UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Centro de Ciências Matemáticas e da Natureza
Instituto de Matemática
Departamento de Métodos Estatísticos

Diogo da Hora Elias

APLICAÇÕES DE PLANEJAMENTO DE EXPERIMENTO PARA DADOS NÃO NORMAIS

Rio de Janeiro
2014

DIOGO DA HORA ELIAS

APLICAÇÕES DE PLANEJAMENTO DE EXPERIMENTO PARA DADOS NÃO NORMAIS

Projeto Final de Curso como parte dos requisitos para
a obtenção do título de ESTATÍSTICO

Flávia Landim – UFRJ
Orientador

Mariane Alves – UFRJ
Banca Avaliadora

Marina Paez – UFRJ
Banca Avaliadora

Rio de Janeiro
2014

Agradecimentos

Agradeço a toda minha família, principalmente aos meus pais, Helio e Maria de Lourdes por todo carinho, paciência e esforço para que eu chegasse até aqui. A Professora Flávia Landim pela atenção, conhecimento e todo o tempo disponibilizado para me auxiliar na construção deste projeto.

Resumo

Neste trabalho, após realizarmos uma revisão dos conteúdos de Planejamentos de Experimentos e Modelos Lineares Generalizados (MLG's), utilizamos três exemplos práticos que são simultaneamente constituídos por um plano fatorial 2^k e uma variável resposta que segue uma distribuição da família exponencial. Em cada exemplo, além da metodologia de MLG's, foram abordadas outras duas alternativas que são amplamente utilizadas, primeiro, o modelo linear normal, segundo, alguma transformação da variável resposta que, supostamente, conduziria a variável resposta à normalidade. Dessa forma, podemos comparar os três métodos e comprovar a eficiência dos Modelos Lineares Generalizados.

No primeiro exemplo, trabalhamos com uma situação hipotética, os dados foram obtidos através de uma simulação que forneceu um plano fatorial 2^4 com uma única replicação e variável resposta com distribuição binomial. O objetivo da simulação foi analisar o uso da regressão logística, que é um dos casos mais populares da aplicação de MLG's, empregada principalmente quando a resposta é uma proporção. Neste exemplo, além do modelo linear normal, utilizamos a transformação de Box-Cox, em ambos os casos foram observadas algumas incoerências discutidas com detalhes no capítulo 4.

O segundo exemplo é um estudo dos fatores sobre a distância alcançada pela bola lançada pela catapulta, baseado em um exercício do livro de Myers et al (2010), também foi realizada uma simulação. Temos um experimento fatorial fracionado 2^{4-1} com três replicações, já que em cada combinação de tratamento a bola foi atirada três vezes, a escolha um plano fracionário foi intencional, pois o fracionamento é uma estratégia muito comum quando não é possível obter realizações para todas as combinações de tratamento.

Nesse caso, a distribuição da variável resposta é gama, por ser contínua e não ser uma proporção, para muitos, seria natural considerar uma variável que segue uma distribuição normal. Inicialmente, utilizamos o modelo linear normal e a transformação logarítmica, já que a resposta é sempre positiva, e finalmente, um modelo de regressão gama com função de ligação logarítmica. O objetivo era avaliar quais fatores que afetam a distância alcançada pela bola e construir um modelo de regressão adequado.

O terceiro exemplo é um estudo sobre a sobrevivência de espermatozóide em um banco de esperma também foi retirado de Myers et al (2010). Nesse estudo, os espermatozóides são armazenados em citrato de sódio e glicerol, a quantidades dessas substâncias variaram juntamente com o tempo de equilíbrio. Dessa forma, temos um plano fatorial 2^3 com apenas uma replicação e a variável resposta é a proporção de espermatozóides sobreviventes. O propósito é encontrar quais fatores (substâncias e o tempo de equilíbrio) afetam a resposta, a partir daí, construir o modelo de regressão. Considerando que a resposta segue distribuição binomial, utilizamos novamente o modelo logístico. Como alternativas usamos o modelo linear normal e a transformação arco seno da raiz quadrada, que estabiliza a variância quando os dados são proporções.

Sumário

CAPÍTULO 1 - Introdução

O primeiro objetivo desse trabalho foi realizar um resumo de tópicos de Planejamento de Experimentos, especialmente os planos fatoriais 2^k, e de Modelos Lineares Generalizados (MLG's). Isto foi feito através de uma revisão bibliográfica. O segundo objetivo foi, através da teoria apresentada nos capítulos iniciais, mostrar aplicações dos MLG's em Planejamentos de Experimentos.

Como o conteúdo da disciplina de graduação "Análise de Regressão" é grande e MLG's é seu último tópico, muitas vezes o tempo que sobra para a apresentação do mesmo deixa a desejar. Além disso, geralmente, em muitas situações envolvendo Planejamentos de Experimentos, o pesquisador, automaticamente, supõe que a variável resposta possui distribuição normal, o que pode levar a conclusões erradas. Pretendemos mostrar que os Modelos Lineares Generalizados podem representar uma boa alternativa nesses casos.

Segundo Montgomery (2007), o desenvolvimento do planejamento de experimentos pode ser dividido em quatro fases. A primeira fase foi liderada pelo trabalho pioneiro de Ronald A. Fisher, entre as décadas de 1920 e 1930, quando este foi o responsável pelas estatísticas e análise de dados em um instituto de pesquisa sobre agricultura. A convivência com cientistas e pesquisadores de diferentes áreas possibilitou o conhecimento necessário para que Fisher estabelecesse os três princípios básicos do planejamento de experimentos: aleatorização, replicação e blocagem. Fisher introduziu princípios e o pensamento estatístico na investigação experimental, incluindo o conceito de plano fatorial e a análise da variância.

A segunda fase foi estimulada pelo desenvolvimento da metodologia de superfície de resposta por Box e Wilson (1951). Eles reconheceram e exploraram o fato de muitos experimentos industriais serem fundamentalmente diferentes daqueles praticados na agricultura de duas maneiras: geralmente, a resposta pode ser observada imediatamente, e o pesquisador pode aprender rapidamente informações cruciais sobre um pequeno grupo de realizações que podem ser usadas para planejar o próximo experimento. No entanto, a aplicação do planejamento estatístico no processo de fabricação ainda não era amplamente difundida. Isso acontecia devido a um treinamento inadequado sobre conceitos básicos de estatística, a falta de recursos computacionais e a não existência de um software compatível para aplicações em planejamento de experimentos.

O aumento do interesse em melhoria da qualidade impulsionou a terceira fase. O trabalho de Genichi Taguchi teve significante impacto para o interesse no uso de planejamento de experimentos. Taguchi defendeu que o processo deve ser insensível em relação à variação de fatores de difícil controle e encontrar níveis das variáveis do processo que levam a média a ter um determinado valor desejado, ao mesmo tempo em que a variabilidade em torno deste valor é reduzida. Quando esses objetivos são alcançados, podemos dizer que o processo é robusto.

A quarta fase apresenta um interesse renovado em planejamento estatístico por pesquisadores e praticantes e o desenvolvimento de novas abordagens para problemas experimentais no campo industrial incluindo alternativas aos métodos de Taguchi que permitem que conceitos de engenharia sejam utilizados de forma mais eficiente.

Segundo Turkman e Silva (2000), os Modelos Lineares Generalizados, introduzidos na década de 1970, tiveram grande impacto no desenvolvimento da estatística aplicada. No início, seu uso esteve confinado a um grupo restrito de pesquisadores, devido a falta de bibliografia acessível e à complexidade inicial do GLIM, primeiro software dirigido

para aplicação desta metodologia. Foram necessários 20 anos para que os MLG's chegassem ao domínio público. Isso ocorreu devido às melhorias no software existente. Hoje, a maioria dos pacotes estatísticos contém módulos apropriados ao estudo destes modelos. Pode-se dizer que o conhecimento da metodologia dos MLG's é imprescindível para qualquer indivíduo que utilize métodos estatísticos.

A importância dos Modelos Lineares Generalizados não é apenas de caráter prático. Do ponto de vista teórico, a sua importância vem, essencialmente, do fato desta metodologia constituir uma abordagem unificada de muitos procedimentos estatísticos usados habitualmente e promover o papel central da verossimilhança na teoria de inferência.

Para alcançar os objetivos desse trabalho, essa monografia foi estruturada em cinco capítulos. O capítulo 1 contém uma introdução e os objetivos do trabalho. No capítulo 2, apresentam-se uma introdução ao planejamento de experimentos com foco para os modelos fatoriais 2^k. No capítulo 3, faz-se uma breve revisão dos modelos lineares generalizados, MLG's. No capítulo 4, são feitas aplicações dos modelos lineares generalizados para o planejamento de experimentos. Finalmente, no capítulo 5, conclusões e considerações finais do trabalho são expostas.

CAPÍTULO 2 - Planejamentos de Experimentos: uma introdução

A discussão a seguir está baseada no livro do Montgomery (2007).

Segundo Montgomery (2007), experimentos são testes ou uma série de testes, nos quais mudanças intencionais são feitas na variável de entrada de um processo ou sistema para que possamos observar e identificar as razões para mudanças que podem ser observadas na variável resposta. Seu objetivo é fazer com que o processo seja minimamente afetado por fontes externas de variabilidade. As principais áreas de utilização de experimentos estão na ciência e engenharia, porém existem aplicações em áreas como: marketing e negócios.

Podemos citar o seguinte exemplo, um engenheiro metalúrgico está interessado em estudar o efeito de dois processos de enrijecimento de ligas de alumínio: solução de óleo ou solução de água salgada. A média de rigidez das ligas em cada processo será determinada para saber qual é o melhor. No entanto, devemos ficar atentos para algumas questões: Outros processos poderiam ser testados? Existem outros fatores que poderiam afetar o enrijecimento? Quantas ligas deveriam ser testadas em cada processo? Como devem ser determinadas as soluções e, em qual ordem, os dados devem ser coletados? Qual método de análise de dados deve ser usado? Qual diferença observada na rigidez entre as duas soluções deve ser considerada importante? Estas questões devem ser respondidas antes do experimento ser executado.

É importante que o experimento seja bem planejado porque os resultados e as conclusões a serem tomadas dependem do modo em que os dados são coletados.

Suponha que no exemplo anterior, o engenheiro não tenha conhecimento suficiente para saber o quanto da diferença entre as médias está relacionado ao processo de enrijecimento e o quanto provém do aquecimento (feito anteriormente). Assim, o método de coleta de dados afetou as conclusões que podem ser tomadas do experimento.

Podemos visualizar um processo como uma combinação de operações, máquinas, métodos, pessoas e recursos que transformam o insumo em produto que tem uma ou mais variáveis respostas. Algumas das variáveis do processo $x_1, x_2, ..., x_p$ são controláveis enquanto outras variáveis não controláveis $z_1, z_2, ..., z_q$.

Experimentos envolvem vários fatores e, muitas vezes, o objetivo do experimentador é determinar a influência que esses fatores exercem na variável resposta do sistema. O experimentador pode usar várias estratégias.

No livro do Montgomery, há o seguinte exemplo do jogo de golfe. Suponha que um praticante deseja melhorar seu desempenho, mas não tem muito tempo para treinar. Primeiramente, foram estabelecidos oito fatores que podem influenciar seu rendimento, mas percebe que quatro podem ser ignorados, pois não possuem nenhum efeito prático.

Os efeitos considerados seriam tipo de bola, tipo de taco, tipo de bebida e tipo de locomoção. Na primeira rodada, percebe-se que as tacadas estão irregulares, portanto, na rodada seguinte muda-se o tipo de taco, mantendo outros fatores no mesmo nível.

O processo continua indefinidamente, sempre mudando o nível dos fatores de acordo com o resultado observado. Essa estratégia é chamada de melhor palpite. No entanto, tem duas desvantagens: não há garantias de sucesso e o primeiro palpite é sempre considerado inaceitável.

Outra estratégia a ser usada seria “um fator a cada vez”, estabelece-se um ponto de partida com cada fator em determinado nível, então se varia cada fator mantendo os

outros fatores constantes. A desvantagem é que a interação entre fatores é desconsiderada.

O método mais adequado para tratar vários fatores é um experimento fatorial. Nesta abordagem, os fatores variam juntos, ao invés de um de cada vez. Se temos *k* fatores, cada um com dois níveis, experimento será 2^k fatorial

Para experimentos com quatro, cinco fatores ou mais não é necessário usar todas as combinações possíveis. Para casos como estes, a técnica a ser utilizada será o experimento fatorial fracionário, que é uma variação do experimento fatorial básico.

Existem dois aspectos em um problema experimental: o planejamento do experimento e a análise estatística dos dados. Esses dois assuntos estão intimamente relacionados porque o método de análise depende diretamente do planejamento empregado.

Os três princípios básicos do planejamento de experimentos são: aleatorização, replicação e blocagem. Através da aleatorização é definida a alocação do material do experimento e ordem na qual as observações do experimento são executadas são determinadas aleatoriamente. Métodos estatísticos exigem que as observações (ou erros) sejam variáveis aleatórias independentemente distribuídas.

Programas de computadores são amplamente usados para ajudar a selecionar e construir o planejamento do experimento. Esses programas, geralmente, apresentam as realizações em ordem aleatória. Essa ordem é criada pelo uso de um gerador de números aleatórios.

A replicação é uma repetição independente de cada combinação de fatores. No exemplo do engenheiro metalúrgico, se cinco ligas de alumínio são tratadas em cada processo de resfriamento, temos cinco replicações. Replicação tem duas importantes propriedades: permite obtenção de uma estimativa do erro experimental, essa estimativa é utilizada para avaliar se diferença observada entre os dados é estatisticamente significante. Se a média amostral é utilizada para estimar a verdadeira média da resposta, a replicação permite obter uma estimativa mais precisa do parâmetro. Por exemplo, considerando como a variância de uma observação individual igual a σ^2 e *n* replicações, a variância da média amostral será igual a $\frac{\sigma^2}{n}$.

Blocagem é uma técnica de planejamento para melhorar a precisão de como é feita a comparação entre os fatores de interesse. Geralmente, a blocagem é usada para diminuir ou eliminar a variabilidade transmitida por fatores que podem influenciar a resposta, mas que não são de nosso interesse (fatores de ruído). Blocagem pode ser definida como um conjunto de condições relativamente homogêneas do experimento.

Podemos destacar um conjunto de estratégias para realizar um experimento. Pode parecer óbvio, mas, muitas vezes, não é simples perceber que o problema precisa ser tratado através da experimentação. É importante solicitar informações de todos os profissionais envolvidos no processo, além disso, devemos fazer uma lista de questões sobre o experimento. Estabelecer exatamente qual o problema contribui para um melhor entendimento do fenômeno a ser estudado e para sua solução.

Devemos estar certos que a variável resposta selecionada fornece informações úteis sobre o processo a ser estudado. Geralmente, a média e o desvio padrão serão a variável resposta. A capacidade de medição é fator importante, porque, caso este seja inadequado, somente os efeitos dos fatores com relativa grandeza serão detectados pelo experimento, ou replicações adicionais podem ser necessárias. Em alguns casos de inconsistência da capacidade de medição, podemos medir a unidade experimental várias vezes e usar a média como resposta observada.

Considerando os fatores que podem influenciar um sistema, podemos classificá-los como: fatores potenciais de planejamento e fatores de ruído. Os fatores potenciais de planejamento são aqueles que podem variar no experimento.

Fatores de ruído podem ter um grande efeito, mas podem não ser interessantes no contexto do experimento. Classificamos fatores desta espécie como: controláveis e incontroláveis. Os fatores de ruído controláveis são aqueles que podem ser identificados.

Na escolha do planejamento, levamos em consideração o número de replicações, seleção de uma ordem adequada para as realizações e determinar se a blocagem e restrições na aleatorização estão envolvidas.

Há vários programas estatísticos que ajudam nessa fase do processo. Informamos o número de fatores, níveis e faixa de variação e o programa apresentará uma seleção de planejamentos.

Quando realizamos um experimento, é vital monitorar o processo cuidadosamente para assegurar que tudo está sendo feito de acordo com o planejamento. Neste momento, erros destroem a validade do experimento. Coleman e Montgomery (1993) sugerem que antes de realizar o experimento, é aconselhável realizar testes ou experimentos-pilotos que informarão a consistência do material e do sistema de medição.

Os resultados da análise estatística devem fornecer resultados mais objetivos que o julgamento natural. Se o experimento foi bem planejado, os métodos estatísticos não precisam ser muito elaborados. Também é importante apresentar os resultados de acordo com o modelo empírico. Análise residual e checagem da adequação do modelo são técnicas que podem ser usadas.

