionado (CATEN, 1995 e BARBETTA; 1998).

2.3.1.2 Metodologias baseadas na proposta de Box e Meyer (1986)

Uma contribuição importante de Taguchi na otimização de produtos e processos é a idéia de que a qualidade de um produto está associada à variabilidade de suas características de qualidade. Taguchi deu início aos estudos de experimentos fatoriais para analisar efeitos de dispersão além dos de localização (LIAO, 2000).

Dando continuidade aos estudos de Taguchi, a literatura apresenta diversas propostas para a modelagem de média e de variância a partir de dados oriundos de experimentos fracionados. A primeira, e uma das mais referenciada por outros autores, foi desenvolvida por Box e Meyer (1986). Segundo os mesmos, é possível utilizar fatoriais fracionados sem repetições para identificar fatores que afetam a variância, além daqueles que afetam a média. Sabe-se que a estimação direta da dispersão através da replicação do experimento pode ser cara e demorada. A estimação sem repetições é recomendada pelos autores em experimentos do tipo *screening*, com o objetivo de identificar os fatores que potencialmente podem afetar a média e a variância. Experimentos do tipo *screening* têm sido utilizados em estágios inicias do desenvolvimento de produtos e processos novos para identificar efeitos de dispersão e de localização em experimentos onde se utiliza o fracionamento sem repetição. Assim, pode-se reduzir o número de fatores candidatos ao modelo de ajustamento, sendo possível estudar detalhadamente os fatores significativos através de uma análise mais formal.

Box e Meyer (1986) apresentam em seu trabalho método exploratório para identificar separadamente efeitos de dispersão e de localização quando não há repetição de tratamentos. Esse procedimento não possui estatística de teste, pois a distribuição da sua estatística não é conhecida. Em sua metodologia, Box e Meyer (1986) propõem identificar os efeitos de localização através do gráfico de probabilidade Normal proposto por Daniel (1959), que utiliza o Gráfico de Probabilidade Normal para identificar os fatores preponderantes. Os

efeitos significativos vão aparecer afastados da linha, onde vão se concentrar os efeitos inertes, os quais não são identificados a *priori*.

Após a identificação dos fatores preponderantes através da análise gráfica, deve-se realizar um experimento fatorial completo com repetição dos fatores com efeito sobre a média da variável resposta. A probabilidade de um efeito ser ativo é calculada através de uma aplicação do teorema de Bayes, sendo o cálculo da probabilidade de um efeito independente dos demais. Depois de identificar os efeitos de localização e incorporá-los no modelo da variância através do cálculo de suas estimativas e análise de seus resíduos, o planejamento pode ser reexaminado para detectar efeitos de dispersão ativos.

Após ajustar o modelo para a variância, Box e Meyer (1986) recomendam a estimação por mínimos quadrados para um ajustamento mais preciso do modelo e apresentam esses estimadores em seu trabalho. Os autores acreditam que, em um estudo preliminar, essa metodologia fornece uma maneira econômica de identificar um pequeno número de efeitos de localização e dispersão significativos.

Baseado na proposta de Box e Meyer (1986) para experimentos fatoriais fracionados sem repetição, Ribeiro; Fogliatto e Caten (2001) propõem uma modelagem simplificada da média e variância. Essa proposta modela a variância das respostas, usando os resíduos do modelo de regressão para a resposta média, sem a repetição dos tratamentos. Ou seja, o procedimento consiste em verificar se os resíduos de um nível de um determinado fator controlável difere significativamente de outro nível; em caso afirmativo, a variância da resposta é dada pelos resíduos e pode ser modelada como função desse fator.

Assim como a metodologia proposta por Ribeiro, Fogliatto e Caten (2001), existem diversas contribuições na literatura de propostas que usam resíduos da modelagem da média da resposta para gerar modelos para a dispersão (BOX; MEYER, 1986; DAVIDIAN; CARROLL, 1987; WANG, 1989; BERGMAN; HYNEN, 1997). Tais metodologias freqüentemente envolvem estatísticas de testes com distribuições de probabilidade desconhecidas, e a identificação dos efeitos de dispersão depende dos fatores que possuem efeitos de localização.

Opondo-se a esses métodos, Bergman e Hynen (1997) propõem a utilização das observações originais e fornecem uma estatística de teste com distribuição conhecida (F-*Snedecor*) para identificar os efeitos de dispersão e localização em experimentos do tipo

screening. Os autores desenvolveram um método para identificar efeitos de dispersão em experimentos fracionados 2^k sem repetição, ajustando, separadamente, modelos de regressão para os níveis alto e baixo dos fatores.

Assim como Bergman e Hynen (1997), Wang (1989) propõe uma metodologia para identificar efeitos de dispersão baseada em uma distribuição de probabilidade conhecida. A estatística de teste proposta pelo autor segue aproximadamente uma distribuição Qui-Quadrado, sendo definida pelo quadrado da diferença da soma dos resíduos padronizados ao quadrado correspondente aos dois níveis do fator considerado. Se essa soma é grande comparada com os valores da distribuição Qui-Quadrado, conclui-se que há indicativo de efeitos de dispersão significativos nos fatores e nas interações analisadas. Esta estatística de teste pode ser calculada utilizando o programa computacional GLIM (WANG, 1989). O teste Qui-Quadrado é utilizado como teste formal, mas é necessário um tamanho de amostra grande e erros Normalmente distribuídos.

A metodologia proposta por Wang (1989) pode ser facilmente aplicada em planejamentos ortogonais de fatoriais fracionados com repetições, viabilizando uma modelagem conjunta da média e da variância da variável resposta. Se esse não é o caso, pode-se identificar efeitos de localização e dispersão separadamente. Conforme Wang (1989) e Box e Meyer (1986), os efeitos de localização devem ser identificados primeiro, pois os resultados da dispersão, sem a eliminação dos efeitos de localização significantes, podem ser equivocados.

Observa-se que Wang (1989) e Bergman e Hynen (1997) desenvolveram uma maneira mais formal para identificar os efeitos significativos, baseados nas distribuições Qui-Quadrado e F, respectivamente. Em contrapartida, Liao (2000) propõe um procedimento para identificar efeitos de dispersão em fatoriais fracionados sem repetição usando a razão do logaritmo da verossimilhança baseado na Normalidade dos erros.

Comparando o poder do seu método com o método de Wang (1989) e Bergman e Hynen (1997), Liao (2000) conclui que o seu método é mais sensível na identificação de efeitos de dispersão, pois os graus de liberdade são ajustados de acordo com o número de efeitos de localização ativos. Os resultados apresentados pelo autor demonstram que seu teste é mais sensível na identificação de efeitos de dispersão significativos no estágio inicial de um experimento.

Uma vez identificados os fatores influentes para a dispersão e para a localização da resposta, pode-se planejar um experimento mais elaborado para caracterizar a relação entre a variável resposta e os fatores pela média de funções matemáticas adequadas. Para tanto, os modelos mistos discutidos por Wolfinger e Tobias (1998) podem ser uma boa escolha.

Alguns métodos assumem que os efeitos de localização podem ser corretamente identificados num experimento sem repetição. Conforme Pan (1999), tais métodos podem deixar de identificar pequenos e médios efeitos de localização. Assim a inadequada identificação de efeitos na média diminui a eficiência do método utilizado para a identificação de efeitos de dispersão. Por outro lado, os efeitos de média não identificados, mesmo sendo pequenos, podem, cumulativamente, invalidar o método usado para identificar os efeitos de dispersão. Logo tais efeitos devem ser interpretados com cuidado, e o autor sugere a repetição do experimento ou a utilização de um fatorial completo.

Contrário aos demais autores citados, Pan (1999) propôs uma metodologia de modelagem a qual exige repetição de tratamentos. Segundo Wang e Lin (2001), Pan (1999) enfatizou o problema de encontrar diretamente fatores principais e/ou interações com efeitos de dispersão através de um exemplo numérico, um estudo de simulação e argumentos matemáticos. Para lidar com isso, ele sugeriu o uso de experimentos com repetição para eliminar efeitos de localização não identificados na determinação de efeitos de dispersão.

Enfim, observa-se que a maioria dos procedimentos de identificação de efeitos de dispersão em experimentos fracionados sem repetição envolve duas etapas (LIAO, 2000). Inicialmente, deve-se aplicar o Gráfico de Probabilidade Normal para identificar os efeitos de localização mais significativos. Posteriormente, calcula-se uma estatística relacionada com o efeito de dispersão, geralmente baseada nos resíduos de um modelo linear ajustado à média. Então, tradicionalmente aplica-se o Gráfico de Probabilidade Normal novamente na estatística calculada para identificar efeitos de dispersão (BOX; MEYER, 1986).

O princípio da parcimônia usado na prática da modelagem sugere que, na maioria dos casos, apenas alguns poucos efeitos principais são significativos. Assim, Liao (2000) afirma que se pode utilizar o Gráfico de Probabilidade Normal para identificar os efeitos que parecem significativos sobre a localização. Entretanto, esse é um método subjetivo. Bergman e Hynen (1997) afirmam que o uso do Gráfico de Probabilidade Normal para identificar os efeitos de localização é adequado apenas quando a dispersão da variável resposta pode variar

com os níveis de alguns fatores do experimento. Nelder e Lee (1998) consideram que ferramentas gráficas são muito úteis, mas deveriam ser utilizadas com métodos mais formais desde o inicio da análise. Se apenas métodos gráficos forem utilizados, efeitos intermediários potencialmente significativos podem ser desconsiderados.

2.3.2 Modelagem conjunta de média e variância

Métodos de modelagem conjunta de média e variância podem ser divididos entre aqueles que utilizam GLM e aqueles que utilizam outros tipos de modelos. McCullagh e Nelder (1989) e Lee e Nelder (1998) apresentam uma proposta de modelagem conjunta utilizando GLM. Já Wang (1989) e Wolfinger e Tobias (1998) propõem a modelagem de planejamentos ortogonais de fatoriais fracionados com repetições e modelos mistos, respectivamente.

Cabe ressaltar que Engel e Huele (1996) aplicam GLM para modelar apenas a dispersão. Para a modelagem da média, os autores assumem um modelo cuja função de ligação é a identidade, sendo assim um caso particular dos GLMs propostos por McCullagh e Nelder (1989) e Nelder e Lee (1998). Conforme Vieira (2004), Engel e Huele (1996) não consideram o uso da técnica de máxima verossimilhança restrita para ajustamento no modelo da média, necessário em experimentos fatoriais altamente fracionados, utilizados por McCullagh e Nelder (1989) e Nelder e Lee (1998). Além disso, Engel e Huele (1996) utilizam um experimento fatorial completo e com repetição para ilustrar a simulação apresentada em seu artigo.

2.3.2.1 Modelos lineares generalizados (GLM)

Lee e Nelder (1998) apresentam um modelo conjunto para média e para dispersão, utilizando GLM para dados que, mesmos transformados, não produzem necessariamente variância constante e linearidade dos efeitos sistemáticos para a média e para a dispersão. Os autores demonstram, através de exemplos, que a análise de todos os dados permite estimar os resíduos individuais como medida de ajustamento, e não apenas a *deviance*.

Suponha uma variável y com média $E(y_i) = \mu_i$ e variância $Var(y_i) = \phi_i Var(\mu_i)$, onde ϕ_i é o parâmetro de dispersão (diferente para cada combinação dos níveis dos fatores) e

$Var(\mu_i)$ é a função da variância, que expressa a parte da variância funcionalmente dependente da média μ_i. Nelder e Lee (1991) propuseram os seguintes modelos para a média e para dispersão: $\eta_i = g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$ e $\xi_i = h(\phi_{ii}) = \mathbf{z}_i'\gamma$, respectivamente, sendo que os efeitos sistemáticos de tais modelos são obtidos por funções de ligação.

Em situações onde não é possível utilizar a Família Exponencial de distribuições, Wedderburn (1974) desenvolveu a *quase*-verossimilhança, que pode ser utilizada para concluir apenas sobre o modelo para média. Quando o parâmetro de dispersão é diferente para cada resposta y_i, a função de *quase*-verossimilhança não é suficiente para a modelagem da média. Para esses caso, utiliza-se a *quase*-verossimilhança estendida (QVE) definida na seção 2.2.6.

Modelo da média

Para cada ϕ_i definido, a QVE é a *quase*-verossimilhança para um modelo com função de variância $V(\mu_i)$. Assim, a maximização da função da QVE com relação a $\boldsymbol{\beta}$ será obtida com os mesmos estimadores de *quase*-verossimilhança, porém com pesos $1/\phi_i$, cujas funções-escore são (VIEIRA, 2001):

$$\frac{\partial Q^+}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = 0 . \qquad (21)$$

Modelo da variância

O modelo para dispersão proposto por Lee e Nelder (1998) é $\xi_i = h(\phi_{ii}) = \mathbf{z}_i'\gamma$. Nessa modelagem, a variável resposta é o resíduo *deviance* d_i. Assim, pode-se definir:

$$E(d_i) = \phi_i \quad \text{e} \quad Var(d_i) = 2\phi_i^2 . \qquad (22)$$

Embora d_i seja geralmente um estimador viesado para ϕ_i, Lee e Nelder (1998) mostram, num estudo de simulação, que a máxima verossimilhança extendida (MVE) produz melhores estimadores do que utilizando os resíduos de Pearson e estimadores de máxima verossimilhança. O uso da MVE para o modelo da dispersão é utilizado para identificar fatores experimentais significantes na mesma escala de ligação e para comparar diferentes funções de ligação.

A Figura 8 apresenta um resumo dos modelos e componentes para a média e para a variância que foram apresentados anteriormente.

Componente	Modelo para Média	Modelo para Variância
Critério de minimização	$-2Q^+$	$-2Q_c^+$
Variável Resposta	y_i	d_i
Valor Esperado	μ_i	ϕ_i
Parâmetro de escala	ϕ_i	2
Função de Variância	Arbitrária	Arbitrária
Função de Ligação	Arbitrária	Log (usualmente)
Preditor Linear	$\eta_i = \mathbf{X}_i\boldsymbol{\beta}$	$\xi_i = \mathbf{Z}_i\boldsymbol{\gamma}$
Peso	$1/\phi_i$	1
Fonte: Lee e Nelder (1998)		

Figura 8: Resumo da modelagem conjunta por GLM para média e variância

Procedimento iterativo de modelagem da média e da variância

Considerando um experimento com repetição de cada combinação dos fatores, deve-se iniciar o procedimento de modelagem, ajustando um modelo saturado para a média, o que implica num subseqüente modelo de dispersão, utilizando apenas contrastes dentro de realizações. Busca-se um modelo parcimonioso para a dispersão, usando, por exemplo, um procedimento de eliminação retroativa (*backward stepwise selection*). Lee e Nelder (1998) também recomendam que sejam utilizados os inversos dos valores ajustados como pesos, como nova pesquisa na busca de um GLM plausível para a média.

Neste contexto, Vieira (2004) sugere o seguinte procedimento iterativo para modelagem da média e da variância baseado nas definições de Lee e Nelder (1998).

- **Ajustamento do modelo da média**: Deve-se escolher: (*i*) a função de ligação, (*ii*) a função de variância e (*iii*) selecionar os coeficientes significativos. Os graus de liberdade restantes serão utilizados para o modelo da dispersão. O parâmetro ϕ_i é considerado constante na primeira iteração. Nas iterações seguintes, utiliza-se $1/\phi_i$, proveniente do modelo de dispersão, para ajustar os coeficientes das equações de *quase*-verossimilhança. O componente de *deviance* d_i, resultante do ajustamento do modelo, será a variável resposta do modelo da dispersão.

- **Ajustamento do modelo para a dispersão**: Considerando d_i como resposta, ajustar o modelo para a dispersão. A escolha mais usual é a distribuição Gama com função de ligação logarítmica e parâmetro de escala 2. O valor ajustado para ϕ_i será utilizado no modelo da média.

O processo iterativo termina quando os parâmetros do modelo da dispersão são iguais em duas iterações consecutivas, a menos de uma tolerância definida no início da modelagem.

Lee e Nelder (1998) concluem que o uso de todos os dados permite o cálculo de estatísticas conhecidas, tal como o uso dos resíduos para verificar o ajustamento do modelo, a estatística *t* para testar a significância dos fatores e o processo iterativo para estimar o modelo para a dispersão. Além disso, os autores afirmam que, sem um procedimento iterativo, tanto os parâmetros do modelo de dispersão como os erros padrões para ambos os modelos (média e dispersão) podem ser subestimados, resultando na seleção equivocada de determinados fatores para compor o modelo final.

2.3.2.2 Outras propostas de modelagem conjunta

Assim como na metodologia de Lee e Nelder (1998), o modelo proposto por Wang (1989) também permite a modelagem conjunta de média e variância, desde que aplicado em planejamentos ortogonais de fatoriais fracionados com repetições. Assim, haverá graus de liberdade suficientes para testes simultâneos dos efeitos. Se este não é o caso, identifica-se efeitos de localização e de dispersão separadamente, conforme descrito na seção anterior.

McGrath e Lin (2001) também afirmam que é necessário utilizar repetições para fazer a modelagem conjunta de média e variância quando se utilizam dados de projetos fatoriais fracionados. Segundo os autores, isso ocorre porque se o modelo para a média não incluir todos os termos significativos, tais termos podem erroneamente aparecer como significativos na modelagem da variância, se o procedimento de Box e Meyer (1986) for adotado, por exemplo. Em outras palavras, os efeitos de localização devem ser estudados e incorporados ao modelo da média antes do estudo da variância, pois a identificação do efeito da variância é sensível ao modelo ajustado para a média. O estudo de efeitos de dispersão na presença de efeitos de localização significativos não incorporados ao modelo da média pode gerar resultados equivocados. Se um par de efeitos de localização significativos não for incluído no modelo da média, sua interação pode surgir como um efeito de dispersão espúreo no modelo para a variância. McGrath e Lin (2001), assim como Box e Meyer (1986), recomendam que os efeitos de localização sejam primeiramente identificados e, posteriormente, sejam utilizados os resíduos do modelo da média para identificar os efeitos de dispersão.

McGrath e Lin (2001) apresentam um detalhamento da análise em Box e Meyer (1986) e derivam uma relação explícita entre os efeitos de localização e dispersão. Eles mostram que, sem repetição das rodadas experimentais, não é possível determinar se um efeito da dispersão ou se dois efeitos de localização são significativos. Isso ocorre porque existe uma relação confusa entre os efeitos de dispersão e localização em um experimento fracionado sem repetições e, sem informações adicionais, esse confundimento não pode ser removido. Assim, deve-se utilizar repetições para avaliar com precisão o efeito da variância. Quando isso não for possível, é possível utilizar pontos centrais adicionais para auxiliar na separação desses efeitos. Alternativamente, pode-se realizar um experimento com todos os fatores fixos, exceto os que se suspeita da presença do efeito da variância. Assim, se forem realizadas k repetições do efeito suspeito, realiza-se o teste $F_{(k\text{-}1)(k\text{-}1)}$ para testar o efeito da variância. Estudos preliminares dos autores mostram que o efeito de dispersão produz correlação entre um par de efeitos de localização. A análise dessa correlação pode ajudar a remover o confundimento entre efeitos da média e da variância.

Outra abordagem da modelagem conjunta é encontrada no trabalho de Wolfinger e Tobias (1998), os quais propõem a modelagem conjunta dos efeitos de localização e dispersão utilizando modelos mistos e assumindo Normalidade. Modelos mistos são usualmente

utilizados quando os dados envolvem alguma estrutura de blocos que afeta a covariância entre as observações, ou seja, existe uma variável que distingue dois grupos. Conforme Wolfinger e Tobias (1998), a aplicação de modelos mistos apresenta diversas vantagens, entretanto, observa-se que as inferências neles baseadas pressupõem dados Normalmente distribuídos e a escolha adequada do modelo. Além disso, modelos mistos não permitem detectar pequenos efeitos de localização na presença de grandes efeitos de dispersão. Por fim, um modelo misto complexo não pode, em alguns casos, ser ajustado para um conjunto pequeno de dados extremamente fracionados.

Diferentemente dos modelos mistos, o GLM pode ser aplicado em experimentos fracionados, além de permitir detectar efeitos de localização e dispersão, independente de sua intensidade.

3 ROTEIRO DE MODELAGEM CONJUNTA DE MÉDIA E VARIÂNCIA

Neste capítulo, inicialmente, será descrito um roteiro para modelagem de dados empregando os GLMs com e sem repetição (seção 3.1 a 3.4) e, posteriormente, será apresentado o procedimento para modelagem conjunta de média e variância utilizando GLM em experimentos fatoriais fracionados sem repetição através de um processo iterativo (seção 3.5). Tal procedimento está baseado na metodologia de modelagem conjunta proposta por Lee e Nelder (1998) e na utilização do pacote computacional "R". Os comandos e notações de possíveis GLM ajustados através desse pacote estão descritos no Apendice B e C, respectivamente. Ao final desta seção serão apresentados dois fluxogramas, os quais resumirão os dois procedimentos de modelagem propostos.

Conforme Nelder e Lee (1991), o ajustamento de um GLM tem dois objetivos: separação dos modelos e parcimônia. A separação será atingida se a função de variância para a média μ for corretamente definida; assim, o parâmetro de dispersão ϕ será livre de influências da média. A parcimônia será obtida pela correta identificação da função de ligação e do preditor linear para ambos os modelos, o da média e o da variância. Salienta-se que nenhuma transformação nos dados é realizada no ajustamento de um GLM.

3.1 ESPECIFICAÇÃO DA VARIÁVEL RESPOSTA E DEFINIÇÃO DA DISTRIBUIÇÃO DE PROBABILIDADE DO COMPONENTE ALEATÓRIO PARA O MODELO DA MÉDIA

A primeira etapa a ser realizada no processo de modelagem conjunta de média e variância é a definição da variável resposta como contínua ou discreta.

São consideradas variáveis contínuas aquelas que podem ser representadas por uma grandeza definida no intervalo dos números reais, por exemplo, volume, custo, resistência etc. Nesses casos, o espaço amostral da variável aleatória é contínuo e as distribuições de probabilidade Normal, Gama e Normal Inversa podem representar as variáveis. Para definir a melhor distribuição de probabilidade que se ajusta aos dados pode-se utilizar informações sobre simetria e comportamento da variância da variável resposta. A distribuição Normal é simétrica em relação à média e possui variância constante. Já as distribuições Gama e Normal Inversa são assimétricas e apresentam variância aumentando com a média.

As variáveis que só podem assumir valores pertencentes a um conjunto finito ou enumerável, sendo geralmente números inteiros, são denominadas variáveis discretas. Por exemplo, peças que podem ser classificadas apenas de duas formas: “defeituosa” ou “não defeituosa”. O número de peças defeituosas de uma amostra aleatória de tal peça segue uma distribuição de probabilidade Binomial. Outro exemplo de variável discreta bastante conhecido é o número de defeitos por unidade de inspeção. Conforme Vieira (2004), a contagem de defeitos ocorre em um intervalo contínuo e geralmente apresenta as seguintes condições: (*i*) independência dos eventos (defeitos); (*ii*) os eventos ocorrem aleatoriamente em qualquer ponto do intervalo; e (*iii*) não podem ocorrer dois ou mais eventos em um mesmo ponto do intervalo. Nessas condições, o número de defeitos em um determinado intervalo contínuo segue a distribuição de probabilidade de Poisson.

3.2 DEFINIÇÃO DA FUNÇÃO DE LIGAÇÃO E DA FUNÇÃO DE VARIÂNCIA

Definida a distribuição de probabilidade da variável resposta, é possível utilizá-la com diferentes funções de ligação e de variância. A função de ligação definirá a relação funcional entre a média dos dados e a sua estrutura linear. A escolha da função de ligação

pode seguir as diretrizes na seção 2.2.3. As análises para verificar a sua adequação são apresentadas na seção 2.2.8.8.

A função de variância é determinada pela distribuição de probabilidade dos dados. Nos casos em que não é possível definir a distribuição exata e utiliza-se a *quase*-verossimilhança para estimar os coeficientes, deve-se definir a função de variância que descreve os dados. Os meios para testar a sua adequação são descritos na seção 2.2.8.8.

3.3 ADEQUAÇÃO DO MODELO

3.3.1 Significância dos coeficientes

Para testar a significância dos coeficientes dos fatores incluídos no modelo, recomenda-se o procedimento de eliminação retroativa (*backward stepwise selection*) dos fatores. Para testar a significância de tais fatores, utiliza-se a estatística *t*, conforme descrito na seção 2.2.7.1.

3.3.2 Análise da *deviance* (ANODEV)

Uma vez que o teste *t* para GLMs é um teste assintótico, recomenda-se analisar os resíduos da *deviance* através da Análise de *Deviance* (ANODEV) descrita na seção 2.2.8.4. O objetivo é avaliar se a inclusão de cada termo é significativa para a redução da *deviance* residual do modelo. A ANODEV testa a significância dos coeficientes através da diferença de *deviance* entre dois modelos, avaliando, assim, o decréscimo provocado na *deviance* devido à inclusão de um determinado termo ao modelo.

Para analisar a significância do modelo como um todo testa-se a *deviance* residual do modelo através da comparação com uma distribuição Qui-Quadrado, com *n-p* graus de liberdade, quando o modelo com *p* parâmetros é considerado correto. Se a *deviance* residual for menor que a estatística tabelada, o modelo pode ser considerado adequado a um determinado nível de significância previamente definido.

3.3.3 Análise gráfica dos resíduos

Por fim, realiza-se a análise gráfica dos resíduos, conforme descrito na seção 2.2.8.7. Para a análise da adequação do modelo da média e o da variância utiliza-se os resíduos *deviance*, *deviance* studentizado e o de Pearson, conforme descrito na seção 2.2.8.7. A análise gráfica desses resíduos, de acordo com sugestões de alguns autores apresentadas no final da seção 2.2.8.7 e resumidos na Figura 9, é usada para avaliar a adequação do modelo. Deseja-se que os gráficos que envolvem os resíduos apresentem uma distribuição aleatória dos mesmos em torno de zero com amplitude constante (DEMÉTRIO, 2001). Para verificar a presença de valores atípicos recomenda-se a utilização da estatística de Cook.

Tipo de gráfico	**Elemento testado no modelo**	**Diagnóstico de não adequação**	**Referência**
Resíduo *deviance* (absoluto) x valores ajustados	Função de variância	Tendência no gráfico	McCullagh e Nelder (1989)
Resíduo *deviance* x regressores	Função de variância	Tendência no gráfico	McCullagh e Nelder (1989)
Probabilidade Normal dos resíduos *deviance*			McCullagh e Nelder (1989)
Resíduos padronizados x valores ajustados	Relação funcional variância/média satisfatório	Tendência	Cordeiro (1986)
Resíduos studentizados padronizados x valores ajustados			Nelder e Lee (1998)
Resíduos absolutos x resíduos ajustados			Nelder e Lee (1998)
Probabilidade Normal dos resíduos *deviance* studentizados	Adequação do modelo e identificação de observações atípicas		Demétrio (2001)
Resíduos studentizados x valores ajustados	Função de ligação		McCullagh e Nelder (1989)
Valor absoluto resíduos studentizados x valores ajustados	Função de variância		McCullagh e Nelder (1989)

Figura 9: Resumo das análises gráficas de resíduos sugeridas na literatura para verificar a adequação de um GLM

Dentre os diversos gráficos sugeridos na literatura, recomenda-se a análise dos seguintes, por serem os mais relevantes e informativos:

- **Resíduos × valores preditos:** para verificar a homogeneidade da variância;

- **Normal Q-Q Plot**: para verificar a aderência dos dados a uma distribuição de probabilidade previamente definida;
- **Distância de Cook**: para verificar a presença de valores atípicos (ver seção 2.2.8.10). Em alguns pacotes computacionais esta estatística é apresentada na forma de tabelas;
- **Valores absolutos dos resíduos studentizados × valores ajustados**: para verificar a adequação da função de variância; e
- **Valores dos resíduos studentizados × valores ajustados**: para verificar a adequação da função de ligação.

Caso o modelo não esteja adequado, deve-se redefinir os fatores a serem incluídos no mesmo. Se necessário, redefinir a função de ligação e de variância.

3.4 QUALIDADE DO AJUSTAMENTO

Para verificar a qualidade do modelo final ajustado, recomenda-se a análise do critério de Akaike (ver seção 2.2.8.9) e da estatística Qui-Quadrado Generalizada de Pearson (ver seção 2.2.8.5).

3.5 MODELAGEM CONJUNTA DE MÉDIA E VARIÂNCIA

A proposta iterativa de modelagem conjunta de média e variância, conforme Lee e Nelder (1998; ver seção 2.3.1.2) considera, basicamente, o ajustamento de um modelo para média e posteriormente um modelo para variância de uma forma sistemática e iterativa. O ajustamento de cada um desses modelos deve seguir as recomendações já descritas neste capítulo.

Em suma, o procedimento proposto por Lee e Nelder (1998) consiste, inicialmente, na modelagem de um modelo saturado para média. Os desvios de tal modelo devem ser utilizados para a modelagem do modelo da variância, o qual deve seguir uma distribuição de probabilidade Gama com função de ligação logarítmica. O inverso do parâmetro de dispersão do modelo da variância deve ser utilizado para ponderar a variável resposta e, assim, gerar um

novo modelo para média através das equações de *quase*-verossimilhança. Os resíduos desse modelo da média serão utilizados para modelagem de um novo modelo para variância. O procedimento iterativo deve ser finalizado quando os parâmetros do modelo da variância forem iguais em duas iterações consecutivas.

A proposta de Lee e Nelder (1998) é aplicada a dados obtidos de experimentos fatoriais completos com repetição. Sendo o objetivo dessa dissertação ajustar um GLM para fatoriais fracionados sem repetições foi necessário alterar algumas etapas propostas pelos autores, as quais serão descritas na seqüência.

3.5.1 Ajustamento do modelo inicial para a média

Lee e Nelder (1998) recomendam que o primeiro modelo da média utilizado para iniciar o processo iterativo seja um modelo saturado. Isso é possível mediante utilização de dados oriundos de experimentos nos quais os tratamentos experimentais são repetidos. Como o objetivo desta dissertação é modelar dados fracionados sem repetição, sugere-se iniciar o processo iterativo pelos fatores com maiores efeitos estimados, que conseqüentemente são os mais importantes no ajustamento. A Tabela 5 apresenta a estimação dos efeitos, considerando um experimento fatorial completo com quatro fatores, e apresenta os tratamentos vinculados a cada efeito (ou seja, que possuem o mesmo efeito, porém com o sinal invertido). O modelo ajustado sob estas condições será denominado “modelo inicial da média”.

Tabela 5: Efeitos dos fatores

Tratamento	Efeito estimado	Tratamentos vinculados
A	-0,4825	-
BC	-0,4625	AD
B	-0,3025	-
C	0,2175	-
AB	0,1225	CD
AC	0,0975	BD
D	-0,0475	-

Posteriormente, deve-se verificar as funções de ligação e de variância mais adequadas e selecionar os coeficientes significativos. Os graus de liberdade restantes são

utilizados para o modelo da variância. O parâmetro de dispersão ϕ_i é considerado constante para essa primeira modelagem. Ao se dar continuidade ao processo iterarivo, ϕ_i é considerado diferente para cada combinação dos níveis dos fatores (ver próxima seção 3.5.3).

O pacote computacional "R" permite ajustar as funções de distribuições de probabilidade com as respectivas funções de ligação apresentadas na Figura 10. As funções de variância são definidas pela distribuição de probabilidade, exceto para as distribuições *quase*, as quais são definidas na Figura 11.

Distribuição de probabilidade	**Funções de ligação canônica**	**Funções de ligação possíveis**
Binomial	Logit	Probit, cauchit, logaritmo e complementar loglog (cloglog)
Normal	Identidade	Logaritmo e inversa
Gama	Inversa	Identidade e logaritmo
Normal Inversa	$1/\mu^2$	Inversa, identidade e logaritmo
Poisson	Logarítmo	Identidade e raiz quadrada

Figura 10: Funções de probabilidade e funções de ligação do "R"

3.5.2 Ajustamento do modelo para a variância

A variável resposta do modelo para a variância é o desvio (d_i) do modelo proveniente do modelo para a média, ajustado anteriormente.

Segundo Lee e Nelder (1998), a escolha mais usual de modelo para a variância é a distribuição Gama com função de ligação logarítmica e parâmetro de escala 2. Nelder e Lee (1991) sugerem a distribuição Gama como escolha natural para a distribuição dos erros, particularmente quando d_i é usado como variável resposta.

O parâmetro de dispersão desse modelo será utilizado no ajustamento do próximo modelo para a média.

3.5.3 Ajustamento do modelo para a média baseado no modelo para a variância

O inverso do parâmetro de dispersão do modelo da variância ($1/\phi_i$) previamente ajustado é utilizado para ponderar a variável resposta e ajustar as equações de *quase-*

verosimilhança. Lembra-se que o GLM utiliza os mínimos quadrados ponderados para o ajustamento dos seus coeficientes.

Só é possível calcular o parâmetro de dispersão ϕ_i para cada observação quando se possui repetição do experimento. Nesses casos é possível estimar a variância de cada uma das respostas e usá-las para ajuste de um modelo, independentemente do modelo da média (VIEIRA, 2004).

O parâmetro de dispersão ϕ é supostamente conhecido para cada observação (CORDEIRO, 1986). A função $a(\phi)$, que identifica a parcela da variabilidade de uma distribuição de probabilidade pertencente à Família Exponencial (ver seção 2.2), é a forma generalizada de $a(\phi) = \phi.\text{w}$, onde *w* é uma constante conhecida (ou seja, um peso conhecido *a priori*) e ϕ o parâmetro de dispersão do modelo.

Nelder e Lee (1991) recomendam utilizar como ponderação da variável resposta o inverso dos valores ajustados (ou seja, preditos pelo modelo). Ao se multiplicar o inverso dos valores ajustados pelo parâmetro de dispersão ϕ, será possível obter uma estimativa da dispersão de cada observação, uma vez que não se dispõe de repetições.

Definida a variável de ponderação e utilizando a variável resposta original sob análise, deve-se ajustar a função de *quase*-verossimilhança mais adequada. O pacote computacional "R" permite ajustar as funções de *quase*-verossimilhança combinadas com as funções de ligação e de variância apresentadas na Figura 11.