Uma vez que os dados foram analisados, podemos tomar conclusões práticas sobre os resultados e recomendar que ações serão realizadas. Métodos gráficos também são utilizados neste momento. Testes de confirmação devem ser usados para validar suas conclusões.

Segundo Montgomery (2007), experimentação é uma importante parte do processo, inicialmente, formulamos hipóteses que servirão de base para execução do processo e para avaliação dos resultados. Planejar um experimento grande e complexo é um erro, o sucesso do experimento depende do nosso conhecimento sobre os importantes fatores, faixa de variação, escolha dos níveis e unidade de medição adequada. Os fatores e suas faixas de variação podem mudar durante o processo.

2.1. Modelos Fatoriais

2.1.1. Definições básicas e princípios

Planejamentos fatoriais são mais eficientes para estudar os efeitos de dois ou mais fatores em um experimento, pois, em cada replicação, todas as combinações dos níveis dos fatores são investigadas. Por exemplo, se existem *a* níveis de um fator A e *b* níveis de um fator B cada replicação contém *ab* combinações.

A variação da resposta produzida por uma variação do fator é o que chamamos de efeito de um fator. Por se referir, frequentemente, aos fatores primários de interesse, chamamos de efeito principal. Os níveis altos e baixos são denotados, respectivamente, por “+” e “-“.

Considerando um plano fatorial 2^2 com uma única replicação, o efeito do fator A pode ser calculado através da diferença entre as respostas médias no nível alto e no nível baixo de A, o procedimento é feito de modo análogo para B

Tabela 2.1 – Plano fatorial 2².

Fatores	Nível baixo de A(-)	Nível alto de A(+)
Nível alto de B(+)	y_{-+}	y_{++}
Nível baixo de B(-)	y_{--}	y_{+-}

$$A = \frac{(y_{+-} + y_{++})}{2} - \frac{(y_{--} + y_{-+})}{2}$$

$$B = \frac{(y_{-+} + y_{++})}{2} - \frac{(y_{--} + y_{+-})}{2}$$

Para fatores com mais de dois níveis, os efeitos são obtidos de forma diferente. Quando a diferença da resposta entre os níveis de um fator não é a mesma para todos os níveis de outro fator, temos interação entre os fatores. A interação pode ser percebida caso haja diferença entre a variação de A no nível mais baixo e mais alto.

$$AB = \frac{(y_{++} - y_{+-})}{2} - \frac{(y_{-+} - y_{--})}{2}$$

O modelo de regressão é uma abordagem muito útil para interpretar a interação entre os fatores. Em um experimento 2² fatorial, temos

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

A resposta e os parâmetros devem ter seus valores estimados, usamos x_1 para representar o fator A, x_2 para representar o fator B, $x_1 x_2$ representa o produto dos dois níveis dos dois fatores. Nos níveis baixos, x_1 e x_2 são representados por -1, nos níveis altos, por +1. Consequentemente, quando os sinais de x_1 e x_2 são iguais , o valor para a interação é +1, caso x_1 e x_2 sejam contrários, temos que $x_1 x_2$ será igual a -1.

É possível mostrar que os coeficientes estimados através do método de mínimos quadrados são iguais à metade da estimativa dos efeitos.

$$\hat{\beta}_1 = \frac{(y_{+-} + y_{++}) - (y_{--} + y_{-+})}{4} = \frac{1}{2} A$$

$$\hat{\beta}_2 = \frac{(y_{-+} + y_{++}) - (y_{--} + y_{+-})}{4} = \frac{1}{2} B$$

$$\hat{\beta}_{12} = \frac{(y_{++} - y_{+-}) - (y_{-+} + y_{--})}{4} = \frac{1}{2} AB$$

A estimativa para o coeficiente $\hat{\beta}_0$ será a média das quatro observações para a resposta.

$$\hat{\beta}_0 = \frac{(y_{++} + y_{-+} + y_{+-} + y_{--})}{4}$$

Quando $\hat{\beta}_{12}$ é pequeno em relação a $\hat{\beta}_1$ e $\hat{\beta}_2$, podemos ignorar a interação. Para um modelo no qual a interação não existe ou pode ser considerada desprezível, temos o seguinte ajuste.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

A representação geométrica da equação é um plano que é chamado gráfico de superfície de resposta, uma ferramenta importante de análise do experimento. Para o modelo sem interação, o gráfico será uma superfície totalmente plana, já quando consideramos a interação, a superfície adquire uma curvatura.

Geralmente, nas situações em que a interação é grande, os efeitos principais têm pequeno significado prático. Vamos supor que o fator A seja pequeno, isso nos induz a conclusão de que o efeito de A deve ser ignorado. Porém, antes devemos examinar os efeitos de A para os diferentes níveis de B. Se houver grande diferença entre as variações de A para o nível alto e o nível baixo, o efeito de A existe, mas depende do nível de B. Portanto, a interação mascara efeitos principais significantes. Para não tomarmos conclusões precipitadas, sempre que houver interação, temos de examinar os níveis de um fator fixando o nível dos outros fatores.

2.1.2. Experimentos a dois fatores

Em um exemplo de um experimento fatorial envolvendo dois fatores, um engenheiro está desenvolvendo uma bateria para usar em um dispositivo que estará sujeito a variações extremas de temperatura. O único parâmetro que pode ser selecionado é o tipo de material da bateria, ele tem três opções. O engenheiro decidiu testar as três opções de material em três níveis de temperatura: 15, 70 e 125 graus Fahrenheit. Portanto, temos um experimento 3^2 fatorial. Quatro baterias são testadas em cada combinação, então são obtidas 36 realizações. Neste problema, o engenheiro quer responder as seguintes perguntas: Quais são os efeitos do tipo de material e da temperatura na vida da bateria. Há um tipo de material que forneceria uma vida longa independentemente da temperatura?

A última questão é importante, porque caso a resposta seja "sim", o engenheiro pode tornar a bateria robusta a variação de temperatura. Este é um problema de projeto robusto de produção, muito utilizado na engenharia.

O modelo para um planejamento fatorial é descrito da seguinte forma:

$$\begin{cases} y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + e_{ijk}, \\ i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases}$$

Na equação acima, *a* é o número de níveis de fator A, *b* é o número de níveis do fator B, *n* é o número de replicações, y_{ijk} é a resposta observada considerando que o fator A está no *i-ésimo* nível e o fator B está no *j-ésimo* nível para a *k-ésima* replicação. A ordem das *abn* observações é feita de modo aleatório , por isso, este é uma planejamento completamente aleatorizado, em que μ é a média geral do efeito, τ_i é o efeito do *i-ésimo* nível de A, β_j é o efeito do *j-ésimo* nível de B, $(\tau\beta)_{ij}$ é o efeito da interação do *i-ésimo* nível de A com o *j-ésimo* nível de B e e_{ijk} é a componente correspondente ao erro aleatório. Temos ao todo *abn* observações.

Supomos que ambos os fatores são fixos e os efeitos do tratamento são definidos como desvios da média geral, então $\sum_{i=1}^{a} \tau_i = 0$ e $\sum_{j=1}^{b} \beta_j = 0$. Os efeitos da interação também são fixos tais como $\sum_{i=1}^{a} (\tau\beta)_{ij} = \sum_{j=1}^{b} (\tau\beta)_{ij} = 0$. Ambos os fatores, A e B, são de interesse.

Dessa forma, estamos interessados em testar os seguintes testes de hipóteses:

1- $H_0: \tau_1 = \tau_2 = \ldots = \tau_a = 0$

H_1: pelo menos um $\tau_i \neq 0$,

2- $H_0: \beta_1 = \beta_2 = ... = \beta_b = 0$

H_1: pelo menos um $\beta_j \neq 0$,

3- $H_0: (\tau\beta)_{ij} = 0$ para todo *i,j*

H_1: pelo menos um $(\tau\beta)_{ij} \neq 0$

Tabela 2.2 – Níveis de um plano fatorial 2².

Observações	Soma	Média
No *i-ésimo* nível do fator A	$y_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$	$\bar{y}_{i..} = \frac{y_{i..}}{bn}$
No *j-ésimo* nível do fator B	$y_{.j.} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}$	$\bar{y}_{.j.} = \frac{y_{i..}}{an}$
Na combinação do *i-ésimo* nível do fator A com o *j-ésimo* nível do fator B	$y_{ij.} = \sum_{k=1}^{n} y_{ijk}$	$\bar{y}_{ij.} = \frac{y_{ij.}}{n}$
Total	$y_{...} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$	$\bar{y}_{...} = \frac{y_{...}}{abn}$

A soma de quadrados total corrigida é escrita como:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

Tabela 2.3 - ANOVA para um experimento a dois fatores

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F_0
A	$a-1$	SS_A	$MS_A = \frac{SS_A}{a-1}$	$\frac{MS_A}{MS_E}$
B	$b-1$	SS_B	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MS_E}$
AB	$(a-1)(b-1)$	SS_{AB}	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$\frac{MS_{AB}}{MS_E}$
Erro	$ab(n-1)$	SS_E	$MS_E = \frac{SS_E}{ab(n-1)}$	
Total	$abn-1$	SS_T		

Assumindo que o modelo apresentado é adequado e que os erros são independentes e normalmente distribuídos com variância constante, a divisão do quadrado médio do determinado fator ou interação pelo MS_E possui distribuição F, cujo grau de liberdade do numerador é igual ao do fator ou interação, para o denominador, o grau de liberdade é o mesmo do resíduo.

Quando rejeitamos a hipótese nula, então pode ser necessário executar comparações individuais entre os níveis para descobrir se essa diferença entre as médias dos tratamentos é significante. Quando a interação é significante, comparações entre as médias de um fator podem ser complexas como vimos anteriormente. Uma solução seria fixar B em determinado nível e aplicar o teste de Tukey, que é baseado na

amplitude *studentizada* e consiste em definir a menor diferença significativa, para a média do fator A em cada nível.

Antes de as conclusões fornecidas pela ANOVA serem validadas, é necessário verificar a adequação do modelo, podemos usar análise de resíduos. Definimos o resíduo em um experimento fatorial como: $e_{ijk} = y_{ijk} - \hat{y}_{ijk}$, já que o valor ajustado é igual à média das observações para a combinação entre o *i-ésimo* nível de A e o *j-ésimo* nível de B.

Da mesma forma que o modelo de regressão linear, usamos o gráfico de resíduos versus valores ajustados são ferramentas que auxiliam a verificação de normalidade e variância constante.

2.1.3. Estimação dos Parâmetros do Modelo

Os estimadores são obtidos via minimização da soma dos quadrados das diferenças entre o valor estimado e observado.

Devido ao modelo possuir *ab+a+b+1* parâmetros, são necessárias *ab+a+b+1* equações normais para encontrar todos os estimadores.

$$\mu : abn\hat{\mu} + bn\sum_{i=1}^{a}\hat{\tau}_i + an\sum_{j=1}^{b}\hat{\beta}_j + n\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{ij} = y_{...}$$

$$\tau_i : bn\hat{\mu} + bn\hat{\tau}_i + n\sum_{j=1}^{b}\hat{\beta}_j + n\sum_{j=1}^{b}(\tau\beta)_{ij} = y_{i..}$$

$$i = 1, 2, ..., a$$

$$\beta_j : an\hat{\mu} + n\sum_{i=1}^{a}\hat{\tau}_i + an\hat{\beta}_j + n\sum_{i=1}^{a}(\tau\beta)_{ij} = y_{.j.}$$

$$j = 1, 2, ..., b$$

$$(\tau\beta)_{ij} : n\hat{\mu} + n\hat{\tau}_i + n\hat{\beta}_j + n(\tau\beta)_{ij} = y_{ij.}$$

$$i = 1, 2, ..., a$$

$$j = 1, 2, ..., b$$

Já que existem equações linearmente dependentes, para obter solução do modelo, devemos impor as seguintes restrições:

$$\sum_{i=1}^{a}\hat{\tau}_i = 0; \sum_{j=1}^{b}\hat{\beta}_j = 0; \sum_{i=1}^{a}(\tau\beta)_{ij} = 0$$

$$\text{e } \sum_{j=1}^{b}(\tau\beta)_{ij} = 0$$

Então, encontramos os estimadores dos parâmetros.

$$\hat{\mu} = \bar{y}_{...}$$

$$\hat{\tau}_i = \bar{y}_{i..} - \bar{y}_{...}; i = 1, 2, ..., a$$

$$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}; j = 1, 2, ..., b$$

$$(\tau\beta)_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}; \begin{matrix} i = 1, 2, ..., a \\ j = 1, 2, ..., b \end{matrix}$$

Os resultados obtidos no modelo para dois fatores podem ser estendidos para o caso geral.

2.1.4. Ajuste de curvas e superfícies de resposta

A ANOVA sempre trata os fatores de um experimento como se fossem qualitativos e categóricos, mas, geralmente, o experimento possui pelo menos um fator quantitativo. Portanto, é útil termos uma relação entre os níveis de um fator e a resposta que é encontrada através do ajuste da curva de resposta. Essa equação pode ser usada para prever a resposta de acordo com os níveis do fator.

Se tivermos pelo menos dois fatores quantitativos, podemos ajustar uma superfície de resposta. Dessa forma, pode se prever a resposta dada uma combinação de níveis dos fatores. Normalmente, métodos de regressão linear são usados para ajustar o modelo.

2.1.5. Blocagem no experimento fatorial

Discutimos o planejamento fatorial em um contexto de experimento completamente aleatorizado. No entanto, às vezes, não é possível que todas as realizações sejam totalmente aleatorizadas. Por exemplo, quando existe um fator de ruído, uma alternativa é realizar o experimento em blocos.

Suponha um experimento fatorial com os dois fatores, A e B, e a presença de interação AB. Para realizar o experimento, é necessário uma matéria-prima especifica que não está disponível em quantidade para executar *abn* combinações de tratamentos num único lote. Observou-se que seria possível gerar ab combinações com apenas um lote.

Conseqüentemente, os lotes são uma restrição a aleatorização. O modelo para o experimento incluindo a blocagem é

$$y_{ijk} = \mu + \tau_i + (\tau\beta)_{ij} + \delta_k + e_{ijk}$$
$$i = 1,2,...,a;\ j = 1,2,...,b; k = 1,2,...,n$$

δ_k representa o efeito do *k-ésimo* bloco.

O modelo supõe desprezível a interação entre tratamentos e blocos. Além da matéria-prima, podem existir outras restrições a aleatorização como: tempo. Por exemplo, o experimento não pode ser feito inteiramente em um dia, então, seria executada uma replicação num primeiro dia e outra replicação em um segundo dia. Cada dia representa um bloco.

2.2. Experimentos 2^K

Uma replicação desse experimento requer 2^k observações, por isso, é chamado de 2^k fatorial. Lembrando que estamos supondo que os fatores são fixos, o experimento é completamente aleatorizado e as suposições de normalidade estão satisfeitas. Os experimentos 2^k são amplamente utilizados em estágios iniciais de pesquisas para selecionar dentre muitos fatores quais são realmente relevantes.

2.2.1. O Planejamento 2^2

Em um experimento 2^2 fatorial com fatores A e B, cada um com dois níveis: alto e baixo. O experimento também pode ser interpretado através da figura abaixo, um quadrado onde cada vértice representa uma combinação de tratamento.

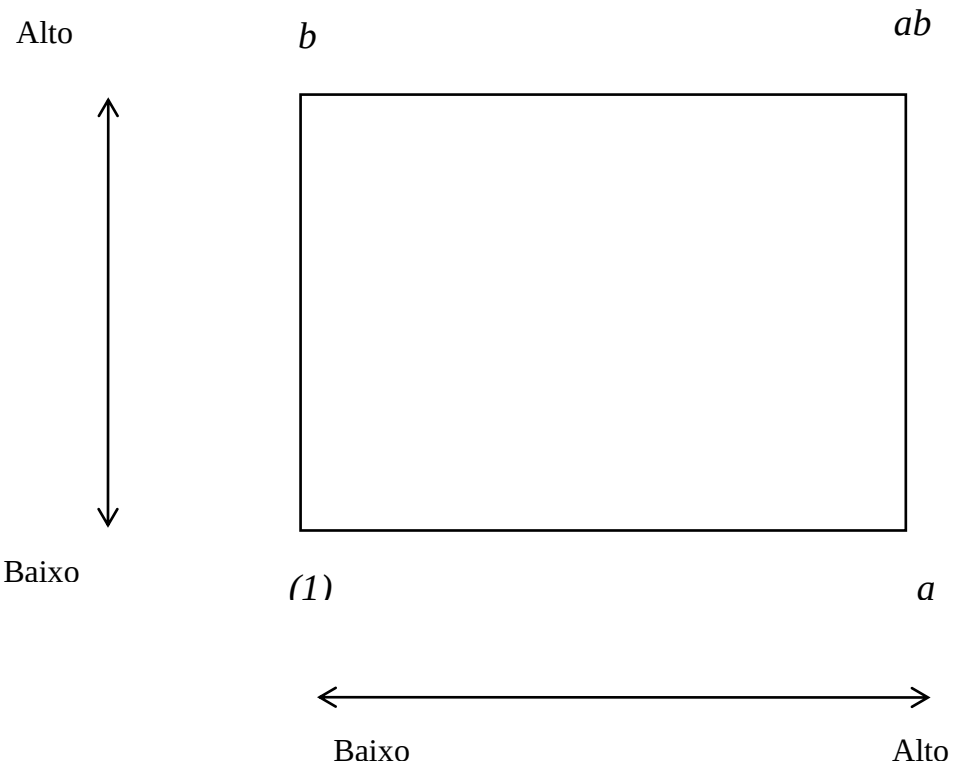

Figura 2.1 – Plano fatorial 2^2

O nível alto de A com o nível baixo de B é representado por *a*, o nível baixo de A com o nível alto de B é representado por *b*, quando A e B estão no nível alto, temos *ab*, e os dois fatores no nível baixo, temos *(1)*. Como se mostrou na figura acima, essa notação representa o total de cada combinação de tratamento. A partir dessa notação, calcularemos os efeitos principais e a interação.

$$A = \frac{1}{2n}([ab - b] + [a - (1)])$$

$$= \frac{1}{2n}(ab + a - b - (1))$$

$$B = \frac{1}{2n}([ab - a] + [b - (1)])$$

$$= \frac{1}{2n}(ab + b - a - (1))$$

$$AB = \frac{1}{2n}([ab - b] - [a - (1)])$$

$$= \frac{1}{2n}(ab + (1) - a - b)$$

Repare que o efeito de A é a média entre os efeitos de A no nível mais baixo e mais alto de B. O efeito de B é calculado de maneira análoga. O efeito da interação AB é a média da diferença do efeito de A no nível mais baixo e mais alto de B, mas também,

pode ser definido como a média da diferença entre o efeito de B no nível mais alto e mais baixo de A.

Existem excelentes pacotes estatísticos capazes de realizar todos os cálculos para um experimento 2^k fatorial, mas podemos fazer isso manualmente. Para determinar as somas de quadrados de A, B e AB, usamos os contrastes, também chamados de efeito total. Por exemplo, o contraste de A, usado para estimação de A, é

$$Contraste_A = ab + a - b - (1)$$

A soma de quadrados é o quadrado do contraste dividido pelo número total de observações.

$$SS_A = \frac{[ab + a - b - (1)]^2}{4n}$$

$$SS_B = \frac{[ab + b - a - (1)]^2}{4n}$$

$$SS_{AB} = \frac{[ab + 1 - a - b]^2}{4n}$$

A ANOVA também auxilia para confirmar a interpretação da magnitude e direção dos efeitos, mas não deve ser a única ferramenta a ser utilizada, pois não contém todas as informações necessárias.

Note que os contrastes para os efeitos A, B e AB são ortogonais. Consequentemente, o experimento fatorial 2^k é ortogonal.

Um experimento fatorial pode facilmente ser expresso por meio de um modelo de regressão. Desta forma, temos o modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

x_1 é uma variável codificada que representa o fator A, x_2 é uma variável codificada que representa o fator B.

Antes de validar o modelo, precisamos realizar a análise de resíduos para verificar as suposições de normalidade e variância constante.

2.2.2. Planejamento 2^3 fatorial

Supondo que agora temos três fatores de interesse: A, B e C, cada um com dois níveis. As oito combinações de tratamento são representadas por meio de um cubo. Cada vértice é uma combinação.

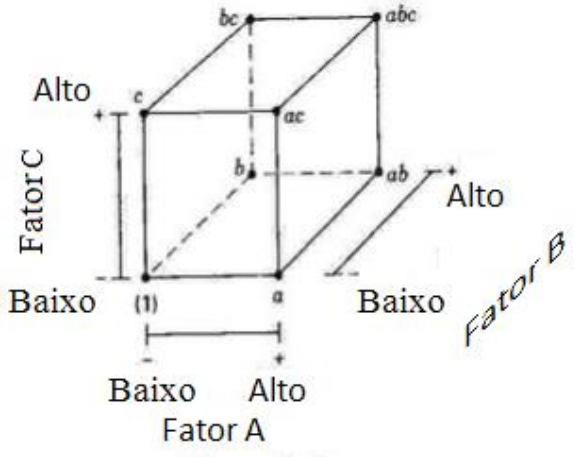

Figura 2.2 – Plano fatorial 2^3

A estimativa do efeito de A pode ser interpretado como uma média de quatro situações: o efeito de A quando B e C estão no nível baixo, *[a-(1)]/n*, o efeito de A quando B está no nível baixo e C está no nível alto, *[ac-c]/n*, o efeito de A quando B está no nível alto e C está no nível baixo, *[ab-b]/n*, e o efeito de A quando B e C estão no nível alto, *[abc-bc]/n*. Dessa forma, o efeito médio de A é

$$A = \frac{1}{4n}[a - (1) + ab - b + ac - c + abc - bc]$$

Do mesmo modo, obtemos os outros efeitos principais:

$$B = \frac{1}{4n}[abc + ab + bc + b - ac - a - c - (1)]$$

$$C = \frac{1}{4n}[abc + ac + bc + c - ab - a - b - (1)]$$

Note que a equação para o efeito de A é um contraste entre as combinações que estão no vértice da face do lado direito e da face do lado esquerdo do cubo na figura 2.2. A interação AB é a diferença da média do efeito de A nos dois níveis de B.

$$AB = \frac{1}{4n}\left[(1) - a - b + ab + c - ac - bc + abc\right]$$

De modo análogo, podemos encontrar os efeitos das interações AC e BC.

$$AC = \frac{1}{4n}\left[(1) - a + b - ab - c + ac - bc + abc\right]$$

$$BC = \frac{1}{4n}\left[(1) + a - b - ab - c - ac + bc + abc\right]$$

Podemos definir a interação ABC como a média da diferença média de AB entre os dois níveis de C.

$$ABC = \frac{1}{4n}[(abc - bc) - (ac - c) - (ab - b) + (a - (1))]$$

$$= \frac{1}{4n}\left[abc - bc - ac + c - ab + b + a - (1)\right]$$

Em todas estas equações, as quantidades entre colchetes são os contrastes nas combinações de tratamentos. A partir dos contrastes, podemos construir uma tabela com os sinais de mais ou menos. Os sinais para as interações são obtidos pela multiplicação das colunas. Por exemplo, os sinais da interação AB são obtidos por meio do produto das colunas A e B em cada linha. Com a exceção da coluna I, que possui apenas sinais positivos, cada coluna possui quantidades iguais de sinais positivos e negativos. A soma do produto de sinais para duas quaisquer duas colunas é zero. Qualquer coluna multiplicada pela coluna I se mantém inalterada. O produto de quaisquer duas colunas forma uma coluna que pertence à tabela.

Estas propriedades são implicações da ortogonalidade do planejamento 2^3 e dos contrastes utilizados para estimar os efeitos. A soma de quadrados para os efeitos são calculados facilmente porque cada efeito tem um contraste que corresponde a 1 grau de liberdade. Em um planejamento 2^3 com n replicações, temos que para qualquer efeito:

$$SQ = \frac{(Contraste)^2}{N};\ N = 2^3 n$$

Da mesma forma que os planejamentos 2^2, pode ser construído um modelo de regressão para os planejamentos 2^3. O modelo considerando todos os efeitos principais e interações como significantes teria o seguinte formato:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_{12} x_1 x_2 + \hat{\beta}_{13} x_1 x_3 + \hat{\beta}_{23} x_2 x_3 + \hat{\beta}_{123} x_1 x_2 x_3$$

As variáveis codificadas x_1, x_2 e x_3 representam respectivamente A, B e C. Os termos $x_1 x_2$, $x_1 x_3$, $x_2 x_3$ e $x_1 x_2 x_3$ são, respectivamente, as interações AB, AC, BC e ABC. Lembrando que quando existir interação, as linhas de contorno da resposta não serão retas.

A análise da variância é uma maneira formal para determinar quais efeitos dos fatores são diferentes de zero. Há vários outros métodos para julgar a significância dos efeitos. O erro padrão dos efeitos é utilizado para construir um intervalo de confiança. O erro padrão é calculado facilmente. Supondo que temos *n* replicações de cada um das 2^k combinações de tratamentos.

$$s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2, \; i = 1, 2, ..., 2^k$$

Essa é uma estimativa da variância para *i-ésima* realização. 2^k estimativas da variância podem ser combinadas para conseguir uma estimativa geral da variância.

$$s^2 = \frac{1}{2^k (n-1)} \sum_{i=1}^{2^k} s_i^2$$

Essa estimativa é também chamada de quadrado médio do resíduo. A variância de cada efeito estimado é

$$V(Efeito) = V\left(\frac{Contraste}{n2^{k-1}} \right) = \frac{1}{\left(n2^{k-1} \right)^2} V(Contraste)$$

Cada contraste é uma combinação linear dos 2^k totais de tratamentos e cada total consiste de *n* observações. Portanto,

$$V(Efeito) = \frac{1}{n2^{k-2}} \sigma^2 \text{ e } V(Contraste) = n2^k \sigma^2$$

O estimador do erro padrão é raiz quadrada da expressão para variância do efeito substituindo-se σ^2 por s^2.

$$se(Efeito) = \frac{2s}{\sqrt{n2^k}}$$

A significância de qualquer efeito é testada por meio da seguinte estatística.

$$t_0 = \frac{Efeito}{se(Efeito)}$$

Essa estatística possui distribuição t de Student com *N-p* graus de liberdade. Onde *N* é o número de observações e *p* é o número de parâmetros do modelo. Portanto, um intervalo de confiança com *100(1-α)%* para determinado efeito é

$$Efeito \pm t_{\alpha/2, N-p} se(Efeito)$$

2.2.3. Planejamento 2^k fatorial

Os métodos apresentados anteriormente podem ser generalizados no planejamento 2^k fatorial, que contém *k* fatores com dois níveis cada um. Um modelo estatístico para o

2^k fatorial inclui *k* efeitos principais, $\binom{k}{2}$ interações de ordem 2, $\binom{k}{3}$ interações de ordem 3,..., e uma interação de ordem *k*, somando um total de $2^k - 1$ efeitos.

Em um procedimento de análise de experimentos dessa espécie, devemos seguir algumas etapas: (1) estimar os efeitos dos fatores, (2) formular o modelo inicial, (a) se for replicado, ajustar o modelo replicado, (b) caso contrário, formular o modelo utilizando o gráfico de probabilidade normal dos efeitos, (3) executar teste estatísticos, (4) refinar o modelo, (5) analisar os resíduos, (6) interpretar os resultados.

A primeira etapa fornece uma noção ao experimentador de quais fatores e interações são importantes e como esses fatores podem ser ajustados para melhorar a resposta. Na terceira etapa utilizamos a análise da variância para testar a significância dos fatores e interações. A quarta etapa consiste em remover as variáveis não significantes do modelo completo, e na quinta etapa, checamos as suposições e a adequação do modelo.

Em certas situações, o refinamento do modelo poderá ocorrera pós a análise dos resíduos, casos em que o modelo é inadequado ou quando as suposições são violadas.

Ocasionalmente, pode ser necessário calcular os efeitos estimados ou a soma de quadrados manualmente, portanto precisamos determinar os contrastes, uma opção é utilizar a tabela de sinais, porém, quando k é grande, isso se torna muito trabalhoso.

Outra alternativa é expandir o lado direito da equação. O sinal em cada parêntese é negativo se o fator está incluso, e positivo, caso contrário. Por exemplo, o contraste AB em plano 2^3, temos

$$Contraste_{AB} = (a-1)(b-1)(c+1)$$
$$= abc + ab + c + (1) - ac - bc - a - b$$

Podemos definir um contraste como:

$$Contraste_{AB...K} = (a \pm 1)(b \pm 1)...(k \pm 1)$$

Desta forma, determinamos o efeito estimado e a soma de quadrados.

$$AB...K = \frac{2}{n2^k}(Contraste_{AB...K})$$

$$SS_{AB...K} = \frac{1}{n2^k}(Contraste_{AB...K})^2$$

2.2.4. Plano 2^k com uma única replicação

Geralmente, os recursos de um experimento são limitados. Consequentemente, o número de replicações é restrito, e em alguns casos, é possível obter apenas uma replicação. Nessas situações, se *y* tem variabilidade alta, podemos tomar conclusões erradas sobre o experimento. Com apenas uma replicação, não há como ter uma estimativa interna do erro (ou "erro puro"). Uma abordagem é assumir que as interações de ordem mais alta são desprezíveis e combinar seus quadrados médios para estimar o erro. No entanto, quando as interações de ordem mais altas são significantes, a combinação de quadrados médios não é adequada. Nesse caso, o método a ser usado é atribuído a Daniel (1959), e consiste em examinar o gráfico de probabilidade normal da estimativa dos efeitos.

Neste gráfico, os efeitos desprezíveis possuem distribuição normal com média zero e variância σ^2 e tendem a cair ao longo da reta, já os efeitos significantes, com médias diferentes de zero, se encontrarão distantes da reta. Dessa forma, o modelo conterá apenas os efeitos que são não nulos. Os efeitos aparentemente desprezíveis serão combinados para estimar o erro.

2.2.5. Propriedades do plano fatorial

Os planos fatoriais 2^k têm muitas propriedades úteis e interessantes. Considerando o caso mais simples, um experimento 2^2 com uma única replicação, tem quatro combinações de tratamento: *(1), a, b* e *ab.* Nesse caso, obtemos o seguinte modelo ajustado.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

Lembrando que as variáveis x_1 e x_2 representam os fatores principais e $x_1 x_2$ é a interação entre os dois fatores. Cada combinação de tratamento pode ser escrita como:

$$(1) = \beta_0 + \beta_1(-1) + \beta_2(-1) + \beta_{12}(-1)(-1) + \varepsilon_1$$
$$a = \beta_0 + \beta_1(1) + \beta_2(-1) + \beta_{12}(1)(-1) + \varepsilon_2$$
$$b = \beta_0 + \beta_1(-1) + \beta_2(1) + \beta_{12}(-1)(1) + \varepsilon_3$$
$$ab = \beta_0 + \beta_1(1) + \beta_2(1) + \beta_{12}(1)(1) + \varepsilon_4$$

Torna-se mais fácil se escrevermos as quatro equações em forma matricial.

y=Xβ+ε, onde $\mathbf{y} = \begin{bmatrix} (1) \\ a \\ b \\ ab \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \end{bmatrix}$, e $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$

O vetor β contém os coeficientes do modelo de regressão obtido pelo método mínimos quadrados. Os erros do modelo são representados por **ε.** Provaremos, no próximo capítulo, que $\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Dessa forma, para encontrarmos **β,** precisamos calcular as matrizes $\mathbf{X}'\mathbf{X}$ e $\mathbf{X}'\mathbf{y}$. Como, o planejamento é ortogonal, a matriz $\mathbf{X}'\mathbf{X}$ é diagonal.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} (1) \\ a \\ b \\ ab \end{bmatrix} = \begin{bmatrix} (1) + a + b + ab \\ -(1) + a - b + ab \\ -(1) - a + b + ab \\ (1) - a - b + ab \end{bmatrix}$$

Consequentemente,

$$= \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}^{-1} \begin{bmatrix} (1) + a + b + ab \\ -(1) + a - b + ab \\ -(1) - a + b + ab \\ (1) - a - b + ab \end{bmatrix} = \begin{bmatrix} \dfrac{(1) + a + b + ab}{4} \\ \dfrac{-(1) + a - b + ab}{4} \\ \dfrac{-(1) - a + b + ab}{4} \\ \dfrac{(1) - a - b + ab}{4} \end{bmatrix}$$

Os estimadores dos coeficientes de regressão são iguais à metade dos efeitos estimados. Isso mostra que a variância de qualquer coeficiente é fácil de encontrar.

$$Var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\sigma^2}{4}$$

Todos os coeficientes possuem a mesma variância. Não há outro experimento de quatro observações com as variáveis codificadas pelos valores ±1 que tenha menor variância. O valor máximo do determinante da matriz $\mathbf{X}'\mathbf{X}$ em um experimento de quatro observações é 256. O volume da região de confiança conjunta que contém todos os coeficientes de regressão é inversamente proporcional à raiz quadrada do determinante de $\mathbf{X}'\mathbf{X}$. Portanto, para construir a menor região de confiança possível, devemos escolher um planejamento com o valor máximo possível do determinante de $\mathbf{X}'\mathbf{X}$.