Distribuição de probabilidade	**Funções de ligação canônica**	**Funções de ligação possíveis**	**Funções de variância possíveis**
Quase	Logaritmo	Logit, probit, cloglog, identidade, inversa, logaritmo, $1/\mu^2$ e raiz quadrada	constante, $\mu(1-\mu)$, μ, μ^2 e μ^3
Quase-Binomial	Identidade (função de variância constante)		
Quase-Poisson	Logaritmo		

Figura 11: Ajustamento de funções de *quase*-verossimilhança no "R"

Conforme McCullagh e Nelder (1989), a função de *quase*-verossimilhança estendida possui as mesmas propriedades da função de *quase*-verossimilhança. Assim, para testar os coeficientes do modelo da média utiliza-se a *quase-deviance* (descrita no item 2.2.8.3), substituindo a função de *quase*-verossimilhança pela de *quase*-verossimilhança extendida

(seção 2.2.6). Para amostras pequenas, ambas as metodologias de estimação fornecem valores aproximados.

3.5.4 Final do processo iterativo

Lee e Nelder (1998) recomendam que o processo iterativo termine quando os parâmetros do modelo da variância forem iguais em duas iterações consecutivas, dada uma tolerância estabelecida.

3.6 FLUXOGRAMA DO ROTEIRO DE MODELAGEM CONJUNTA DE MÉDIA E VARIÂNCIA

Nesta seção é apresentado um fluxograma da metodologia descrita na seção anterior para realizar a modelagem conjunta de média e variância. O fluxograma apresentado na Figura 12 descreve o processo de modelagem utilizando GLMs, e a Figura 13 resume o processo de modelagem conjunta da média e dispersão utilizando GLMs, através do processo iterativo descrito na seção 2.3.2.1.

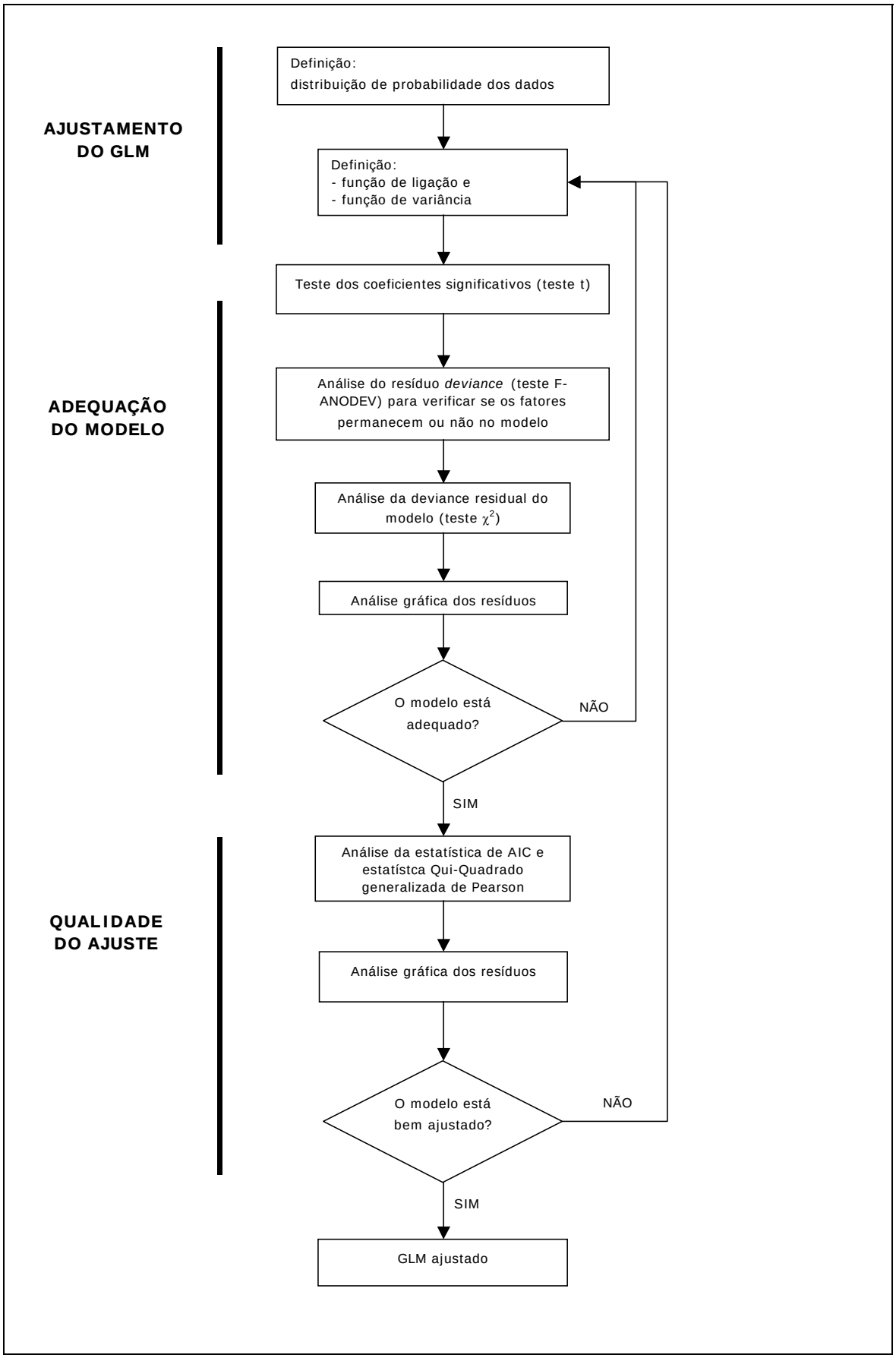


Figura 12: Fluxograma do roteiro de modelagem de um GLM

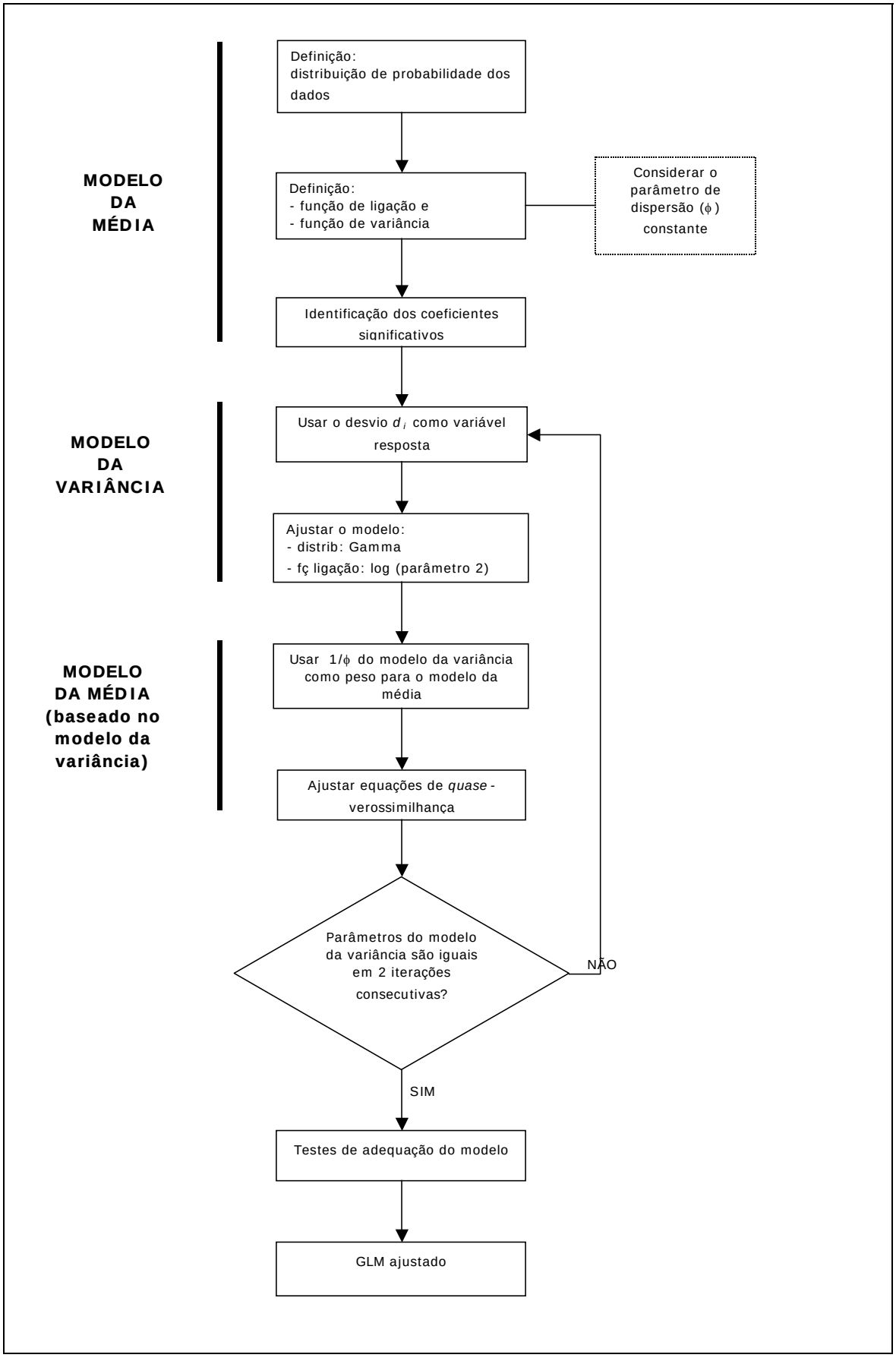


Figura 13: Fluxograma do roteiro de modelagem conjunta de média e variância

4 ESTUDO DE CASO

Este capítulo apresenta a aplicação do roteiro de modelagem descrito no Capítulo 3. O estudo de caso foi realizado com base nos dados apresentados em Pizzolato (2002).

4.1 DADOS PARA O ESTUDO DE CASO

Os dados utilizados são provenientes de um estudo realizado em uma empresa multinacional, fabricante e fornecedora de equipamentos e serviços agropecuários em âmbito mundial. O estudo de caso executado pela autora foi realizado na unidade brasileira situada no interior do estado do Rio Grande do Sul. O produto analisado é um piso plástico utilizado em ambientes de criação de suínos. O processo de injeção do piso plástico é feito em uma máquina injetora, onde é colocada a matéria-prima em forma de grãos. A matéria-prima é aquecida a uma temperatura tal que permita a sua injeção em um molde, o qual possui o formato do produto final. Posteriormente, esta ferramenta de injeção é aberta para permitir a descarga do produto injetado.

Os dados em Pizzolato (2002) são oriundos de um experimento realizado na empresa em junho/2000 e em outubro/2001, como parte de um projeto de melhoria do produto. O objetivo principal do projeto era a obtenção de pisos mais duros e resistentes ao cisalhamento (ou seja, resistentes a tensão que age tangencialmente às faces do piso), sem alteração substancial em sua composição de custos.

A partir do conhecimento do mercado, a equipe técnica da empresa definiu as características de qualidade consideradas importantes para o bom desempenho do produto e fez comparações com marcas já existentes. Uma equipe multifuncional (formada por funcionários dos setores de projeto, de processo e de atendimento ao cliente) analisou as

características importantes para o cliente final. Assim, obteve-se a avaliação das características de qualidade em relação à sua importância, tendo como base o conhecimento do grupo multifuncional, e em relação às demandas definidas pelos clientes. Por fim, as características de qualidade mais importantes demandadas pelo cliente foram traduzidas em variáveis respostas (VRs), com seus respcectivos valores alvo e especificações.

Dentre as VRs apresentadas no trabalho de Pizzolato (2002) optou-se por analisar a variável custo e as variáveis relacionadas ao impacto. A VR custo atende a demanda de oferecer ao cliente um produto de qualidade com um preço adequado. As VRs relacionadas ao impacto medem a resistência da superfície do material ao impacto, sendo que materiais resistentes ao impacto têm menor possibilidade de sofrer escamações, possuindo maior durabilidade.

Os fatores controlados explorados no experimento são: tempo de resfriamento (A), temperatura do fluído (B), percentual de elastômero (C) e percentual de talco (D). Os dois primeiros fatores definem as condições do processo de injeção dos pisos plásticos e os dois últimos são matérias primas utilizadas na composição do produto. Os parâmetros de processo mantidos constantes no experimento são velocidade de injeção e pressão de injeção. Os fatores de ruídos do experimento medidos foram a temperatura do molde e a temperatura do dia (PIZZOLATO, 2002).

O projeto experimental escolhido por Pizzolato (2002) para a coleta de dados foi um projeto fatorial 2^4 dividido em dois blocos, com o objetivo de eliminar o efeito da temperatura no ambiente (fator de ruído) na execução das rodadas experimentais. Além disso, a autora adicionou um ponto central, no qual níveis intermediários dos fatores controláveis foram testados para verificar a falta de ajustamento dos dados a um modelo linear. No caso em estudo não foram realizadas repetições das rodadas experimentais devido à inviabilidade econômica e à dificuldade de interromper a produção para a realização dos ensaios. A matriz experimental, que contempla um projeto 2^4 mais um ponto central, é apresentada na Figura 14 e as variáveis respostas são apresentadas na Tabela 6.

	Fatores Fixos		Fatores Controláveis (FC)				FC Codificados				Fatores Ruído	
Rodada	Pressão Injeção	Velocidade Injeção	(A) Tempo Resfriamento (s)	(B) Temperatura Fluído (ºC)	(C) % Elastômero	(D) % Talco	(A) Tempo Resfriamento (s)	(B) Temperatura Fluído (ºC)	(C) % Elastômero	(D) % Talco	Temperatura Molde (ºC)	Temperatura dia (ºC)
1	60	1200	70	209	0	0	-1	-1	-1	-1	24,8	9
2	60	1200	90	210	5	3	1	-1	1	1	24,5	10
3	60	1200	70	251	5	3	-1	1	1	1	23,3	11
4	60	1200	90	253	0	0	1	1	-1	-1	26,2	11
5	60	1200	70	212	5	0	-1	-1	1	-1	25,3	15
6	60	1200	90	210	0	3	1	-1	-1	1	24,4	15
7	60	1200	70	255	0	3	-1	1	-1	1	28	16
8	60	1200	90	251	5	0	1	1	1	-1	27,4	16
9	60	1200	80	231	2,5	1,5	0	0	0	0	29,3	15
10	60	1200	90	211	0	0	1	-1	-1	-1	26,6	13
11	60	1200	70	210	5	3	-1	-1	1	1	27,5	13
12	60	1200	70	250	0	0	-1	1	-1	-1	31,4	12
13	60	1200	90	250	5	3	1	1	1	1	27,5	12
14	60	1200	70	210	0	3	-1	-1	-1	1	27	9
15	60	1200	90	210	5	0	1	-1	1	-1	25,5	8
16	60	1200	70	250	5	0	-1	1	1	-1	27,8	7
17	60	1200	90	250	0	3	1	1	-1	1	27,8	6

Fonte: Pizzolato *et al.*, 2001

Figura 14: Matriz experimental

Tabela 6: Dados das variáveis respostas utilizados para análise

Rodada Exp.	Custo (R$/m2)	Impacto A (kN)	Impacto B (mm)	Impacto C (J)
1	28,19	1,42	3,76	3,09
2	29,03	0,86	3,76	3,01
3	28,67	2,01	2,13	2,27
4	28,76	2,12	2,55	2,82
5	28,46	1,42	2,27	1,79
6	28,98	1,41	2,77	2,24
7	28,41	1,62	2,29	1,79
8	29,04	1,70	2,40	2,17
9	28,72	0,97	3,64	2,14
10	28,76	0,96	4,88	2,84
11	28,67	1,03	1,68	0,86
12	28,19	1,38	2,43	1,79
13	29,03	0,99	1,59	0,76
14	28,41	1,71	1,53	1,24
15	29,04	1,67	2,06	1,70
16	28,46	2,15	2,38	2,70
17	28,98	0,97	1,42	0,65

Adaptado Pizzolato (2002)

Pizzolatto (2002) ajusta modelos de regressão linear múltipla para as quatro variáveis respostas apresentadas na Tabela 6 (as quais foram utilizadas como referência para a realização deste trabalho). A autora considerou como significativo os fatores com nível de significância inferior à 10% ($p < 0,10$) e analisou o coeficiente de determinação (R^2) para verificar o percentual da variabilidade total das respectivas VRs explicados pelo modelo ajustado, para encontrar os modelos mais adequados. Os resultados de cada uma destas VRs são apresentados a seguir, juntamente com uma breve descrição da mesma:

- **VR Custo (R$/m^2)**: foi obtida através do cálculo do custo para cada combinação dos fatores de controle. Esse modelo explicou 96,25% da variabilidade total da VR Custo (R^2= 96,52%).

Tabela 7: Análise de regressão linear múltipla para a VR custo (R$/m^2)

Parâmetros	Coef.	Erro padrão	Estatística *t*	Valor – *p*
Interseção	28,69	0,014937	1920,94	< 0,0001
(C) elastômero	0,11	0,015397	6,98176	< 0,0001
(D) talco	0,08	0,015397	5,19573	0,0002
(A) tempo	0,26	0,015397	16,8861	< 0,0001

Fonte: Pizzolato (2002).

- **VR Impacto**: Os ensaios da VR impacto foram realizados em equipamento próprio, pela empresa fornecedora da matéria-prima que compõe o produto. Esses ensaios foram sub-divididos em três VRs, devido à necessidade técnica e a forma com que a máquina de impacto media esta grandeza. Estas VRs, com os respectivos modelos ajustados pela autora e o coeficiente de determinação de cada modelo, são apresentadas nas Tabelas 8, 9 e 10.

a) **VR Impacto A:** Carga máxima dada em kilo-newtons: R^2=55,82%.

Tabela 8: Análise de regressão linear múltipla para a VR impacto para carga máxima (kN)

Parâmetros	Coef.	Erro padrão	Estatística *t*	Valor *p*
Interseção	1,43	0,082789	17,3295	< 0,0001
(A) tempo	-0,13	0,085337	-1,50871	0,1595
(B) temperatura	0,15	0,085337	1,80167	0,0990
(D) talco	-0,14	0,085337	-1,62589	0,1323
(AD) tempo × talco	-0,14	0,085337	-1,62589	0,1323
(ABC) tempo × temperatura × elastômero	-0,15	0,085337	-1,75772	0,1066

Fonte: Pizzolato (2002).

b) **Impacto B**: Deflexão da carga máxima, dada em milímetros: R^2=66,44%.

Tabela 9: Análise de regressão linear múltipla para a VR impacto para carga máxima (mm)

Parâmetros	Coef.	Erro padrão	Estatística *t*	Valor *p*
Interseção	2,56	0,159714	16,036	< 0,0001
(B) temperatura	-0,34	0,16463	-2,09561	0,0601
(D) talco	-0,35	0,16463	-2,11087	0,0585
(AB) tempo × temperatura	-0,34	0,16463	-2,08802	0,0608
(BD) elastômero × talco	0,35	0,16463	2,14876	0,0548
(BCD) temperatura × elastômero × talco	-0,33	0,16463	-1,98931	0,0721

Fonte: Pizzolato (2002)

c) **Impacto C**: Energia de Carga Máxima, dada em joules: R^2=98,20%.

Tabela 10: Análise de regressão linear múltipla para a VR impacto pela energia de carga máxima (J)

Parâmetros	Coef.	Erro padrão	Estatística *t*	Valor *p*
Interseção	1,99	0,04134	48,1795	<0,0001
(B) temperatura	-0,11	0,04261	-2,6694	0,0371
(D) talco	-0,38	0,04261	-8,9175	0,0001
(AB) tempo × temperatura	-0,31	0,04261	-7,2748	0,0003
(BC) temperatura × elastômero	0,18	0,04261	4,2534	0,0054
(BD) temperatura × talco	-0,12	0,04261	-2,8434	0,0294
(CD) elastômero × talco	0,20	0,04261	4,6347	0,0036
(ABC) tempo × temperatura × elastômero	-0,20	0,04261	-4,7521	0,0032
(ABD) tempo × temperatura × talco	-0,41	0,04261	-9,7389	0,0001
(ACD) tempo × elastômero × talco	0,14	0,04261	3,1973	0,0187
(BCD) temperatura × elastômero × talco	-0,16	0,04261	-3,6667	0,0105

Fonte: Pizzolato (2002)

4.2 ADAPTAÇÃO DO EXPERIMENTO PARA REALIZAÇÃO DO ESTUDO DE CASO

Como o objetivo do presente estudo é realizar uma modelagem pelos GLMs utilizando projetos fatoriais fracionados, fracionou-se o experimento apresentado na Figura 14, de forma a viabilizar a análise das características da modelagem proposta nesta dissertação. O experimento foi dividido em dois blocos, a partir do contraste de definição ABCD do fatorial completo 2^4. A escolha do contraste de definição se justifica, pois o estudo de Pizzolato (2002) demonstrou que tal interação não é significativa. Assim, o experimento analisado foi um projeto fatorial 2^{4-1}.

Optou-se por analisar a VR denominada "Impacto A", por ter apresentado o pior ajustamento dentre os modelos obtidos por Pizzolato (2002). O ajustamento foi analisado a partir da significância dos fatores que entraram no modelo e o coeficiente de determinação (R^2), que explica o percentual da variabilidade total dos dados explicada pelo modelo. No caso, a VR "Impacto A" apresentou os coeficientes de regressão com menor significância e o menor coeficiente de determinação.

Analisando todas as observações da VR “Impacto A” através da Distância de Cook, identificou-se que as observações oriundas das rodadas experimentais 4, 15 e 16 eram destoantes do restante do banco de dados. Uma vez que tais observações faziam parte da do bloco da fração principal, escolheu-se a fração secundária do fracionamento para a aplicação da modelagem utilizando GLMs. Os dados a serem utilizados na modelagem conjunta da média e da variância são apresentados na Tabela 11.

Tabela 11: Fração do projeto experimental fracionado a ser modelado

Rodada	A	B	C	D	Impacto A
2	1	-1	1	1	0,86
3	-1	1	1	1	2,01
5	-1	-1	1	-1	1,42
8	1	1	1	-1	1,70
10	1	-1	-1	-1	0,96
12	-1	1	-1	-1	1,38
14	-1	-1	-1	1	1,71
17	1	1	-1	1	0,97

4.3 MODELAGEM CONJUNTA DE MÉDIA E VARIÂNCIA

A realização do estudo de caso é apresentada conforme as etapas do método proposto no Capítulo 3 desta dissertação. Nesta seção, apresentam-se os resultados da modelagem conjunta de um experimento fracionado sem repetição, utilizando GLMs através do pacote computacional “R”, um *software* livre para análises estatísticas e gráficas. O cálculo da estatística Qui-Quadrado Generalizado de Pearson foi obtido através do pacote computacional SAS v.8.

Tabela 12: Notação para o nível de significância dos coeficientes

Notação	Nível de Significância (NS)			
***	0	≤ NS <	0,001
**	0,001	≤ NS <	0,01
*	0,01	≤ NS <	0,05
.	0,05	≤ NS <	0,1
		NS >	0,1

Os modelos ajustados na seqüência irão utilizar a notação apresentada na Tabela 12 para definir o nível de significância do respectivo coeficiente.

4.3.1 Ajustamento do modelo inicial para a média

Utilizando os dados da Tabela 11, ajustou-se um modelo para a média com os fatores que possuíam os maiores efeitos. Para tanto, empregou-se um método de eliminação retroativa, respeitando o número de graus de liberdade disponíveis, já que o banco de dados é formado por apenas 8 observações. A melhor distribuição de probabilidade que se ajustou aos dados foi uma Normal Inversa com função de ligação identidade. Os coeficientes do modelo ajustado são apresentados na Tabela 13.

Tabela 13: Modelo inicial para a média

Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	1,38206	0,04827	28,634	<0,0001 **
A	-0,27187	0,04190	-6,488	0,0074 **
BC	0,20726	0,04250	4,876	0,0165 *
B	0,17655	0,04248	4,156	0,0253 *
C	0,13666	0,04245	3,219	0,0486 *

Posteriormente, foi realizada a análise dos resíduos *deviance* para avaliar o efeito da inclusão de cada termo sobre a redução da *deviance* residual do modelo. Para tanto, utilizou-se a Análise de *Deviance* (ANODEV), com resultados apresentados na Tabela 14.

Tabela 14: Análise de d*eviance* (ANODEV) para o modelo inicial para a média

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo		0,20456	7	0,52189		
A	1	0,17560	6	0,31733	35,758	0,0094 **
BC	1	0,05832	5	0,14173	30,695	0,0116 *
B	1	0,06642	4	0,08341	10,194	0,0496 *
C	1	0,20456	3	0,01699	11,611	0,0422 *

Na Tabela 14 observa-se que a *deviance* residual do modelo é 0,016992 com 3 graus de liberdade. Ao se comparar esse resultado com a distribuição Qui-Quadrado com 3 graus de

liberdade ao nível de significância de 5% ($\chi^2_{(3;0,05)} = 7,815$), constata-se que o modelo é adequado.

Analisando o gráfico apresentado na Figura 15, observa-se que os resíduos possuem um comportamento aleatório em relação aos valores preditos e o gráfico da Figura 16 aponta um pequeno desvio à distribuição de probabilidade definida a *priori* (no caso, a distribuição Normal Inversa). A análise da Distância de Cook, apresentada na Figura 17, indica que a observação 5 e 8 destoam em relação as demais (Distância de Cook superior a 0,5). Este fato, entretanto, será desconsiderado, por se tratar de um modelo inicial.

Para verificar a qualidade do modelo final ajustado, analisou-se os valores do critério de Akaike (AIC) e da estatística Qui-Quadrado Generalizada de Pearson. Para o modelo na Tabela 13, obteve-se um valor de AIC igual a (–7,8848) e um valor de χ^2 Generalizado de Pearson igual a 0,0173, indicando, assim, um bom ajustamento.

As instruções de uso do pacote computacional "R" para obtenção dos resultados apresentados encontra-se no Apêndice B.

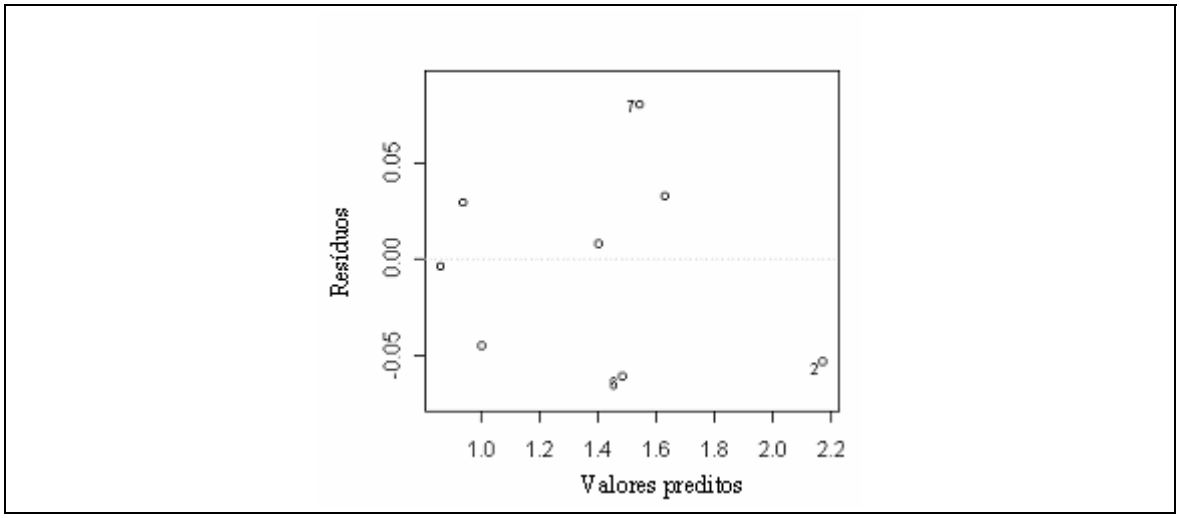


Figura 15: Análise dos resíduos × valores ajustados para o modelo inicial para a média

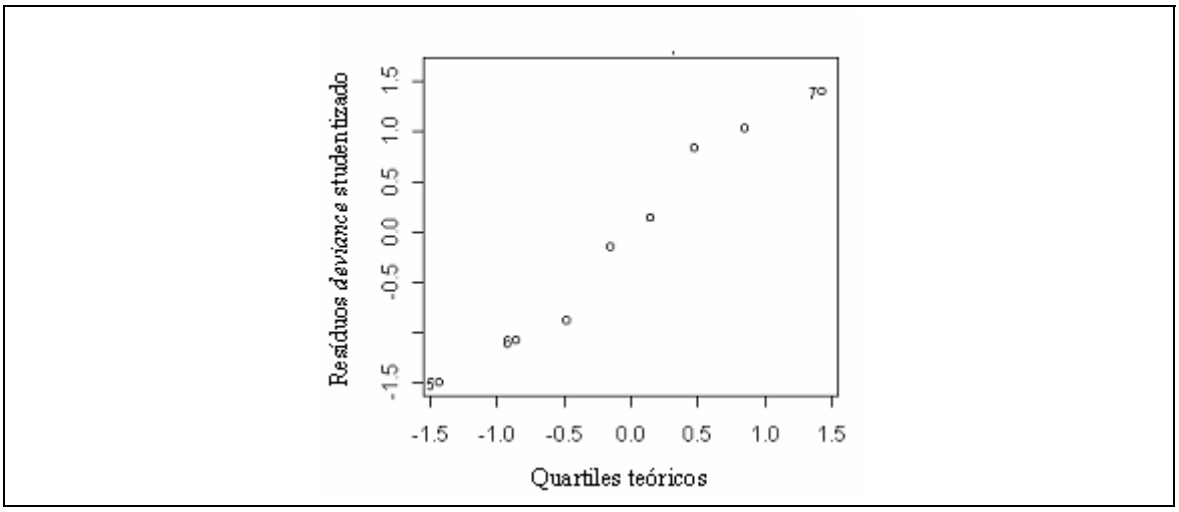


Figura 16: Gráfico de Probabilidade Normal para o modelo inicial para a média

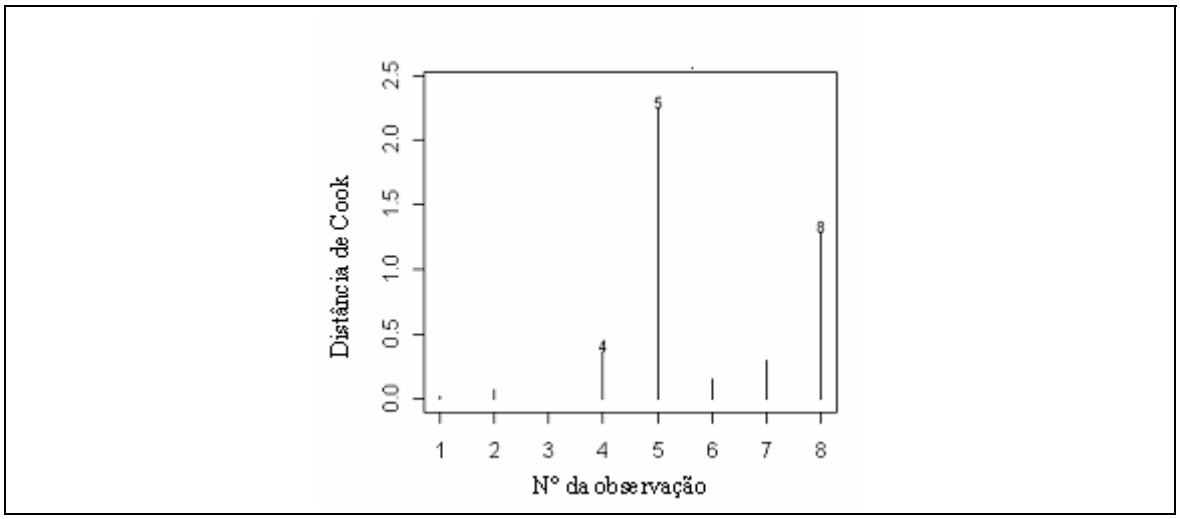


Figura 17: Análise das Distâncias de Cook para o modelo inicial para a média

4.3.2 Processo iterativo

A partir dos resíduos do modelo inicial para a média foi gerado um modelo para a variância. Utilizando o parâmetro de dispersão do modelo da variância, através de uma função de *quase*-verossimilhança com ligação identidade, ajustou-se um novo modelo para a média. Este por sua vez, foi ponderado pela razão do parâmetro de dispersão do modelo anteriormente ajustado para a dispersão pelos valores ajustados para o modelo da média. O processo convergiu (ou seja, gerou dois modelos de dispersão com coeficientes similares) na quarta iteração, sendo que cada processo de iteração gera um modelo para a média e um para a variância. Uma vez que se verificou que os erros obedecem a distribuição Normal Inversa, espera-se que a variância não dependa da média. Os modelos intermediários são apresentados

no Apêndice A. Nas seções seguintes são apresentados, respectivamente, o penúltimo modelo ajustado para a variância (como uma evidência de convergência do processo iterativo), o modelo final ajustado para a média e o final para a variância. Esses modelos finais foram obtidos na quarta iteração.

4.3.3 Ajustamento do penúltimo modelo para a variância

Utilizando como variável resposta o resíduo do terceiro modelo ajustado para a média, ajustou-se novamente um GLM para a variância com distribuição Gama e função de ligação logarítmica. O parâmetro de dispersão desse modelo será utilizado para ponderar a variável resposta "Impacto A" no próximo ajustamento do modelo para a média.

O modelo ajustado pode ser visto na Tabela 15 e a análise da significância de cada termo no modelo pode ser analisada na Tabela 16, através da analise da *deviance*. Observa-se que a inclusão de todos os termos no modelo é significativa e que a *deviance* residual é pequena (0,01341). Esse ajustamento forneceu um AIC igual a –100,92, evidenciando a qualidade do ajuste.

Tabela 15: Penúltimo modelo para a variância

Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	-5,28049	0,02363	-223,511	< 0,0001 ***
A	-0,09946	0,02363	-4,210	0,0249 *
BC	0,52351	0,02363	22,159	0,0002 ***
B	0,19218	0,02363	8,134	0,0039 **
C	-0,31043	0,02363	-13,140	0,0009 ***

Tabela 16: Análise de d*eviance* (ANODEV) para o penúltimo modelo para a variância

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	2,78717		
A	1	0,06766	6	2,71951	15,153	0,0301 *
BC	1	1,68007	5	1,03944	376,258	0,0003 ***
B	1	0,26718	4	0,77227	59,835	0,0045 **
C	1	0,75885	3	0,01341	169,949	0,0010 ***

4.3.4 Ajustamento do modelo final para a média

Nesta etapa, o inverso dos valores ajustados para o último modelo da média gerado vezes o parâmetro de dispersão do modelo do último modelo ajustado para a variância são utilizados novamente como ponderadores para a variável resposta (VR) “Impacto A”, obtendo um novo modelo para a média da VR em questão.

Seguindo a metodologia apresentada no Capítulo 3, procedeu-se ao ajustamento das equações de *quase*-verosimilhança. As equações de *quase*-verosimilhança que melhor se ajustaram aos dados ponderados permanecem sendo aquelas que utilizaram a função de ligação identidade. A Tabela 17 apresenta o modelo final obtido para a média.