Um plano que minimiza a variância dos coeficientes do modelo de regressão é chamado de D-ótimo. O plano 2^k é D-ótimo para ajustar modelo de primeira ordem ou modelos de primeira ordem com interação.

$$Var(\hat{y}(x_1, x_2)) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2)$$

$$= \frac{\sigma^2}{4}(1 + x_1^2 + x_2^2 + x_1^2 x_2^2)$$

A variância máxima da previsão da resposta ocorre quando $x_1 = x_2 = \pm 1$ e é igual a σ^2. Agora, precisamos saber o melhor valor possível da variância da previsão que podemos alcançar.

O menor valor possível da máxima variância da previsão no espaço [-1,1] é $p\sigma^2/N$, onde *p* é o número de parâmetros do modelo e *N* é o número total de observações. Em um experimento 2^2 fatorial com uma única replicação, temos *N=4* e *p=4*. Então, o modelo ajustado para os dados minimiza a máxima variância da previsão sobre a região do plano. Um plano que possui esta propriedade é chamado de G-ótimo. Geralmente, os planos 2^k são G-ótimos para ajustar modelos de primeira ordem ou modelos de primeira ordem com interações.

2.2.6. O uso de variáveis codificadas

Em quase todos os momentos, usamos a variáveis codificadas ao invés de trabalhar com os valores originais. Os resultados obtidos com as variáveis originais podem ser muito diferentes se comparados às analises com variáveis codificadas e, geralmente, os resultados finais são de difícil interpretação.

Na análise das variáveis codificadas, podemos comparar diretamente os coeficientes do modelo, ou seja, não há unidade própria, o efeito da variação de todos os fatores são medidos sobre o mesmo espaço [-1,1] e são estimados com a mesma precisão. Em um modelo com as variáveis originais, os fatores não são ortogonais, portanto os coeficientes têm unidades próprias e suas estimativas possuem precisões diferentes.

Normalmente, preferimos usar a análise em escala codificada, porque isto nos permite observar a importância relativa dos efeitos.

2.3. Experimentos fatoriais fracionados 2^{k-p}

Considerando um plano 2^6 fatorial não replicado, temos 63 graus de liberdade: 6 graus para os efeitos principais, 15 graus para as interações entre dois fatores. Portanto,

apenas 21 graus de liberdade estão associados aos efeitos que provavelmente são os de maior interesse.

Supondo que é possível assumir que as interações de ordem mais altas são desprezíveis, as informações sobre os efeitos principais e interações podem ser obtidas executando apenas uma fração do experimento fatorial completo.

Os fatoriais fracionados são muito usados quando se tem um experimento com muitos fatores a serem considerados e o objetivo é identificar quais fatores possuem os maiores efeitos. Estes são chamados de experimento de seleção e, geralmente, são executados nas etapas iniciais de um projeto quando muitos daqueles fatores considerados inicialmente têm efeito pequeno ou inexistente sobre a resposta. Os fatores classificados como importante são investigados de forma mais especifica nas etapas seguintes.

O uso bem sucedido dos planejamentos fatoriais fracionados é baseado em três idéias chaves.

O princípio dos efeitos esparsos - Quando há varias variáveis, o sistema ou processo, provavelmente, é dirigido por alguns dos efeitos principais e de interações de baixa ordem.

A propriedade da projeção – Os fatoriais fracionados podem ser projetados dentro de planos maiores no subconjunto de efeitos significantes.

Experimentação sequencial – É possível combinar as realizações de dois ou mais fatoriais fracionados para juntá-las sequencialmente em um plano maior para estimar os efeitos dos fatores e interações de interesse.

2.3.1. Definições e Princípios Básicos

Imaginemos um experimento de três fatores cada um com dois níveis. Desse modo, teríamos oito (2^3) combinações de tratamento. Suponha que os recursos disponíveis permitem apenas quatro combinações. Dessa forma, será realizado uma fração ½ do planejamento 2^3 ou um planejamento 2^{3-1}.

Supondo que selecionamos as quatro combinações: *a, b, c* e *abc*. Na tabela de sinais abaixo, as combinações executadas estão na parte de cima e possuem o sinal positivo na coluna ABC. A primeira coluna da tabela é I, que tem apenas sinais (+). Observe que na parte superior da tabela, os sinais da coluna I e ABC são iguais. Consequentemente, a relação de definição desse plano fatorial fracionado é I=ABC. Geralmente, a relação de definição da fração do planejamento é sempre o conjunto de todas as colunas, que são iguais à coluna identidade I.

Tabela 2.5 - Tabela de sinais para o fatorial 2^3.

Combinação de tratamentos	I	A	B	C	AB	AC	BC	ABC
A	+	+	-	-	-	-	+	+
B	+	-	+	-	-	+	-	+
C	+	-	-	+	+	-	-	+
ABC	+	+	+	+	+	+	+	+
AB	+	+	+	-	+	-	-	-
AC	+	+	-	+	-	+	-	-
BC	+	-	+	+	-	-	+	-
(1)	+	-	-	-	+	+	+	-

Também com o auxilio da tabela, vemos que as combinações lineares das observações são utilizadas para estimar os efeitos principais.

$$[A] = \frac{1}{2}(abc + a - b - c)$$

$$[B] = \frac{1}{2}(abc + b - a - c)$$

$$[C] = \frac{1}{2}(abc + c - a - b)$$

As interações entre dois fatores também são estimadas por meio das combinações lineares.

$$[BC] = \frac{1}{2}(abc + a - b - c)$$

$$[AC] = \frac{1}{2}(abc + b - a - c)$$

$$[AB] = \frac{1}{2}(abc + c - a - b)$$

Dessa forma, *[A]=[BC]*, *[B]=[AC]* e *[C]=[AB]*, o que torna impossível distinguir entre A e BC, B e AC, e C e AB. Quando estimamos A, B e C, na realidade, estamos estimando A+BC, B+AC e C+AB. Dois ou mais efeitos com esta propriedade são chamados de efeitos associados.Essa relação pode ser indicada pela seguinte notação.

$$[A] \rightarrow A + BC, [B] \rightarrow B + AC \text{ e } [C] \rightarrow C + AB.$$

Os efeitos associados podem ser encontrados usando a relação de definição, nesse caso, *I=ABC.* Dado o efeito fatorial, o seu efeito associado é obtido multiplicando ambos os lados da relação de definição pelo efeito fatorial. Vamos usar como exemplo o efeito A.

$$A.I = A.ABC = A^2BC = BC$$

Agora suponha que selecionamos a outra fração, que corresponde a parte inferior da tabela: *(1),ab, ac* e *bc.* Nesse caso, a relação de definição será *I=-ABC.* As combinações lineares das observações, digamos [A'], [B'] e [C'], da fração alternativa, são calculadas de modo análogo a [A], [B] e [C]. Consequentemente, obtemos

$$[A]' \rightarrow A\text{-}BC$$
$$[B]' \rightarrow B\text{-}AC$$
$$[C]' \rightarrow C\text{-}AB$$

Desta maneira, quando estimamos A, B e C, na verdade, estamos estimando *A-BC*, *B-AC* e *C-AB.* Na prática, não interessa qual fração é utilizada, pois ambas as frações são da mesma família, as duas juntas formam um experimento 2^3 completo.

Este exemplo é um plano 2^{3-1} de resolução III, ou um plano 2^{3-1}_{III}, onde os efeitos principais estão associados a interações entre dois fatores. Uma definição geral para um fatorial fracionado de resolução R é: um plano é de resolução R se nenhum efeito fatorial com p fatores está associado a outro efeito fatorial com menos de *R-p* fatores.

Outra definição utilizada é: uma fração é de resolução R se o comprimento da menor palavra da relação de definição é R. Neste exemplo, a única palavra na relação de definição é ABC, o comprimento tem três letras.

Nos planejamentos de resolução III, nenhum efeito principal está associado a outro efeito principal, mas os efeitos principais podem ser associados a interações de dois fatores e interações de dois fatores podem estar associadas entre si.

Nos planejamentos de resolução IV, nenhum efeito principal está associado a outro efeito principal e nem com interações de dois fatores, mas interações de dois fatores

estão associadas entre si. Um exemplo seria um plano 2^{4-1} com relação de definição *I=ABCD*, que é um plano 2^{4-1}_{IV}.

Nos planejamentos de resolução V, nenhum efeito principal ou interação de dois fatores estão associados a outro efeito principal ou interação de dois fatores, mas interações de dois fatores estão associadas com interações de três fatores. Um exemplo é um plano fatorial 2^{5-1}, com relação de definição *I=ABCD*, é um plano 2^{5-1}_{V}.

2.3.2. Construção e Análise da Fração ½ de um plano 2^k fatorial.

Para construir uma fração ½ do plano 2^k com a mais alta resolução, escrevemos um plano 2^{k-1} completo, depois, é adicionada uma coluna formada pelo produto dos sinais das colunas à esquerda. A fração alternativa é obtida multiplicando a coluna da fração original por -1. Como exemplo, o plano 2^{3-1}_{III}, obtido utilizando um plano 2^2 fatorial.

Tabela 2.6 - Plano fatorial 2^{3-1}_{III}.

Observações	A	B	C=AB	C=-AB
1	-	-	+	-
2	+	-	-	+
3	-	+	-	+
4	+	+	+	-

Para um plano 2^{k-1} completo, *I=ABC...K*, o que nos leva a *K=ABC...(K-1)*. Dessa forma, a coluna K terá o mesmos sinais da interação *ABC...(K-1)*.

Qualquer interação poderia ser usada para gerar a fração, a coluna correspondente a fator K. No entanto, o uso de qualquer outro efeito diferente de *ABC...(K-1)* não produziria plano de mais alta resolução.

Qualquer plano fatorial fracionado de resolução *R* contém um plano fatorial completo em qualquer subconjunto de *R-1* fatores. Por exemplo, se um pesquisador tem vários fatores de potencial interesse, mas acredita que somente *R-1* fatores são importantes. Portanto, esse plano fatorial fracionado de resolução R é uma escolha apropriada.

2.3.3. Fração ¼ do Planejamento 2^k

Para um número grande de fatores pode ser necessário usar frações ainda menores de experimento 2^k fatorial. Consideremos a fração de um quarto de um plano 2^k fatorial. Este plano contém 2^{k-2} observações é conhecido como fatorial fracionado 2^{k-2}.

A construção de um fatorial fracionado 2^{k-2} consiste na utilização de um fatorial completo com *k-2* fatores, onde serão adicionadas duas colunas com escolhas adequadas de interações envolvendo os primeiros *k-2* fatores. Consequentemente, a fração de um quarto de um plano 2^k tem dois geradores. Se P e Q representam os dois geradores escolhidos, então *I=P* e *I=Q* são chamadas de relações geradoras para o plano. O que determinará a fração ¼ produzida serão os sinais de P e Q. A fração na qual P e Q tem simultaneamente sinal positivo é a fração principal.

A relação de definição completa para o plano consiste de todas as colunas que são iguais a coluna identidade I. Esta tem P, Q e a interação generalizada PQ. Então, a relação de definição completa é *I=P=Q=PQ*. Os efeitos associados são produzidos pela multiplicação de cada efeito pelos elementos da relação de definição.

O pesquisador deve ser cuidadoso na escolha dos geradores de modo que efeitos potencialmente importantes não estejam associados entre si.

Como exemplo, considere um plano 2^{6-2}. Escolhemos os geradores ABCE e BCDF, portanto, a interação generalizada é ADEF. Temos a seguinte relação de definição completa.

$$I=ABCE=BCDF=ADEF$$

Através da quantidade de letras da menor palavra, vemos que o plano possui resolução IV. Os efeitos associados multiplicando a equação acima por cada efeito. Dessa forma,

$$A=BCE=ABCDF=DEF$$

Nesse caso, todos os efeitos principais estão associados a interações de três e cinco fatores. Quando estimamos A, na verdade, estamos estimando A+BCE+ABCDF+DEF. Se as interações com três fatores ou mais são consideradas desprezíveis, o plano nos dará estimativas claras dos efeitos principais.

A construção deste plano se iniciará a partir de um plano fatorial 2^4 completo com os fatores: A, B, C e D. As colunas dos fatores E e F serão obtidas respectivamente por meio das interações ABC e BCD.

Tabela 2.7 - Construção de um plano 2^{6-2}_{IV} com geradores ABCE e BCDF.

Observações	A	B	C	D	E=ABC	F=BCD
1	-	-	-	-	-	-
2	+	-	-	-	+	-
3	-	+	-	-	+	+
4	+	+	-	-	-	+
5	-	-	+	-	+	+
6	+	-	+	-	-	+
7	-	+	+	-	-	-
8	+	+	+	-	+	-
9	-	-	-	+	-	+
10	+	-	-	+	+	+
11	-	+	-	+	+	-
12	+	+	-	+	-	-
13	-	-	+	+	+	-
14	+	-	+	+	-	-
15	-	+	+	+	-	+
16	+	+	+	+	+	+

O plano descrito acima também poderia se tornar um plano 2^{4-1} com duas replicações nos fatores ABDE, BCDF e ADEF, pois estes são os elementos geradores.

2.3.4. Plano Fatorial Fracionado Geral

Um plano fatorial fracionado é uma fração $1/2^p$ de um plano fatorial 2^k. Para a construção desses planos, são necessários p geradores independentes. A relação de definição inclui os p geradores e as $2^p - p - 1$ interações generalizadas. Devemos tomar cuidado na escolha dos geradores, efeitos de potencial interesse não podem estar associados entre si. Para cada efeito, temos $2^p - 1$ efeitos associados.

Seguindo a mesma lógica dos casos especiais apresentados anteriormente, para obter o conjunto de efeitos associados, devemos multiplicar os efeitos fatoriais por cada elemento da relação de definição. As frações obtidas dependem dos sinais dos geradores, a fração principal é aquela na qual todos os sinais são positivos.

Se tivermos um valor alto para k, geralmente assumimos que as interações de ordem mais alta são desprezíveis, simplificando a estrutura de associação. É importante selecionar os p geradores para o plano 2^{k-p} de modo que possamos obter a melhor

estrutura possível dos efeitos associados. Um critério é utilizar geradores que resultem em um plano com maior resolução possível. Como exemplo, considere um plano 2^{6-2}_{IV}, onde são usados os geradores:*E=ABC* e *F=ABCD*. Este plano possui a resolução IV, a maior possível. Imagine se tivéssemos selecionado os geradores: *E=ABC* e *F=ABCD.*

Então, temos a seguinte relação de definição completa: *I=ABCE=ABCDF=DEF*, o que significa que o plano teria resolução III. Este plano não é uma boa escolha, porque sacrificaria desnecessariamente informações sobre as interações. Algumas vezes, somente a resolução é insuficiente para diferenciar os planos.

Três escolhas de geradores para o plano 2^{7-2}_{IV}.

Geradores do Plano A *F=ABC, G= BCD* *I=ABCF=BCDG=ADFG*	Geradores do Plano B *F=ABC, G=ADE* *I=ABCF=ADEG=BCDEFG*	Geradores do plano C *F=ABCD, G=ABDE* *I=ABCDF=ABDEG=CEFG*
AB=CF	*AB=CF*	*CE=FG*
AC=BF	*AC=BF*	*CF=EG*
AD=FG	*AD=EG*	*CG=EF*
AG=DF	*AE=DG*	
BD=CG	*AF=BC*	
BG=CD	*AG=DE*	
AF=BC=DG		

Na tabela acima, temos três planos de resolução IV: A, B e C. Estes planos possuem diferentes geradores, portanto suas estruturas de associação serão distintas. Assumimos que interações com mais de dois fatores são consideradas desprezíveis.