Tabela 17: Modelo final para a média

Fatores	**Estimativa**	**Erro Padrão**	**Estatística t**	**Pr(>|t|)**
(Intercepto)	1,37625	0,04652	29,582	< 0,0001 ***
A	-0,26092	0,04564	-5,717	0,0106 *
BC	0,21480	0,04566	4,704	0,0182 *
B	0,15229	0,04565	3,336	0,0445 *
C	0,12645	0,04565	2,770	0,0696 .

Novamente, realizou-se a análise dos resíduos da *deviance* através da Análise de *Deviance* (ANODEV) apresentada na Tabela 18, onde observa-se que a *deviance* residual do modelo é de apenas 0,00016914, com 3 graus de liberdade. Comparando com uma distribuição Qui-Quadrado com 3 graus de liberdade e nível de significância de 5% , conclui-se que o modelo final obtido para a média é adequado já que a falta de ajuste (*deviance*) não é significativa. O comportamento aleatório dos resíduos apresentados na Figura 18 confirma o bom ajustamento do modelo. O Gráfico de Probabilidade Normal apresentado na Figura 19 mostra um razoável ajuste a distribuição Normal Inversa. A análise da Distância de Cook, apresentada na Figura 20, indica novamente que a observação 5 destoa dos demais dados. Entretanto, ao se retirar tal observação do conjunto de dados, o modelo obtido apresenta pior ajuste.

Tabela 18: Análise de *deviance* (ANODEV) para o modelo final da média

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	0,00078374		
A	1	0,00038193	6	0,00040181	34,8677	0,0097 **
BC	1	0,00019200	5	0,00020981	17,5282	0,0248 *
B	1	0,00009290	4	0,00011691	8,4816	0,0619 .
C	1	0,00008405	3	0,00003286	7,6730	0,0696 .

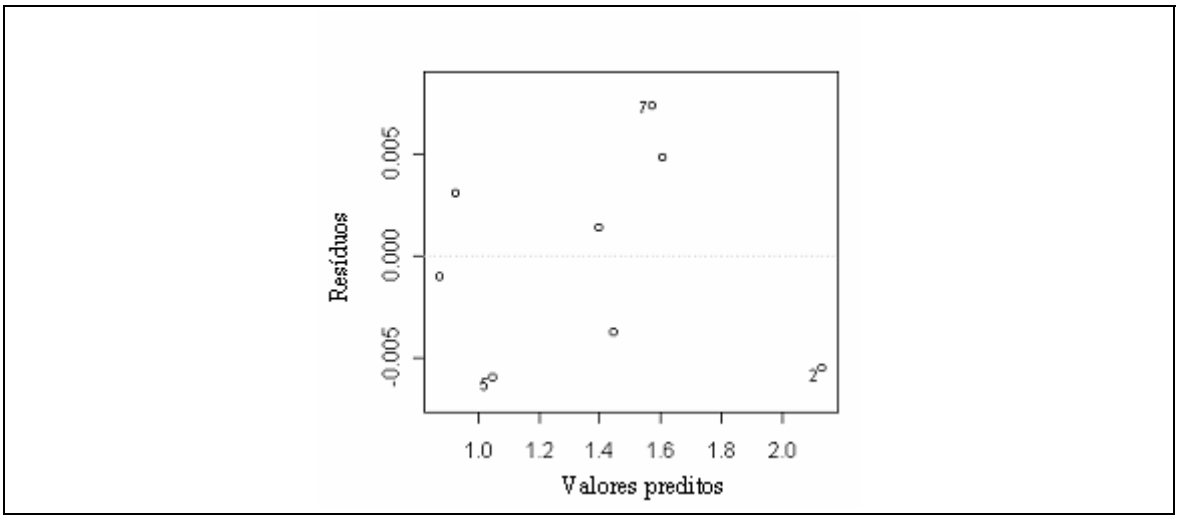


Figura 18: Análise dos resíduos × valores ajustados para o modelo final para a média

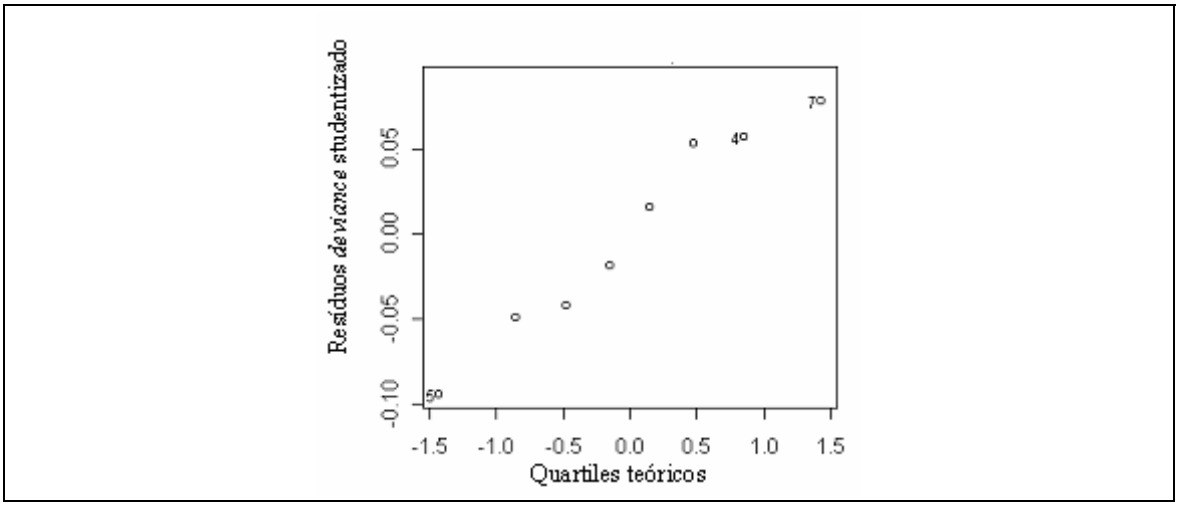


Figura 19: Gráfico de Probabilidade Normal para o modelo final para a média

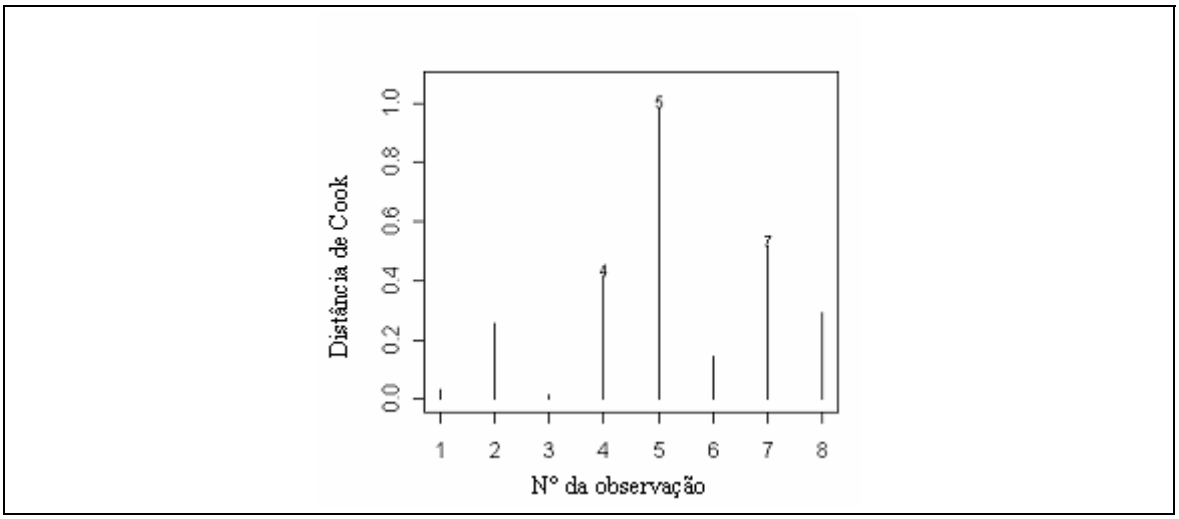


Figura 20: Análise das Distâncias de Cook para o modelo final para a média

4.3.5 Ajustamento do modelo final para a variância

Para ajustar o modelo da variância utilizou-se os valores do resíduo d_i provenientes do modelo final ajustado para a média. Como sugerido por Lee e Nelder (1998), ajustou-se um modelo para variância pressupondo dados distribuídos conforme uma distribuição Gama, com função de ligação logarítmica e parâmetro de escala 2. O modelo final ajustado para a variância por GLM é apresentado na Tabela 19.

Tabela 19: Modelo final para a variância

Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	-5,67876	0,01043	-544,351	<0,0001 ***
A	-0,09986	0,01043	-9,572	0,0024 **
BC	0,53512	0,01043	51,295	<0,0001 ***
B	0,20265	0,01043	19,426	0,0003 ***
C	-0,32567	0,01043	-31,218	<0,0001 ***

Novamente, realizou-se a análise dos resíduos *deviance* através da Análise de *Deviance* (ANODEV) apresentada na Tabela 20, onde observa-se que a *deviance* residual do modelo é de apenas 0,00261, com 3 graus de liberdade, concluido-se que o modelo final obtido para a variância também é relevante. O comportamento aleatório dos resíduos (Figura 21) e o comportamento linear verificado no Gráfico de Probabilidade Normal (Figura 22) mostram um ajuste adequado do modelo para variância. O gráfico apresentado na Figura 23 mostra que as observações 2 e 4 apresentam o valor da Distância de Cook acima do tolerável

(0,5). Porém, também foi verificado que a exclusão de tais observações prejudicaria a qualidade do ajuste do modelo.

Tabela 20: Análise de *deviance* (ANODEV) para o modelo final para a variância

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	2,92308		
A	1	0,07142	6	2,85166	82,033	0,0028 **
BC	1	1,72054	5	1,13113	1976,168	< 0,0001 ***
B	1	0,29463	4	0,83649	338,409	0,0003***
C	1	0,83388	3	0,00261	957,775	< 0,0001 ***

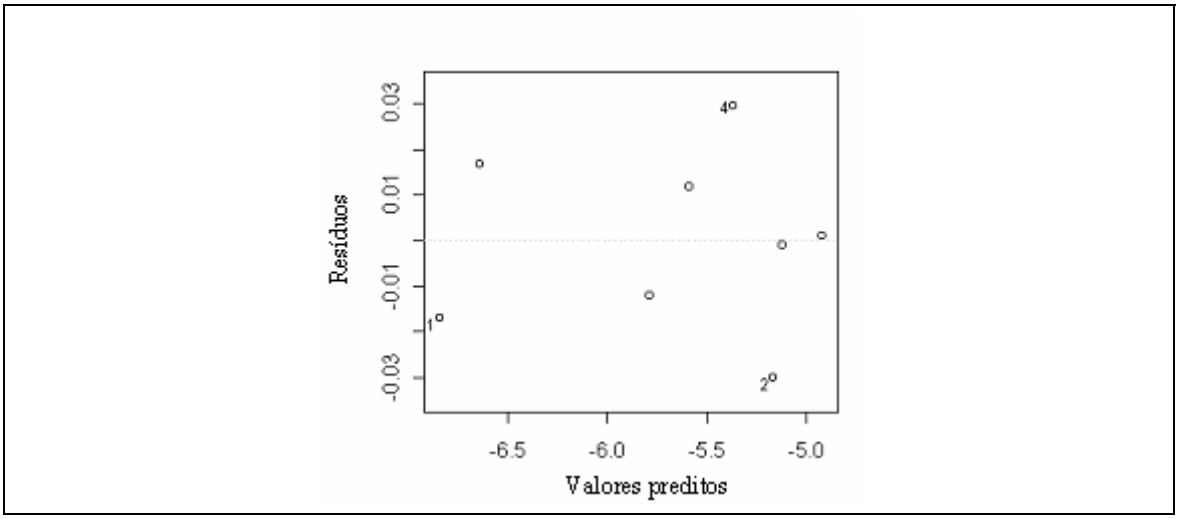


Figura 21: Análise dos resíduos × valores ajustados para o modelo final para a variância

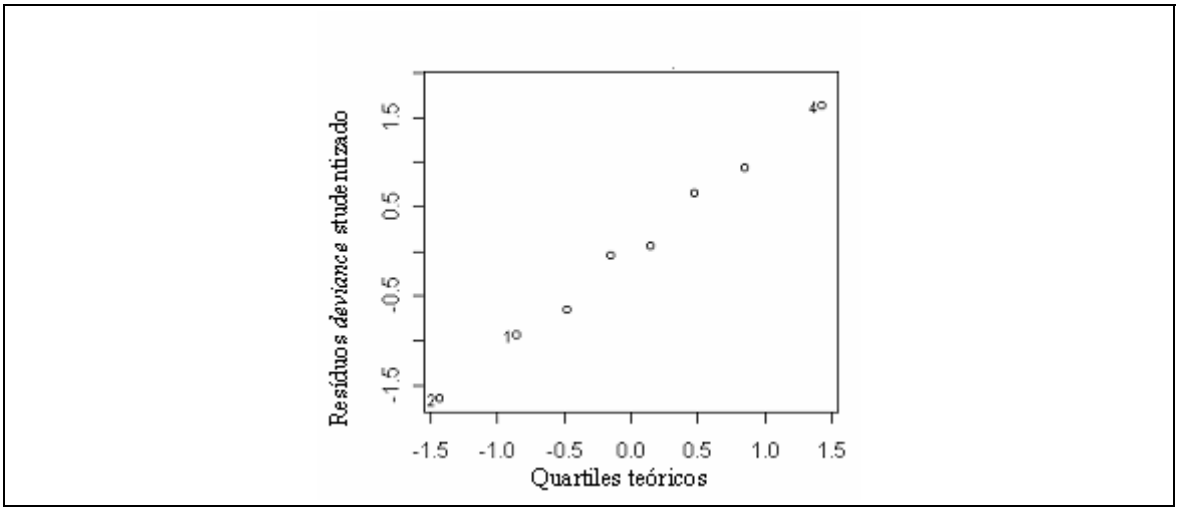


Figura 22: Gráfico de Probabilidade Normal para o modelo final para a variância

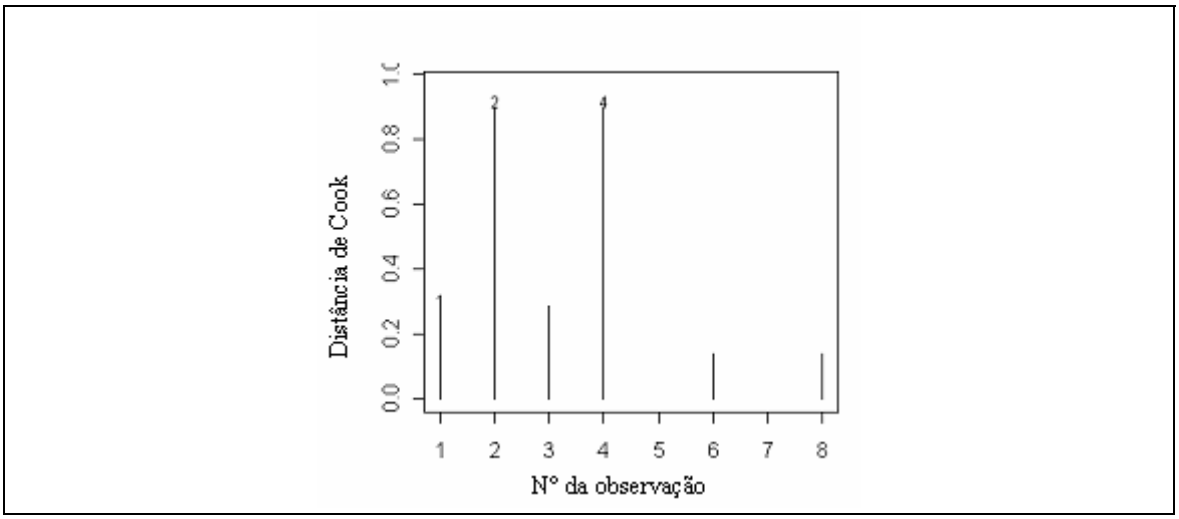


Figura 23: Análise das Distâncias de Cook para o modelo final para a variância

Para verificar a qualidade do modelo final ajustado, analisou-se o critério de Akaike (AIC) e a estatística Qui-Quadrado Generalizada de Pearson. Os valores obtidos foram AIC = –70,261 e χ^2 = 0,0026, indicando um bom ajustamento do modelo aos dados.

4.3.6 Convergência dos modelos ajustados

Conforme a metodologia proposta por Lee e Nelder (1998; ver seção 3.5), há convergência de modelos ajustados de forma conjunta para a média e para variância quando dois modelos para a variância ajustados seqüencialmente possuem parâmetros semelhantes. Ao se comparar o modelo apresentado na Tabela 15 com o da Tabela 19, observa-se uma diferença mínima entre os parâmetros estimados. Assim, os melhores modelos para explicar a média e a variabilidade da variável resposta "Impacto A" são aqueles apresentados, respectivamente, na Tabela 17 e Tabela 19.

Observa-se que ambos os modelos possuem os mesmos fatores significativos (A, BC, B e C), porém em grandezas distintas. Como o fator D não foi significativo, ele é utilizado como uma repetição do experimento durante a modelagem. Objetivando analisar apenas uma fração do experimento realizado por Pizzolatto (2002), analisou-se somente oito observações, impossibilitando a estimação de efeitos de segunda ordem (um vez que a interação de terceira ordem foi utilizada como contraste de definição no fracionamento do experimento analisado).

4.3.7 Modelo ajustado por regressão linear múltipla

Para comparar o modelo ajustado por Pizzolato (2002) com os modelos ajustados por GLM através do processo iterativo, analisaram-se os dados da autora, modelando um GLM com distribuição Normal e função de ligação canônica, o que corresponde a uma regressão linear múltipla. Assim, é possível verificar a adequação desse modelo através das mesmas medidas de diagnóstico utilizadas nos GLMs.

O modelo de regressão utilizando o fatorial completo ajustado por Pizzolato (2002) é apresentado na Tabela 21. Para este modelo observa-se um AIC igual a 18,299 e uma *deviance* residual de 1,28172 (conforme Tabela 22). Tais valores são bem superiores aos encontrados nos GLMs ajustados anteriormente (nas seções 4.3.4 e 4.3.5), mostrando melhor ajuste dos GLMs em relação a regressão múltipla. Conforme Pizzolato (2002), para o modelo desta variável resposta foram considerados como significativos termos com significância de até 0,16, por determinação técnica da equipe de trabalho.

Tabela 21: Modelo de regressão linear múltipla para o fatorial completo sem repetição

Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	143,471	0,08279	17,330	< 0,0001 ***
A	-0,12875	0,08534	-1,509	0,1600
B	0,15375	0,08534	1,802	0,0990 .
D	-0,13875	0,08534	-1,626	0,1320
AD	-0,13875	0,08534	-1,626	0,1320
ABC	-0,15000	0,08534	-1,758	0,1070

Tabela 22: Análise de *deviance* (ANODEV) para o modelo de regressão linear múltipla para o fatorial completo sem repetição

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			16	2,90122		
A	1	0,26522	15	2,63600	2,2762	0,1595
B	1	0,37822	14	2,25777	3,2460	0,0990
D	1	0,30803	13	1,94975	2,6435	0,1322
AD	1	0,30803	12	1,64172	2,6435	0,1322
AB	1	0,36000	11	1,28172	3,0896	0,1065

Supondo que o experimento de Pizzolato (2002) fosse fracionado ao meio e se analisasse a fração secundária (ou seja, os mesmos dados usados para modelar os GLMs ajustados anteriormente), o resultado mostraria que apenas os efeitos A e AD (-BC) são significativos, conforme apresentado na Tabela 23. As análises da significância da inclusão dos fatores no modelo podem ser vistas na Tabela 24.

Sabe-se que, como o banco de dados utilizados é oriundo de um projeto fatorial fracionado ao meio, o efeito AD corresponde ao efeito BC, porém com o sinal invertido (ou seja, AD e BC são efeitos vinculados).

Tabela 23: Modelo de regressão linear múltipla para o fatorial fracionado

Fatores	**Estimativa**	**Erro Padrão**	**Estatística t**	**Pr(>|t|)**
(Intercepto)	0,37625	0,08457	16,274 ***	0,0005
A	0,25375	0,08457	-3,001	0,0576 .
B	0,13875	0,08457	1,641	0,1994
D	0,01125	0,08457	0,133	0,9026
AD (-BC)	- 0,21875	0,08457	-2,587	0,0813 .
ABC	NA	NA	NA	NA

Tabela 24: Análise de *deviance* (ANODEV) para o modelo de regressão linear múltipla para o fatorial fracionado

	GL	**Redução da *Deviance***	**GL restante**	**Resíduo da *Deviance***	**Estatística F**	**Pr(>F)**
Modelo Nulo			7	1,22459		
A	1	0,51511	6	0,70948	9,0035	0,0576 .
B	1	0,15401	5	0,55546	2,6919	0,1994
D	1	0,00101	4	0,55445	0,0177	0,9026
AD (-BC)	1	0,38281	3	0,17164	6,6911	0,0813 .
ABC	1	0,00000	3	0,17164		

Se os efeitos não relevantes fossem retirados do modelo, o modelo final ajustado seria igual ao apresentado na Tabela 25 e as análises, quanto à qualidade do ajustamento, podem ser observados nas Tabela 26. Esse modelo fornece um fator a AIC igual a 5,1169 e uma *deviance* residual de 0,32666. Enfim, o modelo apresentaria apenas os efeitos A e AD (-BC) como significativos.

Tabela 25: Modelo de regressão linear múltipla para o fatorial fracionado ajustado

Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	1,37625	0,09037	15,229	< 0,0001 ***
A	-0,25375	0,09037	-2,808	0,0376 *
AD (-BC)	-0,21875	0,09037	-2,421	0,0601 .

Tabela 26: Análise de *deviance* (ANODEV) para o modelo de regressão linear múltipla para o fatorial fracionado ajustado

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo Nulo			7	1,22459		
A	1	0,51511	6	0,70948	7,8845	0,0376 *
AD (-BC)	1	0,38281	5	0,32666	5,8594	0,0601 .

4.3.8 Regressão linear múltipla x GLM

Comparando um modelo ajustado através de Regressão Linear Múltipla com os modelos ajustados por GLM, a partir de um mesmo banco de dados fracionados sem repetição, é possível afirmar que o GLM é capaz de identificar mais fatores significativos, com maior precisão nas estimativas. Além de apresentar níveis de significância menores (menor Sig.), os modelos ajustados por GLM apresentam menor valor de *deviance* residual, bem como menores erros padrões, indicando maior precisão do GLM; conforme pode ser observado na Figura 24.

Coeficientes	REGRESSÃO LINEAR MÚLTIPLA – fatorial fracionado			GLM (modelo final para a média) – fatorial fracionado		
	Estimativa	Erro Padrão	Sig.	Estimativa	Erro Padrão	Sig.
Intercepto	1,37625	0,09037	<0,0001	1,37625	0,04652	<0,0001
A	-0,25375	0,09037	0,0376	-0,26092	0,04564	0,0106
AD (-BC)	-0,21875	0,09037	0,0601	-0,21480	0,04566	0,0182
B	-	-	-	0,15229	0,04565	0,0446
C	-	-	-	0,12645	0,04565	0,0696
Deviance Residual	0,03266			0,000169		

Figura 24: Comparação dos modelos ajustados por regressão linear múltipla e por GLM

No contexto do estudo de caso utilizado neste trabalho (seção 4.1), observa-se uma alteração do melhor ajuste para a variável resposta “Impacto A” do processo industrial analisado, ao se otimizar o processo a partir dos resultados da modelagem GLM e usando o modelo de regressão linear múltipla (ver Figura 24). O modelo de regressão ajustado para o experimento fracionado identifica como significativos apenas os efeitos tempo de resfriamento (A) e a interação temperatura do fluído e percentual de elastômero (BC). Já o GLM, além de identificar como relevante para o ajuste do processo esses dois efeitos, identificou os efeitos principais da temperatura do fluído (B) e percentual de elastômero (C). Neste caso, onde os interações AD e –BC são vinculadas, é mais interessante ajustar o processo em função do ajuste da interação BC, pois os efeitos principais B e C também são significativos.

5 CONCLUSÕES FINAIS

5.1 CONCLUSÕES

Esta dissertação apresentou uma proposta de modelagem conjunta de experimentos fracionados sem repetição utilizando Modelos Lineares Generalizados (GLM). Tal proposta permite modelar separadamente, mas de forma dependente, um modelo para a média (parâmetro de localização) e um para a variância (parâmetro de dispersão).

Na modelagem de regressão tradicional, utiliza-se o método dos Mínimos Quadrados Ordinários e Máxima Verossimilhança para estimação dos parâmetros do modelo. Tais métodos pressupõem variância constante e Normalidade das respostas. Entretanto, sabe-se que tais suposições são freqüentemente violadas na prática, já que nem todos os fenômenos podem ser bem modelados supondo distribuição Normal. Uma solução simplista geralmente utilizada na prática é a metodologia de transformação proposta por Box & Cox (1964). Entretanto, foi demonstrado que o GLM pode ser uma alternativa para tais situações, pois permite modelar dados oriundos de distribuições de probabilidade pertencentes à Família Exponencial, a qual engloba distribuições discretas, assimétricas e binomiais, entre outras.

Nos últimos anos, foram desenvolvidos diversos procedimentos de modelagem conjunta de média e variância com o intuito de aperfeiçoar os métodos desenvolvidos por Taguchi. Diversos autores consideram que os métodos de Taguchi nem sempre são claros e eficientes. Diante disso, apresentam alternativas de modelagem conjunta, dentre elas a utilização de projetos fatoriais fracionados e do GLM.

A modelagem por GLM de experimentos fracionados sem repetição apresenta algumas vantagens e desvantagens. Na prática, a utilização de experimentos fracionados se

justifica pelo elevado custo e tempo gasto na coleta de dados de experimentos completos. No entanto, o fracionamento pode gerar dúvidas quanto à significância dos efeitos. Sabe-se que dúvidas quanto à eficiência de um fracionamento não se resolvem com repetição, pois os tratamentos repetidos serão os mesmos. A repetição somente permite, assim, aumentar a precisão da estimativa dos coeficientes. No geral, é mais interessante investir em um experimento completo do que na repetição de um experimento fracionado. Sendo o investimento em fatores completos caro e demoro, recomendasse os fatorias fracionados com repetição, quando for possível repetir os tratamentos.

Como descrito anteriormente, o GLM permite modelar dados que não apresentam distribuição Normal, fazendo com que se obtenha modelos mais precisos, como demonstrado no Capitulo 4. Vieira (2004) confirma a superioridade dos GLMs ao afirmar que os mesmos apresentam melhor desempenho na estimativa dos parâmetros, pois resultam em intervalos de confiança menores para as estimativas. Entretanto, sua forma de modelagem é mais complexa, uma vez que além de identificar a distribuição de probabilidade dos dados é necessário determinar a função de ligação mais adequada.

A proposta de modelagem conjunta da média e da variância apresentada neste trabalho baseou-se no artigo de Lee e Nelder (1998), porém, não foi possível aplicar tal metodologia de forma integral, pois os autores utilizaram um fatorial fracionado com repetição e a proposta desta dissertação é utilizar fatoriais fracionados sem repetição. Devido à falta de repetições, este trabalho buscou utilizar os mesmos princípios do trabalho de Lee e Nelder (1998), mostrando alternativas para lidar com a falta de dados a fim de iniciar o processo iterativo e estimar os parâmetros de dispersão de cada uma das observações .

Utilizando os fatores que possuíam os maiores efeitos para iniciar o processo iterativo e o inverso dos valores ajustados para estimar o parâmetro de dispersão, os modelos convergiram em apenas quatro iterações, apesar do conjunto de dados utilizado ser pequeno. Além disso, os modelos finais obtidos apresentaram um bom ajuste, apesar de não haver repetições de tratamentos. Entretanto, é relevante observar que as estatísticas utilizadas para verificar a adequação e qualidade do ajustamento são recomentadas para amostras grandes. A literatura encontrada sobre GLM não menciona estatísticas para analisar a modelagem de uma amostra pequena.

O estudo de caso apresentado mostrou a superioridade da modelagem por GLM em relação à utilização de modelos de regressão tradicionais, ou seja, a capacidade de identicar mais efeitos significativos, além de identificar efeitos com maior precisão. Sendo o GLM indicado para situações onde os dados não se ajustam a distribuição de probabilidade Normal

A fim de facilitar a compreensão das modelagens citadas e desenvolvidas na literatura, o trabalho apresenta uma revisão bibliográfica sobre projetos fatoriais fracionados e sobre GLM, além de um roteiro de modelagem através de tal metodologia. A modelagem foi ilustrada através de um estudo de caso utilizando uma rotina computacional programada no pacote "R", cujos comandos são apresentados nos Apêndices B e C.

5.2 SUGESTÕES PARA TRABALHOS FUTUROS

Uma linha futura de investigação seria propor a modelagem de modelos generalizados não lineares e de Modelos Generalizados Aditivos (GAM – *Generalized Additive Models*), conforme delineado a seguir.

Weisberg *apud* Vieira (2004) afirma que, caso não seja encontrado um GLM adequado para a média, pode-se utilizar os modelos não lineares generalizados. Neles tem-se a mesma estrutura dos GLMs, com exceção do polinômio de regressão, que é não-linear. Para ajustar o polinômio adequado, deve-se conhecer a relação funcional entre os fatores e a variável resposta.

Já os GAMs representam um método para ajustar os relacionamentos entre duas ou mais variáveis através de um gráfico de dispersão (*Scatterplot*) de um conjunto de dados, permitindo o ajustamento da tendência, sazonalidade e efeito de variáveis de confundimento (HASTIE; TIBSHIRANI, *apud* MYERS, MONTGOMERY e VINIG, 2002). Os GAMs são utilizados quando se espera um relacionamento complexo entre variáveis, as quais não são facilmente modeladas por modelos lineares padrão e modelos não-lineares. Também utilizam-se os GAMs quando não há nenhuma razão *a priori* para usar um modelo particular e se deseja sugerir uma forma funcional adequada para ajustar os dados.

REFERÊNCIAS

BARBETTA, P. A. **Construção de Modelos para Médias e Variâncias na Otimização Experimental de Produtos e Processos**. Tese de Doutorado, Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis, 1998.

BERGMAN, B.; HYNEN, A. Dispersion Effects from Unreplicated Designs in the 2^{k-p} Series. **Technometrics**, v. 39, Issue 2, p.191-199, 1997.

BOX, G. E. P.; COX, D.R. An Analysis os Transformations. **Journal of the Royal Statistical Society**. Series B, v. 26, p. 211-252, 1964.

BOX, G. E. P.; HUNTER, J.S. The 2^{k-p} Fractional Factorial Designs Part I. **Technometrics**, v. 42, Issue 1, p.28-48, 2000.

BOX, G. E. P.; HUNTER, W. G.; HUNTER. **J. S. Statistics for Experiments – An Introduction to Design, Data Analysis and Model Building**. John Wiley & Sons, New York, 1978.

BOX, G. E. P.; MEYER, R. D. Dispersion Effects from Fractional Designs. **Technometrics,** v. 28, Issue 1, p.19-27, 1986.

CATEN, C. S. **Método de Otimização de Produtos e Processos Medidos por Múltiplas Características de Qualidade.** Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia de Produção, UFRGS, Porto Alegre, 1995.

CARROLL, R. J.; RUPPERT, D. Discussion of "Signal-to-noise rations, performance criteria and transformation, by G. Box. **Technometrics**, v. 30, p.30-31, 1988.

COOK, R. D. e WEISBERG, S. **Applied Regression Including Computing and Graphics**, John Wiley&Sons, New York, NY,1999.

CORDEIRO, G. M. **Modelos Lineares Generalizados**. Anais do VII Simpósio Nacional de Probabilidade e Estatística (SINAPE), Campinas, SP, 1986.

CORDEIRO, G. M.; NETO, E. A. L. **Modelos Paramétricos**. Anais do XVI Simpósio Nacional de Probabilidade e Estatística (SINAPE), Caxambu, MG, 2004.

COSTA, S. C. **Modelos Lineares Generalizados Mistos para Dados Longitudinais**.. Tese de Doutorado, Escola Superior de Agricultura "Luiz de Queiroz", USP, Piracicaba, 2003.

DANIEL, C. Use of half-Normal plots for interpreting factorial two-level experiments. **Technometrics**, v. 1, p. 311-341, 1959.

DAVIES, O. L.; HAY, W. A. The construction and uses of fractional factorial designs in industrial research. **Biometrics**, v. 6, Issue 3, p. 233-351, 1950.

DAVIDIAN, M.; CARROLL, R. J. Variance function estimation. **Journal of the American Statistics Association,** v. 82, Issue 400, p. 1079-1091, 1987.

DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados em Experimentação Agronômica**. Anais da 46° Reunião Anual da RBRAS e 9° SEAGRO, ESALQ/USP, Piracicaba, 2001.

ENGEL, J.; HUELE, A. F. A Generalized Linear Modelling Approach to Robust Design. **Technometrics,** v.38, p. 365-373, 1996.

GUNTER, B. A. A Perspective on Taguchi Methods. **Quality Process**, v. 20, Issue 6, p.44-52, 1987.

LEE, Y; J.A. NELDER. Generalized linear models for the analysis of quality-improvement experiments. **The Canadian Journal of Statistics**, v. 26, Issue 1, p. 95-105, 1998.

LIAO, C. T. Identification of dispersion effects from unreplicated 2^{n-k} fractional factorial designs. **Computational Statistics & Data Analysis**, v. 33, p. 291-298, 2000.

MCGRATH, R. N.; LIN, D. K. Confounding of Location and Dispersion Effects in Unreplicated Fractional Factorials. **Journal of Quality Technology**, v. 33, Issue 2, p.129-139, 2001.

MCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. Chapman and Hall, London-New York, 1989.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 5° Ed., John Wiley & Sons, New York, 2001.

MONTGOMERY, D. C.; PECK, E. A. **Introduction to Linear Regression Analysis.** 2° Ed., John Wiley & Sons, New York, 1991.

MYERS R. H.; MONTGOMERY, D. C. A Tutorial on Generalized Linear Models. **Journal of Quality Technology**, v.29, Issue 3, p.274-291, 1997.

MYERS, R. H.; MONTGOMERY, D. C.; VINING, G. G. **Generalized Linear Models – With Applications in Engineering and the Sciences**. John Wiley & Sons, New York, 2002.

NAIR, V. N. Testing in industrial experiments with ordened categorical data (with discussions). **Technometrics**, v.28, p.283-311, 1986.

NELDER, J. A.; LEE, Y. Generalized Linear Models for the analysis of Taguchi-TyPe experiments. **Applied Stochastic Models and Data Analysis**, v.7, p. 107-120, 1991.