Observamos que o plano C possui a estrutura de associação menos extensa, por isso é considerado a melhor escolha para um plano 2^{7-2}_{IV}. O plano C é o que apresenta menor número de palavras com comprimento mínimo (4 letras)**.**

Na análise de um fatorial fracionado, os efeitos fatoriais podem ser estimados da mesma forma que um plano fatorial completo, usando a tabela de sinais que permite a construção da fração como vimos anteriormente. Considerando o *i-ésimo* efeito fatorial e seja C_i o seu contraste correspondente, obtido usando os sinais (+) e (-) da tabela, apenas os contrastes dos efeitos estimáveis são calculados, o número de observações é $N=2^{k-p}$.

Então, $Efeito_i = \frac{2C_i}{N} = \frac{C_i}{(N/2)}$. A soma de quadrados é $SS_i = \frac{C_i^2}{N}$. No plano 2^{k-p}, temos $2^{k-p}-1$ efeitos estimáveis.

CAPÍTULO 3 - Introdução aos Modelos Lineares Generalizados

Este capítulo está baseado nos seguintes textos: Turkman e Silva (2000) e Myers et al (2010).

Muitas vezes, em estudos estatísticos, um dos principais objetivos é analisar a influência que uma ou mais variáveis explicativas têm sobre uma variável de interesse. Geralmente, isto é feito por meio de um estudo de modelo de regressão que relacione essa variável de interesse com as variáveis explicativas.

Devido à sua grande utilidade e estrutura flexível, o modelo linear normal dominou a modelagem estatística até meados do século XX. No entanto, este modelo não é adequado para todas as situações. Isto acontece quando não há normalidade da variável resposta e a variância não é constante.

Para tratar esse problema, uma alternativa é a transformação dos dados, porém pode ser complicado trabalhar em escala transformada, e, muitas vezes, há obtenção de valores sem sentido para o modelo. Também foram desenvolvidos modelos para situações em que a normalidade não era adequada, por exemplo: os modelos probit e logit e modelos lineares para dados de contagens.

Todos estes modelos apresentam algo em comum: a variável resposta segue uma distribuição da família exponencial. A partir disso, Nelder e Wedderburn (1972) apresentaram os Modelos Lineares Generalizados, que é uma abordagem unificada para todos os modelos mencionados anteriormente. Portanto, o modelo linear normal é apenas um caso particular dos MLG's.

Os MLG's têm desempenhado um papel importante na análise estatística. O seu uso está sendo estimulado devido às aplicações de modelos logísticos e log-lineares em ciências sociais e medicina, e também, em modelos para dados de sobrevivência. O número de publicações sobre o assunto aumentou muito. No entanto, o método possui algumas limitações como: manutenção da estrutura linear, restrição a variáveis com distribuição da família exponencial e pressupõe independência dos resíduos.

3.1. Família Exponencial

Em modelos de regressão, sempre estamos interessados em situações que há uma variável resposta *Y* e um vetor $\mathbf{x}=(x_1,...,x_k)^T$ de *k* variáveis explicativas. A variável *Y* e as covariáveis podem ser contínua, discreta ou dicotômica.

Assumindo que temos *n* unidades experimentais, os dados podem ser representados na forma matricial.

$$\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X}=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

Sabemos que os modelos lineares generalizados pressupõem que variável resposta tenha uma distribuição pertencente à família exponencial. Seja Y uma variável aleatória, cuja função de densidade de probabilidade se pode escrever na forma.

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

Na equação acima θ é o parâmetro de localização e ϕ é o parâmetro de dispersão, *a(.) ,b(.) e c(.)* são funções reais conhecidas. Normalmente, *a(*ϕ*)* é da forma $\phi\,\omega$, ω é

uma constante conhecida, *b(.)*deve ser diferenciável. Além disso, existem as seguintes propriedades.

$$\text{(I)}\ \mu = E(y) = \frac{db(\theta_i)}{d\theta_i}$$

$$\text{(II)}\quad V(y) = \frac{d^2 b(\theta_i)}{d\theta_i^{\,2}} a(\phi)$$

$$V(y) = \frac{d\mu}{d\theta_i} a(\phi)$$

$$\text{(III)}\ \text{var}(\mu) = \frac{V(y)}{a(\phi)} = \frac{d\mu}{d\phi_i}$$

As distribuições binomial, poisson, normal são membros desta família. Portanto, a seguir, representamos cada uma no formato de modelo linear generalizado.

(a) Distribuição normal

$$f(y_i, \theta_i, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y-\mu))^2$$

Portanto, temos

$$\theta_i = \mu \ \ ;\ b(\theta_i) = \frac{\mu^2}{2}\ ;\ a(\phi) = \sigma^2$$

$$h(y_i, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$$

$$E(y) = \frac{db(\theta_i)}{d\theta_i} = \mu\ ,\ V(y) = \frac{d^2 b(\theta_i)}{d\theta_i^{\,2}} a(\phi) = \sigma^2$$

(b) Distribuição binomial

$$f(y_i, \theta_i, \phi) = \binom{n}{y} \pi^y (1-y)$$

$$= \exp\left\{ y \ln\left[\frac{\pi}{1-\pi}\right] + n\ln(1-\pi) + \ln\binom{n}{y} \right\}$$

$$\theta_i = \ln\left[\frac{\pi}{1-\pi}\right];\ \ \pi = \frac{\exp(\theta_i)}{1+\exp(\theta_i)};$$

$$b(\theta_i) = n\ln(1-\pi);\ a(\phi) = 1;\ h(y_i, \phi) = \ln\binom{n}{y}$$

$$E(y) = \frac{db(\theta_i)}{d\theta_i} = \frac{db(\theta_i)}{d\pi}\frac{d\pi}{d\theta_i}$$

$$\frac{d\pi}{d\theta_i} = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} - \left[\frac{\exp(\theta_i)}{1+\exp(\theta_i)}\right]$$

$$= \pi(1-\pi)$$

$$E(y) = \left(\frac{n}{1-\pi}\right)\pi(1-\pi)$$
$$= n\pi$$
$$V(y) = \frac{dE(y)}{d\theta_i} = \frac{dE(y)}{d\pi}\frac{d\pi}{d\theta_i}$$
$$= n\pi(1-\pi)$$

(c) Distribuição Poisson

$$f(y_i, \theta_i, \phi) = \frac{\lambda^y e^{-\lambda}}{y!}$$
$$= \exp[\, y \ln \lambda - \lambda - \ln(\, y!)]$$
$$\theta_i = \ln(\lambda)\text{, então } \lambda = e^{\theta_i}$$
$$b(\theta_i) = \lambda\,;\ a(\phi) = 1;\ h(y_i, \phi) = -\ln(\, y!)$$
$$E(y) = \frac{db(\theta_i)}{d\theta_i} = \frac{db(\theta_i)}{d\lambda}\frac{d\lambda}{d\theta_i}$$
$$d\lambda = \exp(\theta_i) = \lambda$$
$$E(y) = 1.\lambda = \lambda$$
$$V(y) = \frac{dE(y)}{d\theta_i} = \lambda$$

No modelo de regressão linear clássico, temos que **Y=Xβ+ε**, sendo **β** o vetor de parâmetros e **ε** o vetor de erros aleatórios que se supõe possuir distribuição normal com média zero e variância $\boldsymbol{\sigma}^2$**.** Consequentemente, E(***Y***)=$\boldsymbol{\mu}$ com $\boldsymbol{\mu}$=***X*β,** ou seja o valor esperado da variável resposta é uma função linear das covariáveis.

Em MLG's, a estrutura de linearidade se mantém, mas a função que relaciona o valor esperado e o vetor de covariáveis pode ser qualquer função diferenciável. O valor esperado μ_i se relaciona com o preditor linear $\eta_i = x_i^{'}\beta$ da seguinte forma.

$$\mu_i = h(\eta_i) = h(x_i^{'}\beta),\ \eta_i = g(\mu_i)$$

A função *h* é monótona e diferenciável, $g = h^{-1}$ é a função de ligação. A escolha da função de ligação depende da resposta e do estudo que se deseja realizar. Podemos investigar várias funções por meio de programas computacionais avançados. No entanto, dependendo da distribuição considerada, teremos uma função de ligação natural. Esta afirmação considera o parâmetro canônico (ou de localização natural) igual ao preditor linear, ocorre o que chamamos de função de ligação canônica. Consequentemente, temos $\theta_i = x_i^{'}\beta$.

Tabela 3.1- Distribuições e suas respectivas funções de ligação canônica.

Distribuição	Função de ligação canônica
Normal	$\eta_i = \mu_i$ (identidade)
Binomial	$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ (logit)
Poisson	$\eta_i = \ln(\lambda_i)$ (logarítmica)

3.2. Modelos para dados binários ou na forma de proporções

Suponha que temos *n* variáveis resposta binárias e independentes. Portanto, $Y_i \sim B(1, \pi_i)$.

$$f(y_i \mid \pi_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

Neste caso, a função de ligação canônica é a função *logit*. Dessa forma a probabilidade de sucesso, ou seja, $P(y_i = 1) = \pi_i$ está relacionada com o vetor de covariáveis $x_i^{'}$ através de

$$\pi_i = \frac{\exp(x_i^{'}\beta)}{1+\exp(x_i^{'}\beta)}$$

Pode se perceber que a função $F : \mathbb{R} \to [0,1]$, é a função de distribuição logística. Por este motivo, o modelo linear generalizado definido por uma variável binomial com função de ligação canônica é conhecido como modelo de regressão logística.

Repare que devido ao fato de $\pi_i \epsilon [0,1]$, a princípio, não só a função de distribuição logística como qualquer outra função de distribuição pode ser candidata a função inversa da função de ligação. Por exemplo, podemos supor que a relação entre a probabilidade de sucesso e o vetor de covariáveis é da forma

$$\pi_i = \Phi(\eta_i) = \Phi(x_i^{'}\beta)$$

A função $\Phi(.)$ é a função de distribuição de uma variável aleatória *N(0,1)*. Assim, temos uma função de ligação *probit*.

$$g(\eta_i) = \Phi^{-1}(\eta_i)$$

Outra função de distribuição que também é muito utilizada como inversa da função de ligação é a distribuição de Gumbel. Dessa, obtemos como função de ligação a função *complementar log-log*.

$$F(x) = 1 - exp(-exp(x)), x \epsilon \mathbb{R}$$

A escolha da função de ligação depende da situação. Como as funções *logit* e *probit* não se afastam uma da outra após um ajuste adequado, as duas possuem funcionalidades bem semelhantes. Porém, a função complementar *log-log* pode se comportar de modo diferente, pois tem um crescimento mais abrupto.

3.3. Modelo para dados de contagens

Situações em que a variável resposta é representada por dados de contagem é muito comum na prática. Exemplos disso são: número de acidentes, número de chamadas telefônicas, o número de elementos numa fila de espera, entre outros. O modelo Poisson, como se sabe, desempenha um papel fundamental na análise deste tipo de dados, e tendo uma particularidade, seu valor médio é igual à sua variância.

Nesse caso, a função logarítmica é a função de ligação que geralmente se utiliza. Considerando que a resposta é representada por Y_i tem distribuição Poisson com média μ_i.

$$f(y_i) = \frac{e^{-\mu_i}\mu^{y_i}}{y_i!}, y_i = 0,1,...$$

$$\ln(\mu_i) = x_i^{'}\beta$$

3.4. Metodologia dos Modelos Lineares Generalizados

Há três etapas essenciais que devemos seguir ao tentar modelar dados por meio de um MLG: a formulação, o ajuste e a validação e seleção do modelo.

A formulação do modelo é o momento em que é feita a escolha da distribuição para variável resposta. Para isso, é necessário examinar os dados cuidadosamente. Por exemplo, as distribuições gama e normal inversa são apropriadas para modelar dados de natureza contínua e que possuam assimetrias.

Além disso, faremos a escolha das covariáveis e formulação da matriz de especificação, portanto precisamos entrar em contato com o problema a ser estudado e ter atenção, principalmente, a codificação para variáveis qualitativas.

Devemos escolher uma função de ligação que produza as propriedades estatísticas desejadas para o modelo, mas, apenas a conveniência matemática não determina a escolha da função de ligação.

Na fase do ajuste do modelo, estimamos os coeficientes β's associados às covariáveis. Além disso, obtêm-se intervalos de confiança e teste de bondade de ajuste são realizados.

A fase de seleção e validação do modelo tem por objetivo encontrar modelos com um número moderado de parâmetros e ainda assim apresente adequação aos dados. Nesta etapa, também detectamos se existe discrepância entre os dados e os valores previstos e investigamos a existência de *outliers* e observações influentes. Na seleção do melhor modelo para explicar o problema levamos em consideração três fatores: adequabilidade, parcimônia e interpretação.

3.4.1. Inferência

No modelo de regressão linear, quando os erros são normalmente distribuídos e independentes, teste estatísticos e intervalos de confiança são baseados nas distribuições *t* e *F*, os parâmetros do modelo são estimados através do método de mínimos quadrados. Da mesma forma que o modelo linear normal, a inferência em modelos lineares generalizados é, essencialmente, baseada na verossimilhança. Com isso, o método de máxima verossimilhança não é apenas o método para encontrar as estimativas dos parâmetros, mas também, é utilizado para a construção dos testes de hipótese e estatísticas de qualidade do ajuste. O método de máxima verossimilhança consiste em escolher o vetor $\boldsymbol{\beta}$ que torne máxima a função de log-verossimilhança.

$$\frac{\partial \ln(L)}{\partial \boldsymbol{\beta}} = 0$$

No entanto, em um MLG, a maximização da função de verossimilhança nos conduz a um sistema não linear. Para sua resolução, há um algoritmo numérico, que é uma

modificação do método de Newton-Raphson, conhecido por vários nomes, porém, o chamaremos de Método de Mínimos Quadrados Reponderados (IRLS).

Considerando que o logaritmo da função de verossimilhança é

$$l(y,\beta) = \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i)] / a(\phi) + c(y_i,\phi)$$

Usando uma função de ligação canônica, temos

$$\eta_i = \theta_i = g[E(y_i)] = g(\mu_i) = x_i'\beta$$

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta}$$

$$= \frac{1}{a(\phi)} \sum_{i=1}^{n} \left[y_i - \frac{db(\theta_i)}{d\theta_i} \right] x_i$$

$$= \frac{1}{a(\phi)} \sum_{i=1}^{n} [y_i - \mu_i] x_i$$

Como $a(\phi)$ é constante, $\sum_{i=1}^{n} [y_i - \mu_i] x_i = 0$

Este é um sistema de p equações, uma para cada parâmetro do modelo. Estas equações são chamadas de equação escore de máxima verossimilhança. Em notação matricial, escrevemos as equações são escritas como

$$\mathbf{X'(y - \mu) = 0}$$

Sendo $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_k)$ o valor final que o algoritmo produz como estimativa de β e se as suposições do modelo estão corretas, inclusive a escolha da função de ligação, é possível mostrar que assintoticamente

$$E(\hat{\beta}) = \beta$$

O apêndice B contém maiores detalhes sobre como resolver as equações escore para os casos em que a função de ligação é canônica.

3.4.2. Propriedades do Estimador de Máxima Verossimilhança

Geralmente, os estimadores de máxima verossimilhança possuem melhores propriedades estatísticas que estimadores de mínimos quadrados. Isso ocorre porque, através da máxima verossimilhança, a estimativa exige que as observações sejam normalmente distribuídas. Estimadores de máxima verossimilhança são não viciados ou assintoticamente não viciados.

A segunda derivada da função log-verossimilhança é chamada de matriz hessiana, que possui dimensões (p x p). A negativa da matriz hessiana é a matriz de informação de Fisher. Uma outra definição para a matriz da informação de Fisher é a variância do escore. A matriz **V** = diag { σ_i^2 } , onde, σ_i^2 é função de μ_i .

$$G_{ij} = \frac{\partial L(\hat{\boldsymbol{\beta}})}{\partial \beta_i \partial \beta_j}$$

$$-G(\hat{\boldsymbol{\beta}}) = I(\hat{\boldsymbol{\beta}})$$

$$I(\hat{\boldsymbol{\beta}}) = Var\left\{ \frac{1}{a(\phi)} [\mathbf{X}'(\mathbf{y} - \boldsymbol{\mu})] \right\} = \frac{\mathbf{X}'\mathbf{V}\mathbf{X}}{[a(\phi)]^2}$$

$$Var(\hat{\boldsymbol{\beta}}) = I^{-1}(\hat{\boldsymbol{\beta}}) = [\mathbf{X}'\mathbf{V}\mathbf{X}]^{-1}[a(\phi)]^2$$

Considerando o caso da distribuição normal, temos que

$$\sigma_i^2 = \sigma^2; a(\phi) = \sigma$$

$$Var(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

A diferença existente entre este algoritmo e o algoritmo de Newton-Raphson para resolver sistemas de equações não lineares, reside na utilização da matriz de informação de Fisher ao invés da matriz hessiana. A vantagem desta substituição deve-se ao fato de, em geral, a matriz de informação é mais fácil de ser calculada, além do fato desta ser sempre definida positiva.

As iterações param quando é atingido um determinado critério. Geralmente, a convergência é alcançada após poucas iterações. Quando o procedimento parece não convergir, isto pode ser devido a uma má estimativa inicial, ou até mesmo, a não existência do estimador de máxima verossimilhança dentro da região de valores admissíveis para **β**.

Para calcular a iteração de ordem zero, que inicia o processo, pode-se calcular a estimativa de mínimos quadrados não ponderados. Para que o algoritmo se processe sem problemas, é necessário que a matriz de informação tenha inversa em cada iteração.

Um problema importante é sobre a existência e unicidade dos estimadores de máxima verossimilhança. Inicialmente, não há garantia que a função de verossimilhança tenha um único máximo, ou mesmo que tenha um máximo. Outro aspecto interessante é saber se a verossimilhança tem um máximo na fronteira do espaço admissível para o parâmetro **β**, pois tal máximo pode levar a problemas de natureza computacional.

3.4.3. Testes de Hipótese e Intervalo de Confiança das Estimativas dos Parâmetros.

Para grandes amostras, a distribuição do estimador de máxima verossimilhança é aproximadamente normal. Além disso, as variâncias e covariâncias do conjunto destes estimadores podem ser encontradas a partir da segunda derivada parcial da função log-verossimilhança em relação aos parâmetros do modelo, a matriz hessiana.

Para testar estas hipóteses, podemos construir a estatística de Wald, que compara o estimador de máxima verossimilhança do parâmetro com o valor proposto, supondo que a diferença entre os dois é normalmente distribuída.

Se os elementos da matriz hessiana são calculados considerando o estimador de máxima verossimilhança, $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, a matriz de covariância aproximada dos coeficientes de regressão é

$$\hat{Var}(\hat{\boldsymbol{\beta}}) = -G(\hat{\boldsymbol{\beta}})^{-1} = (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}$$

A raiz quadrada dos elementos da diagonal principal da matriz de covariância estimada de $\hat{\boldsymbol{\beta}}$ é um estimador do erro padrão dos coeficientes de regressão. Dessa forma, a estatística para testar a hipótese de nulidade do parâmetro é

$$H_0 : \beta_j = 0,\ H_1 : \beta_j \neq 0$$

$$\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)} \sim N(0,1)$$

Observe que $\hat{se}(\hat{\beta}_j)$ é o erro padrão estimado para $\hat{\beta}_j$. Uma alternativa é considerar que o quadrado desta estatística possui uma distribuição qui-quadrado com um grau de liberdade.

Baseado neste último resultado, podemos obter um intervalo de confiança para cada coeficiente estimado. Um intervalo de confiança para $\hat{\beta}_j$ de aproximadamente 100(1-α) é

$$\hat{\beta}_j - z_{\alpha/2}\hat{se}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j - z_{\alpha/2}\hat{se}(\hat{\beta}_j)$$

3.4.5. Qualidade do ajuste

Quando se trabalha com muitas covariáveis, existe o interesse de obter um modelo com o menor número de variáveis explicativas possível, que forneça uma boa interpretação do problema em questão, e se ajuste bem aos dados. A seleção do modelo consiste em procurar o melhor modelo. Para tal, modelo terá de atingir o equilíbrio entre três fatores: ajuste, parcimônia e interpretação. Durante o processo de seleção, devemos considerar algumas espécies de modelo. Um modelo completo é um modelo que se ajusta completamente aos dados, isto é, para cada observação tem-se um parâmetro. No entanto, em algumas situações, usar um modelo mais simples pode ser mais fácil de ser compreendido e o seu ajuste pode ser tão bom quanto do modelo completo, esse seria o modelo reduzido.

O teste de razão de verossimilhança pode ser usado para comparar o modelo completo com um modelo reduzido. O teste compara o logaritmo do valor das funções de verossimilhanças do modelo completo e do modelo reduzido, a estatística resultante deste é a deviance.

Para grandes amostras, quando o modelo reduzido é correto, a estatística do teste segue uma distribuição qui-quadrado com graus de liberdade igual á diferença entre o numero de parâmetros entre o modelo completo e o modelo reduzido. Portanto, considerando o nível de significância *α*, se a estatística do teste exceder o quantil *(1-α)* da distribuição qui-quadrado, a hipótese de que o modelo reduzido é adequado é rejeitada

$$D(\boldsymbol{\beta}) = 2[\ln L(MC) - \ln L(MR)]$$
$$\cdot = 2\ln \frac{L(MC)}{L(MR)}$$

A deviance pode ser utilizada para testar subconjuntos de parâmetros do modelo. A estatística funciona de forma análoga a diferença de soma de quadrados em modelos regressão linear. Por isso, podemos reescrever o preditor linear como:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$
$$\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$$

O modelo completo contém p parâmetros, $\boldsymbol{\beta}_1$ têm *p-r* desses parâmetros, e os *r* parâmetros restantes estão em $\boldsymbol{\beta}_2$. As matrizes $\mathbf{X}_1$ e $\mathbf{X}_2$ contém todas as variáveis associadas a esses parâmetros. Portanto, o preditor linear para o modelo reduzido é

$$\boldsymbol{\eta} = \mathbf{X}_1\boldsymbol{\beta}_1$$

Assumimos que o modelo reduzido é ajustado e sua deviance é $D(\boldsymbol{\beta}_1)$, para o modelo completo, temos $D(\boldsymbol{\beta})$. A deviance para o modelo reduzido nunca é menor que a deviance para o modelo completo, porque contém menos parâmetros. No entanto, se a deviance do modelo reduzido é muito maior que a deviance para o modelo reduzido,

podemos rejeitar a hipótese de que todos os parâmetros em $\boldsymbol{\beta}_2$ são iguais a zero. Formalmente, a diferença de deviance e as hipóteses do teste são

$$D(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) = D(\boldsymbol{\beta}_1) - D(\boldsymbol{\beta})$$

Consequentemente, temos que os critérios de decisão para o teste são: se $D(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) \geq \chi^2_{\alpha,r}$, rejeitamos a hipótese nula, se $D(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) < \chi^2_{\alpha,r}$, não rejeitamos a hipótese nula.

Uma medida alternativa de adequação é a estatística de Pearson, X^2. Esta estatística testa a adequabilidade de um modelo comparando o valor calculado com o quantil de probabilidade 1-α de uma distribuição χ^2 com *n-p* graus de liberdade.

$$X^2 = \sum_{i=1}^{n} \frac{(\theta_i - e_i)^2}{e_i}$$

Os valores observados para variável resposta são representados por θ_i. Os valores estimados para reposta são os e_i's.

3.4.6. Intervalo de Confiança para Resposta média

De modo semelhante ao modelo linear normal, pode-se construir um intervalo de confiança para o preditor linear. No entanto, em MLG's, não há igualdade entre a resposta média estimada e o preditor linear. Sendo $\mathbf{x}_0' = [1, \mathbf{x_{01}}, \mathbf{x_{02}}, ..., \mathbf{x_{0k}}]$ um conjunto de valores para as variáveis explicativas, então o preditor linear em $\mathbf{x}_0$ é $\mathbf{x}_0'\boldsymbol{\beta}$. Para encontrar a variância estimada do preditor linear neste ponto, usaremos um resultado já demonstrado, lembre-se que a matriz da covariância aproximada para a estimativa dos parâmetros do modelo é $\hat{Var}(\hat{\boldsymbol{\beta}}) = -G(\hat{\boldsymbol{\beta}})^{-1} = (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}$. Dessa forma, a variância estimada do preditor linear é $\hat{Var}(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'\hat{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0 = \mathbf{x}_0'(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{x}_0$. Assim, o intervalo de confiança com 100(1-α) % é

$$L(\mathbf{x}_0) \leq \mathbf{x}_0'\boldsymbol{\beta} \leq U(\mathbf{x}_0)$$

$U(\mathbf{x}_0) = \mathbf{x}_0'\boldsymbol{\beta} + z_{\alpha/2}\sqrt{\mathbf{x}_0'(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{x}_0}$ e $L(\mathbf{x}_0) = \mathbf{x}_0'\boldsymbol{\beta} - z_{\alpha/2}\sqrt{\mathbf{x}_0'(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{x}_0}$.

Considerando a função de ligação g é a resposta média estimada $\hat{\boldsymbol{\mu}}_0$ em $\mathbf{x}_0$, $\hat{\boldsymbol{\mu}}_0 = g^{-1}(\mathbf{x}_0'\hat{\boldsymbol{\beta}})$. O intervalo de confiança para a resposta média estimada é $g^{-1}(L(\mathbf{x}_0)) \leq \hat{\boldsymbol{\mu}}_0 \leq g^{-1}(U(\mathbf{x}_0))$.

3.4.7. Análise de Resíduos

A análise de resíduos é utilizada, não apenas para avaliar a qualidade de ajuste do modelo no que se refere à escolha da distribuição, da função de ligação, como também para ajudar a identificar observações mal ajustadas não identificadas pelo modelo.

Um resíduo deve mostrar a discrepância entre o valor observado e o valor ajustado pelo modelo. Na regressão linear, o resíduo, usado também para detectar violações a suposições de não homogeneidade da variância, pode ser representado por $y_i - \hat{\mu}_i$. No entanto, em MLG's, esta representação não é apropriada porque a variância da variável resposta não é constante. Por isso, neste caso, é mais conveniente a utilização de dois tipos de resíduos: resíduos de Pearson e resíduos deviance.

O resíduo de Pearson corresponde a cada contribuição de cada observação para a estatística de Pearson generalizada. A desvantagem deste tipo de resíduo é que, geralmente, sua distribuição é bastante assimétrica para modelos não normais.

$$r_P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}ar(y_i)}}$$

A deviance também pode ser escrita da seguinte forma $D(\boldsymbol{\beta}) = \sum_{i=1}^{n} d_i$. Dessa forma, os resíduos deviance são representados por cada componente d_i. Os resíduos deviance têm a propriedade de considerar o sinal da diferença entre o valor observado da resposta e a resposta média estimada, além de a soma dos quadrados ser a própria deviance.

$$d_{i,r} = [\text{sgn}(y_i - \hat{\mu}_i)]\sqrt{d_i}\ ,\ i = 1,2,...,n$$

$$\sum_{i=1}^{n} d_{i,r}^2 = D(\boldsymbol{\beta})$$

Uma questão interessante é saber qual tipo de resíduos é mais adequado para a validação do modelo. Pierce e Schafer (1986), em um estudo sobre os resíduos em modelos baseados em distribuições da família exponencial, sugerem que os resíduos deviance são muito próximos daqueles gerados pela melhor transformação de normalização possível. Com isso, é recomendável usar os resíduos deviance na construção dos gráficos de diagnósticos.

Segundo McCullagh e Nelder (1989), devemos construir os gráficos de resíduos deviance x valores ajustados transformados, a transformação varia de acordo com a distribuição da resposta. A tabela a seguir detalha cada transformação. Além disso, é recomendável construir o gráfico de probabilidade normal dos resíduos deviance. A interpretação dos resultados é análoga ao modelo linear normal.

Tabela 3.2 – Distribuição e suas respectivas transformações.

Distribuição	Transformação
Normal	$\hat{\mu}$
Binomial	$2sen^{-1}\sqrt{\hat{\mu}}$
Poisson	$2\sqrt{\hat{\mu}}$
Gama	$2\log\hat{\mu}$

CAPÍTULO 4 - Aplicações

Neste capítulo, apresentaremos exemplos em que os MLG's são adequados no ambiente do Planejamento de Experimentos, mais especificamente, em planejamentos fatoriais 2^k.

Nos exemplos selecionados, consideraremos três situações: (i) os dados seguem uma distribuição normal (apesar de sabermos que essa hipótese não é verdadeira), (ii) verificando que os dados não são normais usaremos uma transformação que normaliza os dados e, finalmente, (iii) os dados seguem uma distribuição da família exponencial diferente da normal. No final, comparamos os resultados dos ajustes das três situações. Os cálculos e gráficos foram feitos com auxilio do programa R.

4.1 "Simulação de dados binomiais''

Neste exemplo, optamos por trabalhar com um conjunto de dados simulados. Trata-se de um planejamento fatorial 2^4, detalhado na tabela abaixo, foi obtido através de uma simulação, na qual, fixamos os coeficientes e consideramos uma regressão logística. A simulação dos dados foi realizada com base na seguinte equação $p = \frac{e^{-1+1,5A-1,4B-1,2C-1,1AC+0,9AD+0,7BD}}{1+e^{-1+1,5A-1,4B-1,2C-1,1AC+0,9AD+0,7BD}}$, de tal modo que os efeitos significativos são: A, B, C, AC, AD e BD.

Na tabela a seguir apresentamos os valores das probabilidades em função das combinações de níveis dos fatores arredondadas para duas casas decimais. A simulação das respostas foi feita usando-se a função *rbinom* do *R* e considerando-se 100 observações para cada combinação dos níveis dos fatores.

Tabela 4.1 – Simulação dos dados binomiais

Matriz do plano fatorial						
Observação	A	B	C	D	*p*	Resposta
1	-1	-1	-1	-1	0,65	0.61
2	1	-1	-1	-1	0,98	0.98
3	-1	1	-1	-1	0,03	0.02
4	1	1	-1	-1	0,45	0.46
5	-1	-1	1	-1	0,6	0.57
6	1	-1	1	-1	0,35	0.33
7	-1	1	1	-1	0,02	0.01
8	1	1	1	-1	0,01	0.02
9	-1	-1	-1	1	0,07	0.09
10	1	-1	-1	1	0,99	0.98
11	-1	1	-1	1	0,02	0.04
12	1	1	-1	1	0,95	0.92
13	-1	-1	1	1	0,06	0.09
14	1	-1	1	1	0,45	0.5
15	-1	1	1	1	0,01	0.01
16	1	1	1	1	0,17	0.18

4.1.1. Modelo Linear Normal

Sabemos que a resposta segue uma distribuição binomial, mesmo assim, a tentativa de ajuste utilizando o método de mínimos quadrados é interessante, porque, na prática, esse método é muito usado, até mesmo quando não é o mais apropriado. Além disso, queremos abordar algumas limitações que este possui em relação ao MLG's.

Como o experimento não é replicado, usamos o gráfico de probabilidade normal dos efeitos para detectar quais fatores são significantes para o modelo.

Podemos perceber na figura 4.1 que os pontos mais distantes da reta são: A, B, C, 1, 3 e 4. As interações AC, AD e BD estão representadas, respectivamente por 1,3 e 4. Dessa forma, temos que os efeitos considerados como significantes são: A, B, C, AC, AD e BD. Os efeitos restantes serão descartados do modelo.