NELDER, J. A.; LEE, Y. Letters to the Editor – Joint Modeling of Mean and Dispersion. **Technometrics**, v.40, Issue 2, p.168-171, 1998.

NELDER, J. A.; PREGIBON, D. An extended quasi-likelihood function. **Biometrika**, v. 74, p. 221-232, 1987.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society**. Series A, v. 135, p. 370-84, 1972.

ORELIEN, J. G. Model Fitting in PROC GENMOD. **Statistical Science, Inc**. p. 264, 2000.

PAN, G. The impact of unidentified location effects on dispersion-effects identification from unreplicated factorial designs. **Technometrics,** v. 41. Issue, 4, p.313-326, 1999.

PIERCE, D. A.; SCHAFER, D. W. Residual in Generalized Linear Models. **Journal of Statistical American Association**, v. 81, n. 396, p.977-986, December, 1986.

PINTO, E. R.; LEON, A. C. M. P. Síntese da Modelagem Conjunta da média e dispersão de Nelder e Lee para aplicação à metodologia de Taguchi. **Anais do XXXV Simpósio Brasileiro de Pesquisa Operacional (SOBRAPO)**. Natal, RN, 2003.

PIZZOLATO, M. **Método de Otimização Experimental da Qualidade e Durabilidade de Produtos: um estudo de caso em produto fabricado por injeção de plástico**. Dissertação de Mestrado. Programa de Pós-Graduação em Engenharia de Produção, UFRGS, Porto Alegre, 2002.

PIZZOLATO, M.; FOGLIATTO, F.S. & CATEN, C. S. Otimização Experimental da Qualidade de um Produto. **CD Rom 3° Congresso Brasileiro de Gestão e Desenvolvimento de Produto.** Florianópolis, SC, 2001.

RIBEIRO, J. L. D; FOGLIATTO, F. S.; CATEN, C. S. Minimizing Manufacturing and Quality Costs in Multiresponse Optimization. **Quality Engineering**, v. 13, Issue 2, p.191-201, 2001.

SILVA, E. L. **Metodologia de Pesquisa e Elaboração de Dissertação**. Laboratório de Ensino a Distância, Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis, 2000.

SIMMONS, G. F. **Cálculo com Geometria Analítica Volume I**. McGraw-Hill, São Paulo, 1987.

TAGUCHI, G.; ELSAYED, E. A.; HSIANG, T. C. Traduzido por LOVERRI, R. C. **Engenharia da Qualidade em Sistemas de Produção**. McGraw-Hill, São Paulo, 1990.

VIEIRA, A. F. C. **Análise da Média e da Variância em Experimentos Fatoriais não replicados para a Otimização de Processos Industriais.** Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia de Produção, PUC, Rio de Janeiro, 2004.

WANG, P. C. Tests for dispersion effects from orthogonal arrays. **Computational Statistics & Data Analysis**, v. 8, p. 109-117, 1989.

WANG, P. C.; LIN, D. F. Dispersion effects in signal-response data from fractional factorial experiments. **Computational Statistics & Data Analysis**, v. 38, p. 95-111, 2001.

WEDDERBURN, R. W. M. Quasi-Likelihood Functions, Generalized Linear Models and the Gauss Newton Method. **Biometrika**, 61, pp. 439-447, 1974.

WOLFINGER, R. D.; TOBIAS, R. D. Joint Estimation of Location, Dispersion, and Random Effects in Robust Design. **Technometrics**, v. 40, Issue 1, p.62-71, 1998.

YIN, R. K. **Estudo de caso – Planejamento e Métodos**. 2° Ed., Bookman, Porto Alegre, 2001.

APÊNDICE A

Nesse Apêndice são apresentados os modelos intermediários obtidos no processo iterativo do estudo de caso.

Modelo inicial para a media

Variável resposta:	Impacto A
Variável de peso:	-
Distribuição dos dados:	Normal Inversa
Função de Ligação:	Identidade

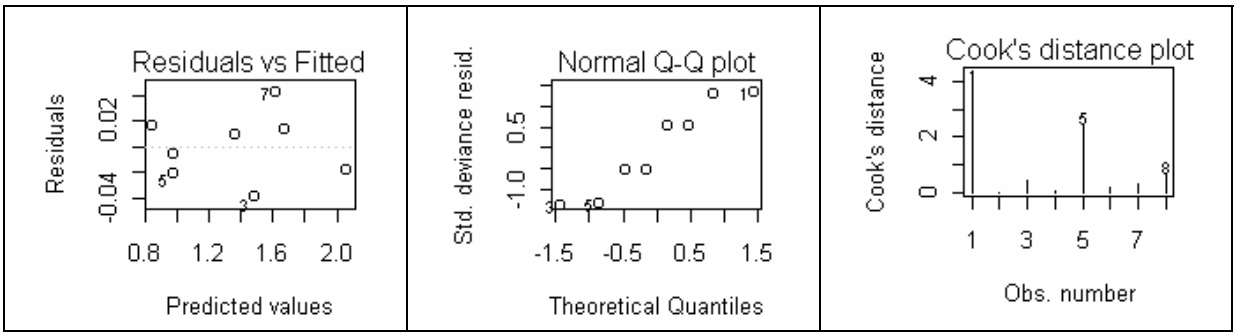
Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	1,38206	0,04827	28,634	9,35e-05 **
A	-0,27187	0,04190	-6,488	0,00743 **
BC	0,20726	0,04250	4,876	0,01649 *
B	0,17655	0,04248	4,156	0,02533 *
C	0,13666	0,04245	3,219	0,04861 *

Parâmetro de dispersão para a distribuição de probabilidade dos dados	0,005720736
AIC	-7,8848

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo		0,20456	7	0,52189		
A	1	0,17560	6	0,31733	35,758	0,009361 **
BC	1	0,05832	5	0,14173	30,695	0,011592 *
B	1	0,06642	4	0,08341	10,194	0,049605 *
C	1	0,20456	3	0,01699	11,611	0,042230 *

N°da observação	Resíduos do Modelo	Valores preditos (ajustados)
1	-0,003812696	0,8630515
2	-0,053328132	2,1743967
3	0,007883781	1,4067838
4	0,032611207	1,6306645
5	-0,044960772	1,0042391
6	-0,06101496	1,4865505
7	0,080044349	1,5479714
8	0,029272773	0,9428182

Análise Gráfica



1°iteração: 1°modelo para a variância

Variável resposta:	Resíduos do modelo inicial da média
Variável de peso:	-
Distribuição dos dados:	Gama
Função de Ligação:	Logarítmica

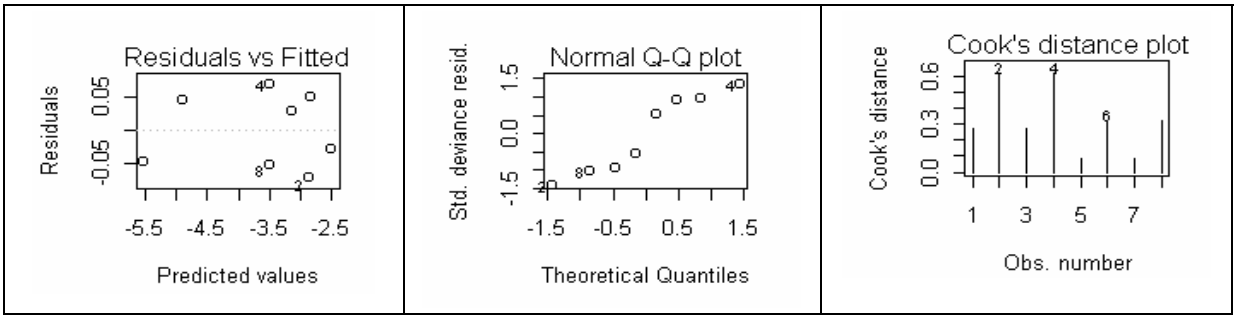
Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	-3,58887	0,02959	-121,30	1,24e-06 ***
A	-0,31619	0,02959	-10,69	0,001751 **
BC	0,59492	0,02959	20,11	0,000269 ***
B	0,42023	0,02959	14,20	0,000756 ***
C	-0,60100	0,02959	-20,31	0,000261 ***

Parâmetro de dispersão para a distribuição de probabilidade dos dados (ϕ)	0,007002596
AIC	-70,261

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	55,835		
A	1	0,7168	6	48,667	102,37	0,0020566 **
BC	1	11,153	5	37,514	159,27	0,0010728 **
B	1	0,9998	4	27,515	142,78	0,0012607 **
C	1	27,305	3	**0,0210**	389,92	0,0002838 ***

N°da observação	Resíduos do Modelo	Valores preditos ($1/\mu_i$)
1	-0,04773734	-5,521219
2	-0,07188478	-2,858537
3	0,04626456	-4,88883
4	0,06859539	-3,490926
5	0,02754305	-3,129377
6	0,05014692	-2,846367
7	-0,02805831	-2,496989
8	-0,05188197	-3,478755

Análise Gráfica



1 ° iteração: 1 ° modelo para a média

Variável resposta:	Impacto A
Variável de peso:	$1/\mu_i$ (inicial média) * ϕ (1° variância)
Distribuição dos dados:	Quase
Função de Ligação:	Identidade

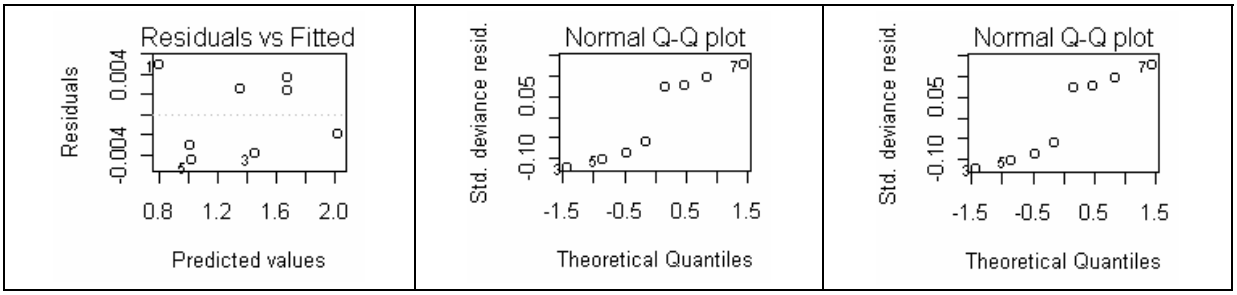
Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	1,37663	0,02609	52,756	0,000359 **
A	-0,2523	0,02579	-9,786	0,010281 *
BC	0,22173	0,02554	8,681	0,013012 *
B	0,14046	0,02609	5,383	0,032824 *
C	0,11344	0,02554	4,441	0,047140 *
AB	0,07823	0,02579	3,034	0,093637 .

Parâmetro de dispersão para a distribuição de probabilidade dos dados	4,847281e-05
AIC	-

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	0,0114788		
A	1	0,0039093	6	0,0075695	80,6497	0,01217 *
BC	1	0,0045692	5	0,0030003	94,2629	0,01044 *
B	1	0,0014672	4	0,0015331	30,2691	0,03149 *
C	1	0,0009900	3	0,0005431	20,4241	0,04564 *
AB	1	0,0004462	2	0,0000969	9,2041	0,09364 .

N°da observação	Resíduos do Modelo	Valores preditos ($1/\mu_i$)
1	0,004874118	0,7973028
2	-0,00202207	2,0263869
3	-0,003817691	1,4584643
4	0,002334981	1,678149
5	-0,004518516	1,0138825
6	0,002445546	1,3560306
7	0,00363943	1,675044
8	-0,0030708	1,0077927

Análise Gráfica



2° iteração: 2° modelo para a variância

Variável resposta:	Resíduos 1°modelo média
Variável de peso:	-
Distribuição dos dados:	Gama
Função de Ligação:	Logarítmica

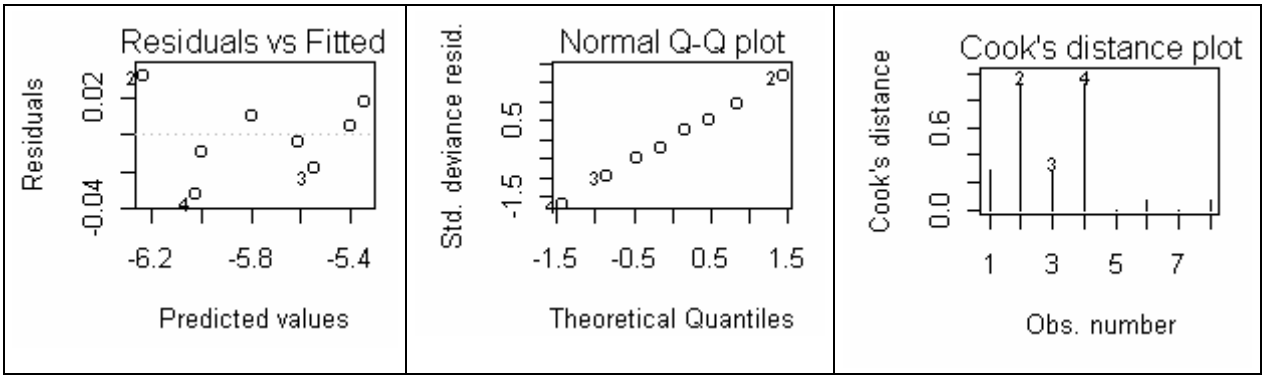
Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	-5,74608	0,01107	-519,078	1,58e-08 ***
A	0,10403	0,01107	9,397	0,002553 **
BC	-0,07338	0,01107	-6,629	0,006993 **
B	-0,26931	0,01107	-24,328	0,000152 ***
C	-0,04241	0,01107	-3,831	0,031340 *

Parâmetro de dispersão para a distribuição de probabilidade dos dados (ϕ)	0,000980322
AIC	-120,5

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	0,71356		
A	1	0,09305	6	0,62051	94,918	0,0022973 **
BC	1	0,03092	5	0,58959	31,543	0,0111596 *
B	1	0,57227	4	0,01732	583,755	0,0001554 ***
C	1	0,01438	3	0,00294	14,671	0,0313546 *

N°da observação	Resíduos do Modelo	Valores preditos ($1/\mu_i$)
1	-5,341771	0,018008857
2	-6,235205	0,031738943
3	-5,549826	-0,018227825
4	-6,02715	-0,032424912
5	-5,403714	0,004145397
6	-6,003631	-0,009839456
7	-5,611769	-0,004156912
8	-5,795576	0,009775269

Análise Gráfica



2° iteração: 2° modelo para a media

Variável resposta:	Impacto A
Variável de peso:	$1/\mu_i$ (1° média) * ϕ (2° variância)
Distribuição dos dados:	Quase
Função de Ligação:	Identidade

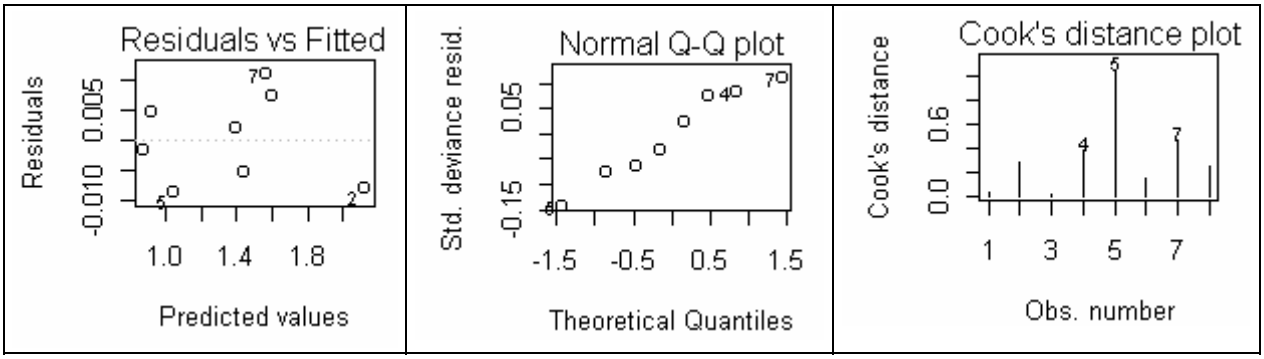
Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	1,3722	0,04624	29,674	8,4e-05 ***
A	-0,26002	0,04546	-5,719	0,0106 *
BC	0,21359	0,04540	4,705	0,0182 *
B	0,15178	0,04551	3,335	0,0446 *
C	0,12730	0,04538	2,805	0,0676 .