Dessa forma, o ajuste tem o seguinte formato:

$$\hat{p} = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 B + \hat{\beta}_3 C + \hat{\beta}_{13} AC + \hat{\beta}_{14} AD + \hat{\beta}_{24} BD$$

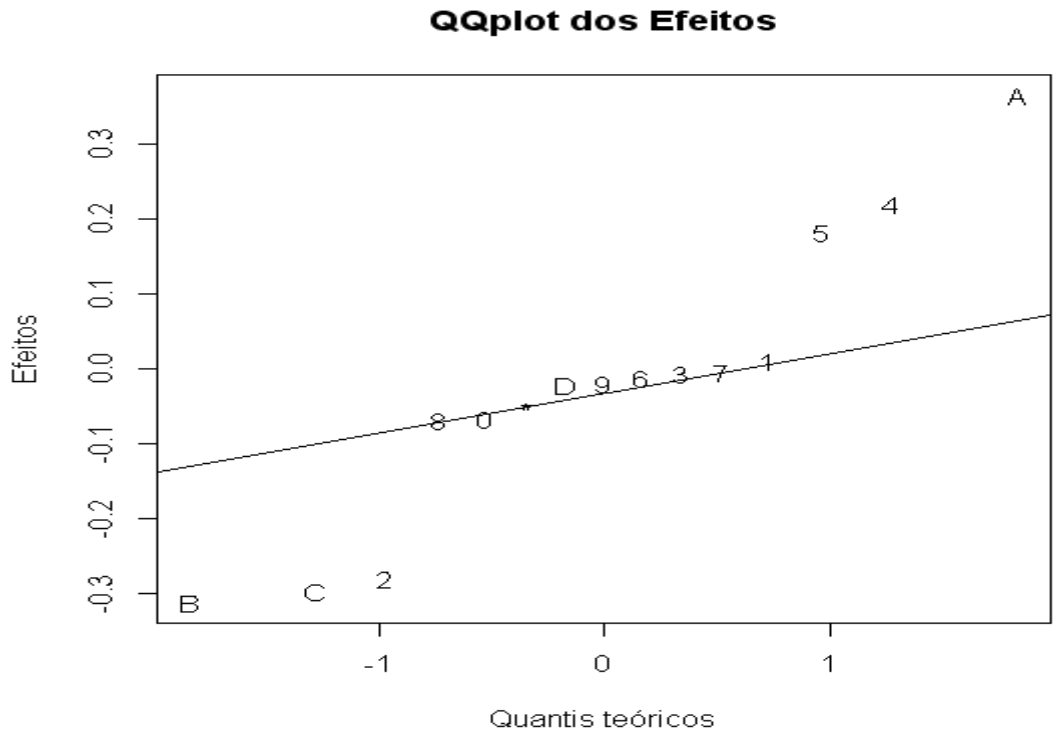

Figura 4.1 - Gráfico de Probabilidade Normal dos Efeitos

Através das tabelas 4.2 e 4.3, percebe-se que todos os efeitos selecionados pelo gráfico de probabilidade normal são realmente significantes. O p-valor da estatística F para o ajuste é muito pequeno, o que nos leva a hipótese nula de que todos os coeficientes são nulos. O R^2 Múltiplo e o R^2 Ajustado estão próximos de 1, indicando que o modelo linear é adequado. Para confirmação, é necessário analisar os resíduos.

Tabela 4.2 -Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	0,36375	0,01915	18,996	1,43E-08
A	0,18250	0,01915	9,531	5,33E-06
B	-0,15625	0,01915	-8,160	1,89E-05
C	-0,14875	0,01915	-7,768	2,80E-05
AC	-0,14000	0,01915	-7,311	4,51E-05
AD	0,11000	0,01915	5,745	0,000278
BD	0,09125	0,01915	4,765	0,001022

Tabela 4.3 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
A	1	0,53290	0,53290	90,835	5,33E-06
B	1	0,39062	0,39062	66,584	1,89E-05
C	1	0,35402	0,35402	60,345	2,80E-05
AC	1	0,31360	0,31360	53,455	4,51E-05
AD	1	0,19360	0,19360	33,000	0,00028
BD	1	0,13322	0,13322	22,709	0,00102
Resíduos	9	0,05280	0,00587		

R^2 Múltiplo: 0,9732, R^2 Ajustado: 0,9553
Estatística F: 54,49 com 6 e 9 graus de liberdade, p-valor: 1,4e-06

Análise de Resíduos

Na figura 4.3, a maioria dos pontos está localizada bem próxima à reta. Portanto, neste gráfico, visualmente, não há indícios para rejeitarmos a hipótese de normalidade, para confirmação realizamos o teste de Anderson-Darling, que verifica se um conjunto de dados segue determinada distribuição de probabilidade, neste caso, estamos interessados em testar se os resíduos estão normalmente distribuídos, no apêndice C, há mais detalhes sobre este teste. A estatística do teste, *A*, é igual 0,1797 com um p-valor de 0,8999. Dessa forma, realmente, podemos considerar que os resíduos estão normalmente distribuídos. Na figura 4.2, existe a suspeita de uma tendência similar à função cúbica ou senoidal, porém, devido o baixo número de observações, não podemos afirmar que existem inadequações do modelo.

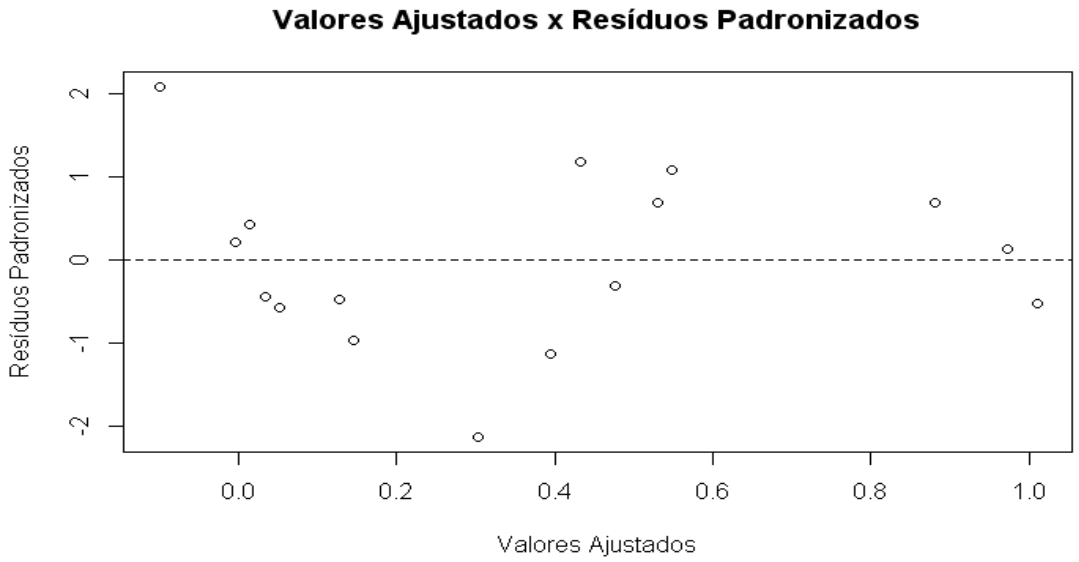

Figura 4.2 - Valores Ajustados x Resíduos Padronizados

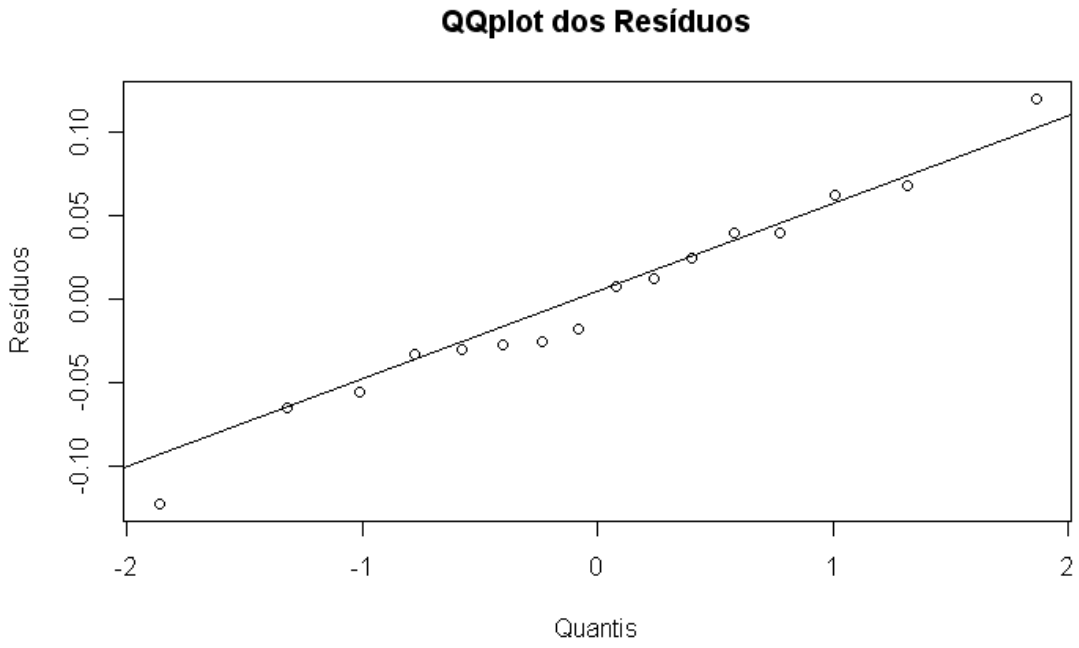

Figura 4.3 - Gráfico de Probabilidade Normal dos Resíduos

4.1.2. Transformação de Box-Cox

Dados que representam proporções geralmente não apresentam uma distribuição simétrica. Para um melhor ajuste do modelo aos dados e para estabilizar a variância, usaremos a transformação de Box-Cox.
A teoria subjacente a esta transformação usa o método de máxima verossimilhança.

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}, \lambda \neq 0 \\ \dot{y} \ln y, \lambda = 0 \end{cases}$$

$\dot{y} = \ln^{-1}\left[(1/n)\sum \ln y\right]$ é a média geométrica das observações. O estimador de máxima verossimilhança de λ é o valor em que o soma do quadrado dos erros é mínimo.

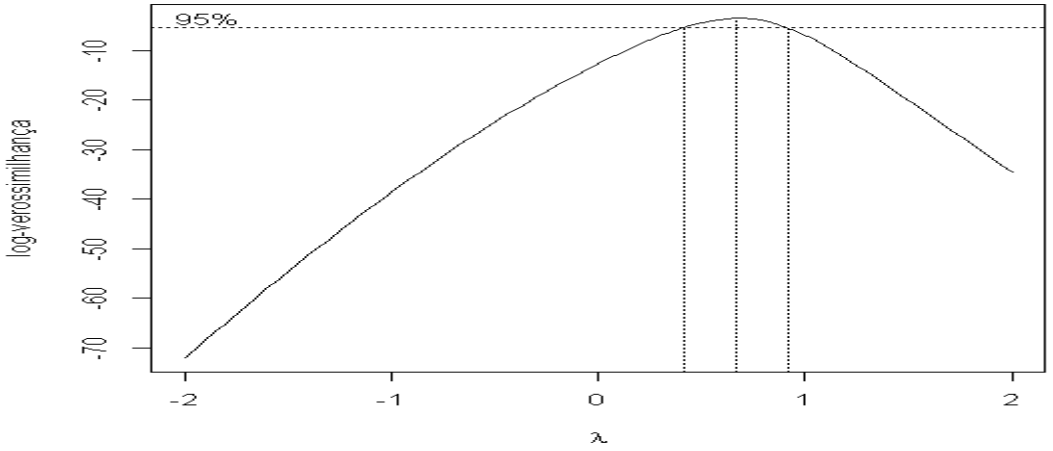

Figura 4.4 - Lambda x Logaritmo da Função de Verossimilhança

O valor de lambda que maximiza a função de verossimilhança é aproximadamente igual a 0,667. Usaremos novamente o gráfico de probabilidade normal dos efeitos para selecionar aqueles que serão significantes para o modelo. Na figura 4.5, observa-se que os efeitos selecionados são os mesmos em relação ao ajuste contendo os dados originais.

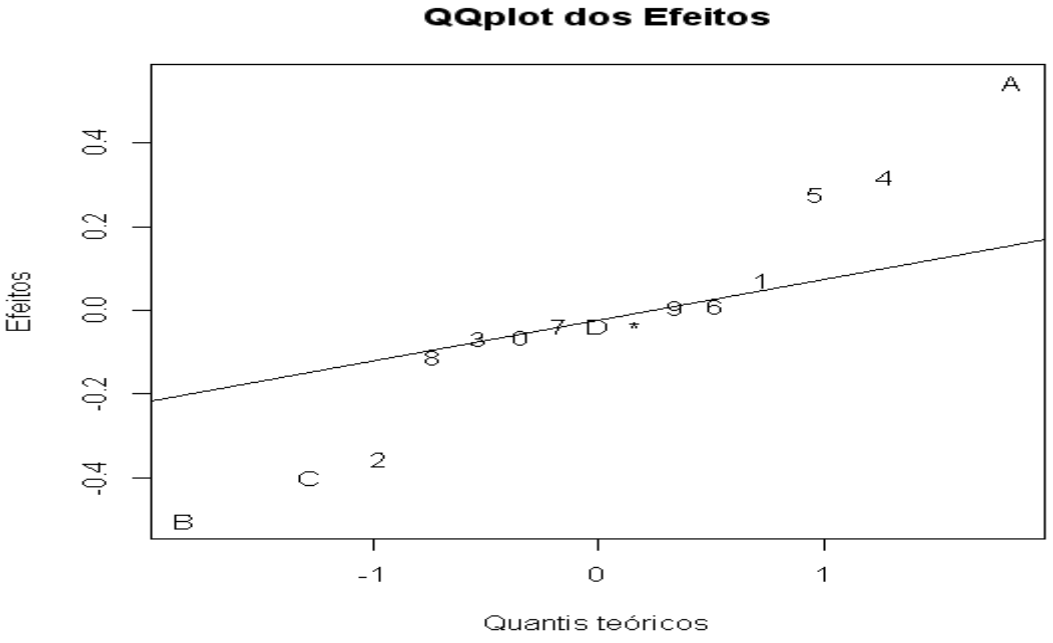

Figura 4.5 – Gráfico de Probabilidade Normal dos Efeitos

Tabela 4.5 - Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	-0,83373	0,02921	-28,538	3,88E-10
A	0,27405	0,02921	9,381	6,08E-06
B	-0,25044	0,02921	-8,572	1,27E-05
C	-0,20107	0,02921	-6,882	7,21E-05
AC	-0,17676	0,02921	-6,051	0,00019
AD	0,16025	0,02921	5,485	0,000387
BD	0,13979	0,02921	4,785	0,000994

R^2 múltiplo: 0,9707, R^2 ajustado:0,9512

Estatística F: 49,75 com 6 e graus de liberdade, p-valor: 2,138e-06

Tabela 4.6 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
A	1	1,20165	1,20165	87,995	6,08E-06
B	1	1,00352	1,00352	73,486	1,27E-05
C	1	0,64684	0,64684	47,367	0,00007207
AC	1	0,49993	0,49993	36,609	0,0001904
AD	1	0,41089	0,41089	30,089	0,0003875
BD	1	0,31265	0,31265	22,895	0,009945
Resíduos	9	0,1229	0,01366		

Nas tabelas 4.5 e 4.6 observa-se que todos os efeitos selecionados inicialmente como significantes permanecem no ajuste após realizar a análise da variância e estimação dos coeficientes. O R^2 múltiplo e o R^2 possuem valores acima de 0,9. O p-valor da estatística F é muito pequeno. Todas estas medidas indicam uma boa adequação do ajuste.

Análise de Resíduos

Na figura 4.6, após a transformação, ainda há uma tendência muito similar a que foi observada no ajuste para os dados originais, mas isso não nos permite concluir que o modelo não é adequado. Na figura 4.7 os pontos não estão distribuídos tão próximos da reta em comparação ao ajuste para os dados originais. Além disso, no teste de Anderson-Darling para os resíduos com uma estatística igual a 0,738 com um p-valor de 0,0432, ou seja, a hipótese de normalidade dos resíduos deve ser rejeitada. Portanto, a transformação da variável resposta não se mostrou eficiente para este caso.