Parâmetro de dispersão para a distribuição de probabilidade dos dados (ϕ)	0,0001222322
AIC	-

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	0,0089688	36,5424	0,00908 **
A	1	0,0044667	6	0,0045021	17,5823	0,02474 *
BC	1	0,0021491	5	0,0023530	8,3834	0,06274 .
B	1	0,0010247	4	0,0013283	7,8669	0,06758 .
C	1	0,0009616	3	0,0003667	36,5424	0,00908 **

N°da observação	Resíduos do Modelo	Valores preditos ($1/\mu_i$)
1	-0,001566493	0,8741272
2	-0,007990878	2,1248871
3	0,002118677	1,3941579
4	0,007271911	1,6048564
5	-0,008526468	1,0467119
6	-0,00536683	1,4431202
7	0,010959442	1,5667426
8	0,004626673	0,9230895

Análise Gráfica



3°iteração: 3°modelo para a variância

Variável resposta:	Resíduos 2°modelo média
Variável de peso:	-
Distribuição dos dados:	Gama
Função de Ligação:	Logarítmica

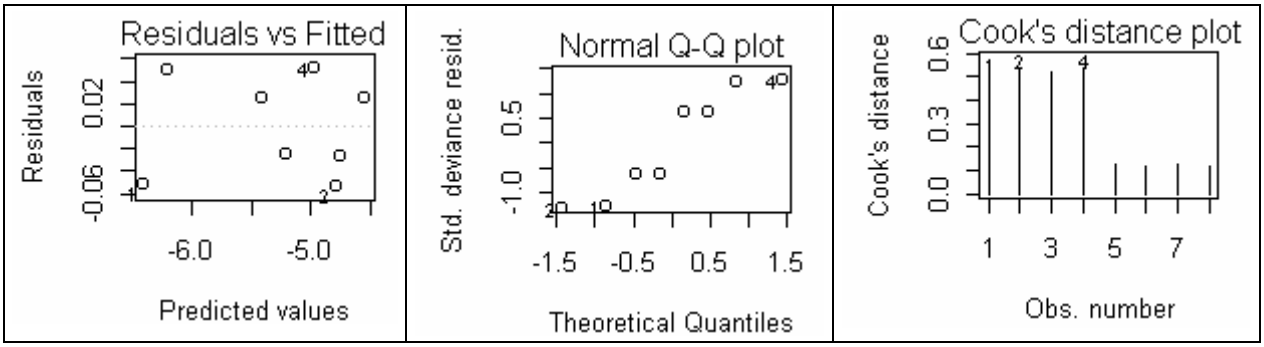
Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	-5,28049	0,02363	-223,511	1,97e-07 ***
A	-0,09946	0,02363	-4,210	0,024480 *
BC	0,52351	0,02363	22,159	0,000201 ***
B	0,19218	0,02363	8,134	0,003885 **
C	-0,31043	0,02363	-13,140	0,000952 ***

Parâmetro de dispersão para a distribuição de probabilidade dos dados (ϕ)	0,004465202
AIC	-100,92

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	2,78717		
A	1	0,06766	6	2,71951	15,153	0,0300667 *
BC	1	1,68007	5	1,03944	376,258	0,0002993 ***
B	1	0,26718	4	0,77227	59,835	0,0044927 **
C	1	0,75885	3	0,01341	169,949	0,0009747 ***

N°da observação	Resíduos do Modelo	Valores preditos ($1/\mu_i$)
1	-0,05238352	-6,406071
2	-0,05320892	-4,77577
3	0,05061526	-6,207155
4	0,05138556	-4,974685
5	-0,02627902	-4,738185
6	-0,02546846	-5,201941
7	0,02582648	-4,53927
8	0,02504324	-5,400856

Análise Gráfica



3° iteração: 3° modelo para a media

Variável resposta:	Impacto A
Variável de peso:	$1/\mu_i$ (2° média) * ϕ (3° variância)
Distribuição dos dados:	Quase
Função de Ligação:	Identidade

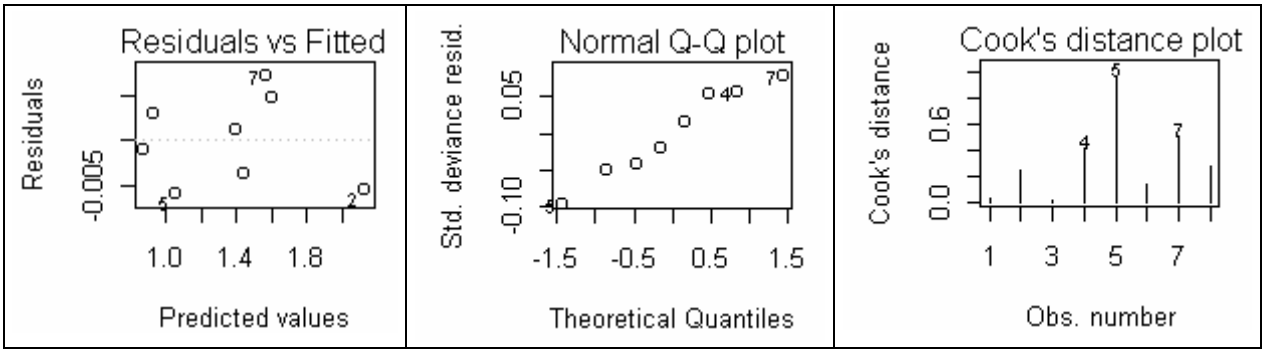
Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	1,37625	0,04654	29,573	8,5e-05 ***
A	-0,26094	0,04566	-5,715	0,0106 *
BC	0,21479	0,04568	4,702	0,0182 *
B	0,15228	0,04567	3,335	0,0446 *
C	0,12645	0,04566	2,769	0,0696 .

Parâmetro de dispersão para a distribuição de probabilidade dos dados (ϕ)	5,637907e-05
AIC	-

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	0,0040304		
A	1	0,0019645	6	0,0020658	34,8449	0,009708 **
BC	1	0,0009862	5	0,0010796	17,4923	0,024908 *
B	1	0,0004782	4	0,0006014	8,4816	0,061880 .
C	1	0,0004323	3	0,0001691	7,6679	0,069621 .

N°da observação	Resíduos do Modelo	Valores preditos ($1/\mu_i$)
1	0,8746898	-0,001049903
2	2,1307117	-0,005533526
3	1,396571	0,001325922
4	1,6088305	0,004808971
5	1,0513636	-0,005967332
6	1,4482349	-0,003795559
7	1,5732448	0,007300716
8	0,9263537	0,003035613

Análise Gráfica



4° iteração: 4° modelo para a variância – MODELO FINAL PARA A VARIÂNCIA

Variável resposta:	Resíduos 3°modelo média
Variável de peso:	-
Distribuição dos dados:	Gama
Função de Ligação:	Logarítmica

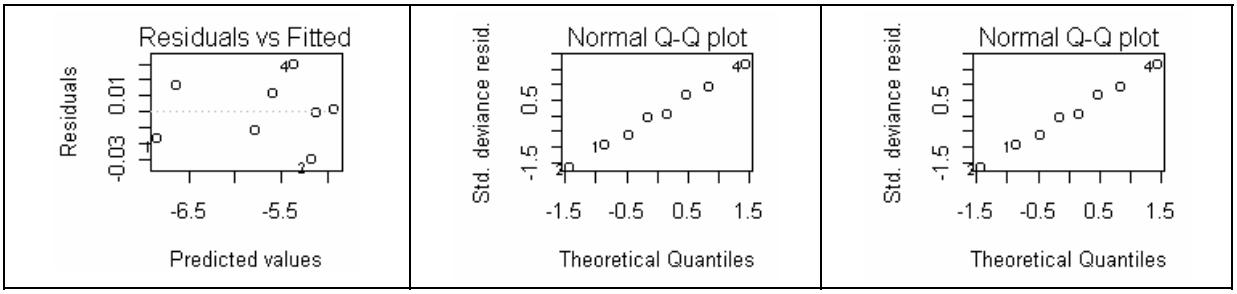
Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	-5,67876	0,01043	-544,351	1,37e-08 ***
A	-0,09986	0,01043	-9,572	0,002419 **
BC	0,53512	0,01043	51,295	1,63e-05 ***
B	0,20265	0,01043	19,426	0,000298 ***
C	-0,32567	0,01043	-31,218	7,22e-05 ***

Parâmetro de dispersão para a distribuição de probabilidade dos dados (ϕ)	0,0008706426
AIC	-120,37

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	2,92308		
A	1	0,07142	6	2,85166	82,033	0,0028428 **
BC	1	1,72054	5	1,13113	1976,168	2,506e-05 ***
B	1	0,29463	4	0,83649	338,409	0,0003505 ***
C	1	0,83388	3	0,00261	957,775	7,412e-05 ***

N°da observação	Resíduos do Modelo	Valores preditos ($1/\mu_i$)
1	-0,016942091	-6,842067
2	-0,029975976	-5,166804
3	0,016752783	-6,642353
4	0,029388751	-5,366518
5	-0,00097953	-5,120476
6	0,011805034	-5,585705
7	0,000978886	-4,920762
8	-0,011898738	-5,78542

Análise Gráfica



4° iteração: 4° modelo para a media – MODELO FINAL PARA A MÉDIA

Variável resposta:	Impacto A
Variável de peso:	$1/\mu_i$ (3° média) * ϕ (4° variância)
Distribuição dos dados:	Quase
Função de Ligação:	Identidade

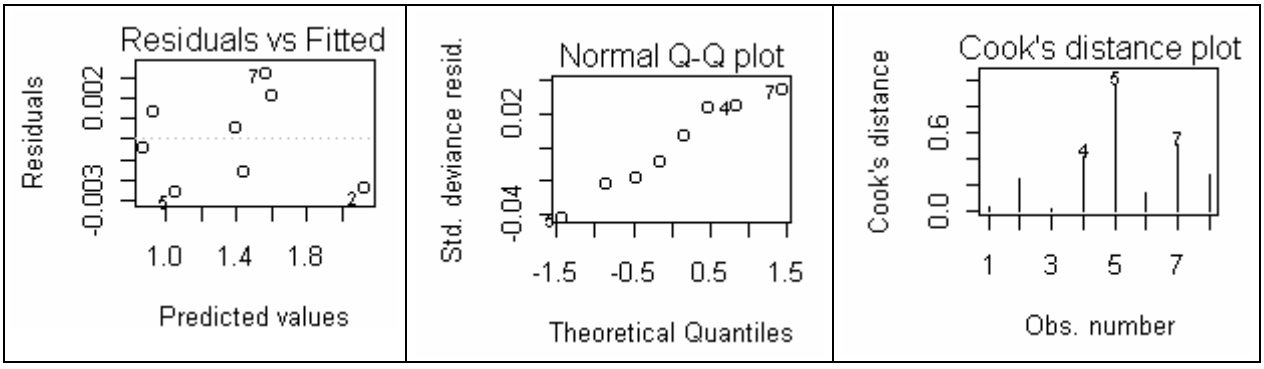
Fatores	Estimativa	Erro Padrão	Estatística t	Pr(>|t|)
(Intercepto)	1,37625	0,04652	29,582	8,48e-05 ***
A	-0,26092	0,04564	-5,717	0,0106 *
BC	0,21480	0,04566	4,704	0,0182 *
B	0,15229	0,04565	3,336	0,0445 *
C	0,12645	0,04565	2,770	0,0696 .

Parâmetro de dispersão para a distribuição de probabilidade dos dados (ϕ)	1,095364e-05
AIC	

	GL	Redução da *Deviance*	GL restante	Resíduo da *Deviance*	Estatística F	Pr(>F)
Modelo nulo			7	0,00078374		
A	1	0,00038193	6	0,00040181	34,8677	0,009699 **
BC	1	0,00019200	5	0,00020981	17,5282	0,024840 *
B	1	0,00009290	4	0,00011691	8,4816	0,061880 .
C	1	0,00008405	3	0,00003286	7,6730	0,069568 .

N°da observação	Resíduos do Modelo	Valores preditos ($1/\mu_i$)
1	-0,0004638	0,8746997
2	-0,0024398	2,1306963
3	0,000586	1,3965298
4	0,00212	1,6088662
5	-0,0026302	1,0514002
6	-0,0016723	1,4482039
7	0,0032175	1,5732302
8	0,0013375	0,9263738

Análise Gráfica



APÊNDICE B

Este Apêndice traz os comandos do pacote computacional “R” utilizados para obtenção dos modelos apresentados nesta dissertação.

O pacote “R” apresenta uma linguagem e ambiente para computação estatística e gráficas. Além disso, fornece uma ampla variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento ...) e gráficos de qualidade e é altamente extensível. Esse pacote é disponibilizado como *Software* Livre sob os termos da Licença Pública Geral GNU da *Free Software Foundation* na forma de código fonte. Assim, pode ser livremente copiado e distribuído entre usuários, bem como possui código livre, tornando-se, um software flexível e de fácil utilização. O pacote computacional “R” pode ser adquirido através do *site*: http://www.r-project.org/.

Segue abaixo os comandos e saídas (*output*) do primeiro modelo gerado durante o procedimento iterativo (seção 4.3.1). Os comandos do “R” estão antecedidos pelo símbolo “>” e destacados em negrito. Os resultados são mostrados logo abaixo dos mesmos. Eventuais comentários sobre a programação estão descritos em caixas de texto.

> read.table("H:/R2/Bloco2.dat",head=T,sep="")->b2

O "R" lê banco de dados em formato *.dat*. O nome das variáveis não podem estar separados por espaços. É necessário dar um nome ao banco de dados a ser utilizado. Neste caso o nome atribuído foi "b2".

> b2

Ao digitar o nome do banco de dados, o mesmo é listado na tela.

	n	A	B	C	D	ABCD	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ImpactoA
1	2	1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	-1	0.86
2	3	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	-1	1	2.01
3	5	-1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	1	1.42
4	8	1	1	1	-1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1.70
5	10	1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	0.96
6	12	-1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1	1.38
7	14	-1	-1	-1	1	-1	1	1	-1	1	-1	-1	-1	1	1	1	1.71
8	17	1	1	-1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	0.97

> inicialmodel<-glm(ImpactoA~A+BC+B+C,b2,family=inverse.gaussian(link="identity"))

Para a modelagem de um GLM é necessário atribuir um nome para identificar o modelo a ser ajustado (neste caso o nome foi "inicialmodel"). A variável resposta é separada das variáveis explicativas por um "~", depois deve digitar o nome do banco de dados, a distribuição de probabilidade dos dados e a função de ligação. Os GLMs possíveis de serem ajustados no "R"são apresentados na Figura 10 e na Figura 11. O Apêndice C apresenta a notação utilizada pelo pacote computacional.

> summary(saturado)

O comando "summary" exibe o modelo que foi ajustado.

```
Call:
glm(formula = ImpactoA ~ A + BC + B + C, family = inverse.gaussian(link = "identity"),
    data = bloco2)

Deviance Residuals:
       1        2        3        4        5        6        7
8
-0.003813  -0.053328   0.007884   0.032611  -0.044961  -0.061015   0.080044   0.029273


Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  1.38206    0.04827     28.634 9          35e-05 ***
A            -0.27187   0.04190     -6.488            0.00743 **
BC           0.20726    0.04250     4.876             0.01649 *
B            0.17655    0.04248     4.156             0.02533 *
C            0.13666    0.04245     3.219             0.04861 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for inverse.gaussian family taken to be 0.005720736)
Null deviance: 0.521894 on 7 degrees of freedom
Residual deviance: 0.016992 on 3 degrees of freedom
AIC: -7.8848

Number of Fisher Scoring iterations: 5

> anova(inicialmed,test="F")

Através do comando "anova" é possível realizar a Análise da *Deviance* (ANODEV).

Analysis of Deviance Table
Model: inverse.gaussian, link: identity
Response: ImpactoA
Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			7	0.52189		
A	1	0.20456	6	0.31733	35.758	0.009361 **
BC	1	0.17560	5	0.14173	30.695	0.011592 *
B	1	0.05832	4	0.08341	10.194	0.049605 *
C	1	0.06642	3	0.01699	11.611	0.042230 *

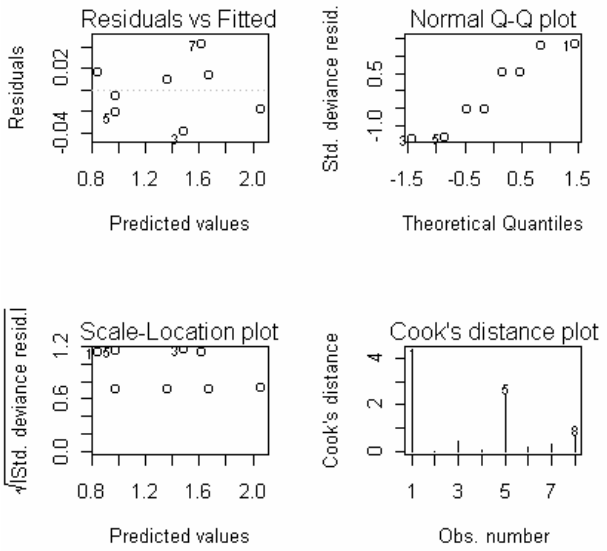
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> par(mfrow=c(2,2))

Para que os gráficos a serem gerados futuramente apareçam na forma de uma matriz 2x2.

> plot(inicialmod)

Por "default" o pacote gera os gráficos apresentados abaixo.



APÊNDICE C

Neste Apêndice são listada as notações de possíveis GLM que podem ser ajustados no pacote computacional “R”.

Nome da função de probabilidade	Notação "R"
Binomial	binomial
Normal	gaussian
Gama	Gamma
Normal Inversa	inverse.gaussian
Poisson	poisson
Função de *quase*-verossimilhança	quase
Função de *quase*-verossimilhança Binomial	quasebinomial
Função de *quase*-verossimilhança Poisson	quasepoisson

Figura 25: Notações das funções de probabilidade do "R"

Nome da função de ligação	Notação "R"
Identidade	indetity
Logaritmo	log
Inversa	inverse
Logit	logit
Probit	probit
Cauchit	cauchit
Complementar log-log	clog-log
Raiz quadrada	sqrt
Inverso da média ao quadrado $\left(1/\mu^2\right)$	1/mu^2

Figura 26: Notações das funções de ligação do "R"

Nome da função de variância	Notação "R"
constant	constant
$\mu(1-\mu)$	mu(1-mu)
μ	mu
μ^2	mu^2
μ^3	mu^3

Figura 27: Notações das funções de variância para as funções *quase*-verossimilhança