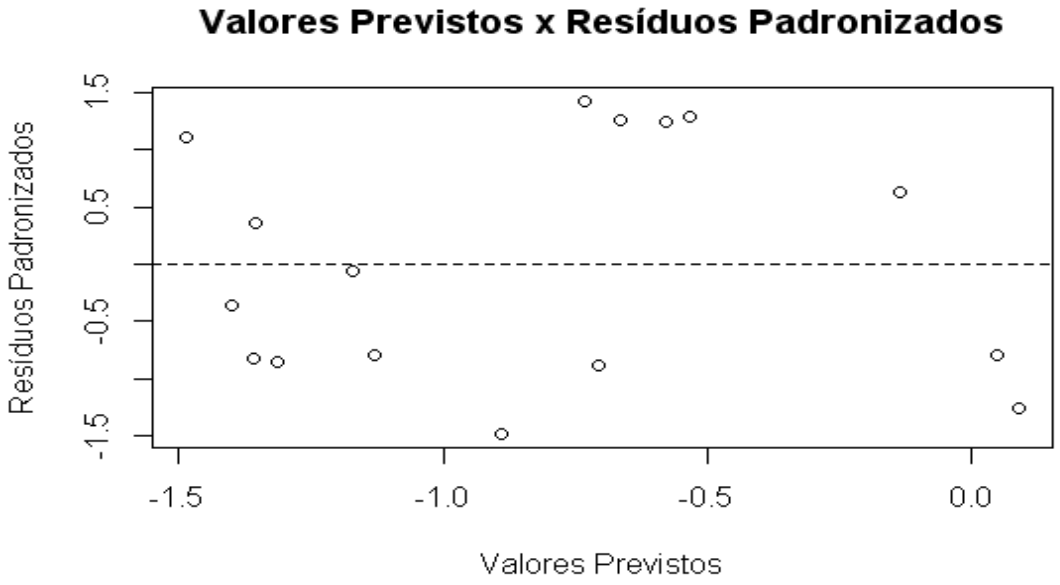

Figura 4.6 - Valores Previstos x Resíduos Padronizados

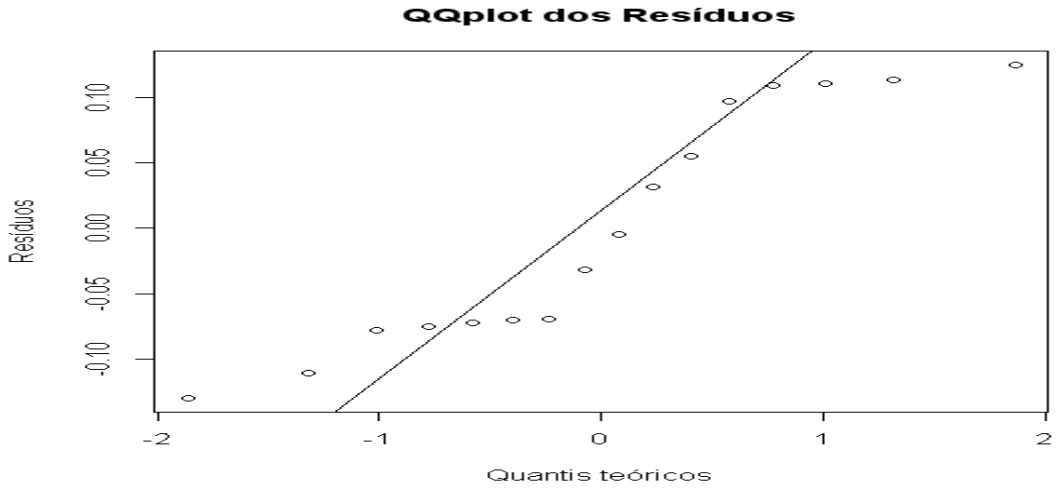

Figura 4.7 - Gráfico de Probabilidade Normal dos Resíduos

4.1.3. Modelos Lineares Generalizados

O modelo será ajustado por meio de uma distribuição binomial com função de ligação logit. Como o experimento foi simulado a partir da regressão logística, espera-se que este ajuste se mostre o mais adequado. O gráfico de probabilidade normal também é utilizado na seleção dos efeitos, porém, ao invés, dos efeitos, usamos diretamente os coeficientes. No método de mínimos quadrados, sabe-se que os coeficientes são iguais a metade dos efeitos, mas não temos conhecimento de qual é esta relação quando estamos trabalhando com modelos lineares generalizados.

No caso de MLG's, ao invés de utilizarmos os efeitos, para selecionar os fatores significantes, usamos os coeficientes estimados. Isso ocorre porque, ao contrário do modelo linear normal, não sabemos qual a relação entre os efeitos e a estimativa dos coeficientes. No entanto, observamos que os coeficientes selecionados como significantes são os mesmos dos dois casos anteriores.

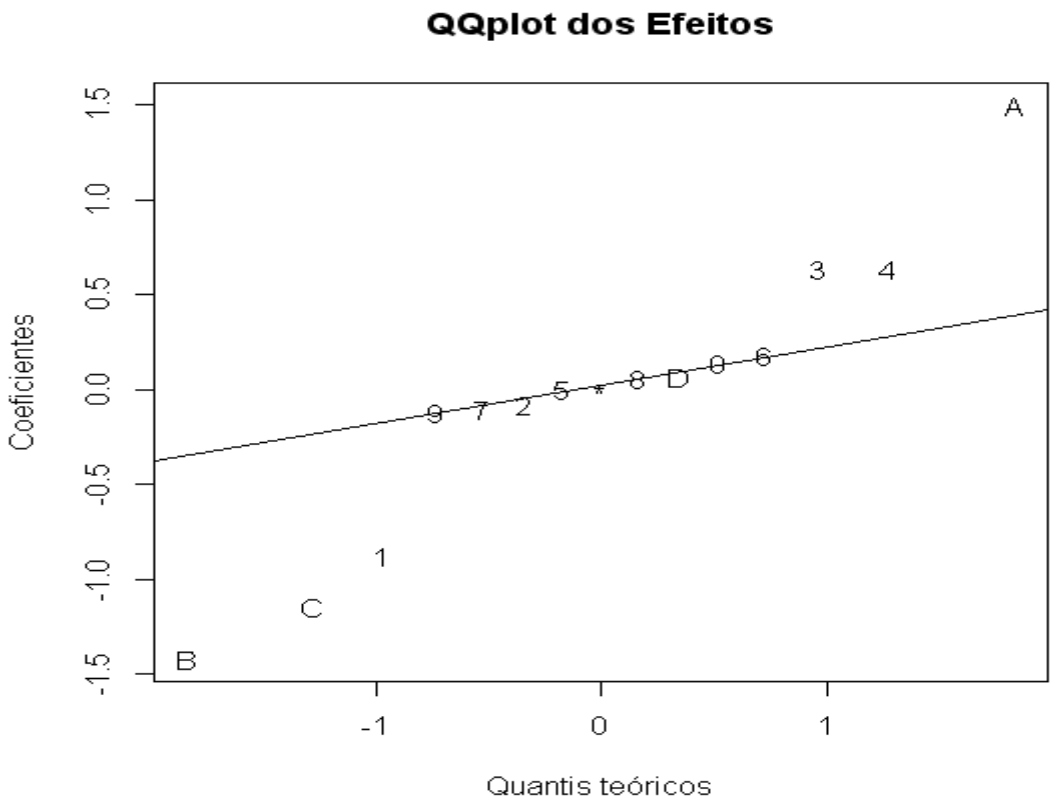

Figura 4.8 – Gráfico de Probabilidade Normal dos Efeitos

As tabelas mostram que todos os fatores selecionados realmente são significantes e as estimativas dos coeficientes do ajuste estão bem próximas aos valores propostos. A diferença de deviance entre o modelo reduzido e modelo completo é igual a 5,75, considerando uma distribuição qui-quadrado com 9 graus de liberdade, temos um p-valor aproximadamente igual a 0,764. Consequentemente, não rejeitamos a hipótese de que o modelo reduzido é adequado. A estatística qui-quadrado de Pearson é igual a 5,62, possuindo um p-valor próximo de 0,777.

Tabela 4.7 - Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor z	Pr(>|z|)
Intercepto	-0,95828	0,08646	-11,084	< 2e-16
A	1,39735	0,09687	14,424	< 2e-16
B	-1,33996	0,10345	-12,953	< 2e-16
C	1,09441	0,08882	-12,321	< 2e-16
AC	-0,98609	0,08687	-11,351	< 2e-16
AD	0,79891	0,08503	9,396	< 2e-16
BD	0,55823	0,08262	6,756	1,42E-11

Tabela 4.8 - Análise da Deviance

Efeito	Grau de Liberdade	Deviance	Grau de Liberdade Restantes	Diferença de Deviance
Nulo			15	1017,29
A	1	240,214	14	777,08
B	1	204,272	13	572,80
C	1	220,412	12	352,39
AC	1	141,844	11	210,55
AD	1	157,437	10	53,11
BD	1	47,364	9	5,75

Análise de Resíduos

Na figura 4.9, visualmente, não é possível identificar nenhuma tendência. Na figura 4.10, os pontos estão dispostos ao longo da reta, por meio do teste de Anderson-Darling, foi obtida a estatística igual a 0,2226 com p-valor de 0,7921, portanto, a hipótese de normalidade dos resíduos deviance não deve ser rejeitada.

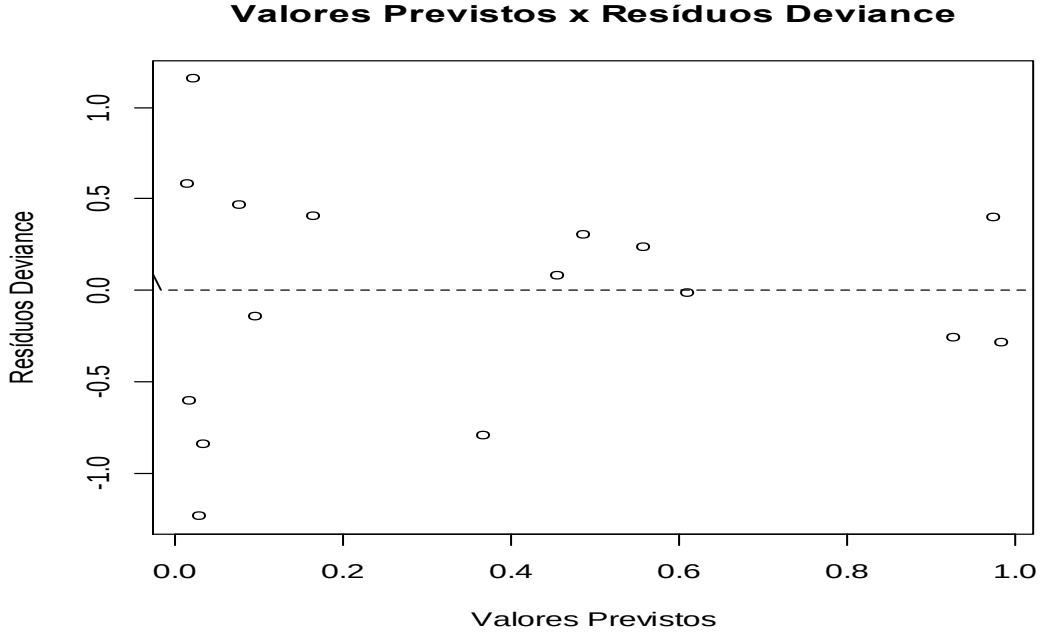

Figura 4.9 - Valores Previstos x Resíduos Deviance

Figura 4.10 – Gráfico de Probabilidade Normal dos Resíduos

4.1.4. Conclusões

O ajuste obtido através MLG apresentou os melhores resultados. Além dos valores pontuais previstos serem os mais próximos da variável resposta, a amplitude dos intervalos de confiança para a resposta é menor. Os resultados para o modelo linear normal e apresenta graves problemas, pois foram obtidos alguns valores previstos negativos e maiores que 1, sendo a variável resposta uma proporção. Na transformação de Box-Cox, também apresenta valores previstos acima de 1 e alguns valores para os limites inferiores não são possíveis de serem calculados. Para comprovar a conclusões utilizamos o erro quadrático médio, que mede a diferença entre os valores estimados e os verdadeiros valores da variável resposta. Observamos que o EQM para o MLG é muito menor do que nos outros dois modelos.

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Tabela 4.7 - Valores previstos e os limites inferiores e superiores do intervalo de confiança para a resposta média.

	Modelo Linear Normal			Transf. Box-Cox			MLG		
Resposta	Prev	LI	LS	Prev	LI	LS	Prev	LI	LS
0,61	0,548	0,433	0,663	0,518	0,384	0,664	0,611	0,525	0,690
0,98	0,972	0,857	1,087	1,049	0,876	1,231	0,974	0,956	0,985
0,02	0,053	-0,062	0,168	0,044	0,001	0,118	0,034	0,020	0,056
0,46	0,477	0,362	0,592	0,366	0,248	0,498	0,456	0,371	0,543
0,57	0,528	0,413	0,643	0,479	0,349	0,622	0,558	0,471	0,641
0,33	0,394	0,279	0,510	0,384	0,264	0,518	0,368	0,287	0,456
0,01	0,033	-0,082	0,148	0,028	Erro	0,095	0,028	0,016	0,047
0,02	-0,101	-0,216	0,015	0,001	Erro	0,044	0,013	0,008	0,021
0,09	0,143	0,028	0,258	0,121	0,046	0,217	0,094	0,063	0,139
0,98	1,009	0,894	1,125	1,090	0,915	1,275	0,984	0,971	0,991
0,04	0,016	-0,100	0,131	0,030	Erro	0,099	0,021	0,012	0,037
0,92	0,882	0,767	0,997	0,871	0,709	1,043	0,927	0,889	0,952
0,09	0,123	0,008	0,238	0,098	0,030	0,189	0,077	0,050	0,117
0,5	0,432	0,317	0,547	0,414	0,291	0,551	0,485	0,395	0,575
0,01	-0,004	-0,12	0,111	0,016	Erro	0,077	0,017	0,010	0,030
0,18	0,304	0,189	0,420	0,261	0,157	0,380	0,165	0,114	0,231

Tabela 4.28 - EQM

Modelo	Modelo Linear Normal	Transf. Box-Cox	MLG
EQM	0,0033	0,0040	0,00021

4.2. "Estudo dos fatores sobre a distância alcançada pela bola lançada pela catapulta. ''

Este exemplo foi baseado em exemplo do texto de MYERS et al (2010). No exemplo, Schubert et al. (1992) conduziram um experimento usando uma catapulta para determinar os efeitos do gancho (A), o comprimento do braço (B), ângulo de partida (C) e o ângulo final (D) para a distância que alcança a bola lançada pela catapulta. Para cada combinação de fatores, a bola foi atirada três vezes. No planejamento realizado, considerou-se um experimento fatorial fracionado 2^{4-1}com três replicações, conforme a tabela a seguir.

Baseado, neste exemplo, foi feita uma simulação, na qual fixamos os coeficientes considerando o modelo com a variável resposta tendo uma distribuição gama e função de ligação logarítmica. A simulação dos dados foi realizada com base na seguinte equação:

$$y = e^{4+0,6A+0,06B+0,35C-0,045D+0,085AB+0,075BC}.$$

Tabela 4.9 – Plano fatorial do estudo sobre a distância alcançada pela bola lançada pela catapulta.

A	B	C	D	*Y*		
-1	-1	-1	-1	20,9	20,2	23,6
-1	-1	1	1	34,7	27,3	34,8
-1	1	-1	1	13,4	19,7	14,4
-1	1	1	-1	54,9	50,5	40,1
1	-1	-1	1	40,2	42,8	49,6
1	-1	1	-1	100,3	96,4	82,1
1	1	-1	-1	57,5	68,3	60,2
1	1	1	1	121,6	124,8	109,9

4.2.1. Modelo Linear Normal

Como o experimento é replicado, não é necessário construir o gráfico de probabilidade normal dos efeitos para selecionar os efeitos significativos. Existem graus de liberdade para calcular todos os efeitos. Portanto, para este caso, um modelo, considerando todos os efeitos seria:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

Tabela 4.10 - Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	54,508	1,249	43,65	<2E-16
A	24,967	1,249	19,99	9,63E-13
B	6,767	1,249	5,42	5,69E-05
C	19,608	1,249	14,90	8,44E-11
D	-1,742	1,249	-1,40	0,18219
AB	4,142	1,249	3,32	4,37E-03
AC	7,767	1,249	6,22	1,23E-05
BC	3,75	1,249	3,00	0,00843

Através da tabela 4.10, conclui-se que o efeito de D não é significante. Portanto, um novo ajuste foi construído descartando este efeito. Neste segundo ajuste, com as estatísticas na tabelas 4.11 e 4.12, observa-se que todos os efeitos são significativos. As estatísticas de qualidade do ajuste, R^2 múltiplo e R^2 ajustado são bem próximos de 1, indicam boa adequação. O p-valor para estatística F é muito pequeno, portanto a hipótese de que todos os coeficientes são nulos deve ser rejeitada. Observando a tabela ANOVA, pode se chegar à mesma conclusão, pois o quadrado médio de qualquer um dos fatores é bem superior ao quadrado médio dos resíduos.

Tabela 4.11 - Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	54,508	1,283	43,65	<2E-16
A	24,967	1,283	19,99	4,69E-13
B	6,767	1,283	5,42	6,21E-05
C	19,608	1,283	14,90	5,28E-11
AB	4,142	1,283	3,32	0,00494
AC	7,767	1,283	6,22	1,29E-05
BC	3,75	1,283	3,00	0,00950

R^2 Múltiplo=0,9753, R^2 Ajustado=0,9666

Estatística F=112,1 para 6 e 17 graus de liberdade

p-valor=1,02E-12

Tabela 4.12 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
A	1	14960	14960	378,625	4,69E-13
B	1	1098,9	1098,9	27,812	6,21E-05
C	1	8310.5	8310.5	210,331	5,28E-11
AB	1	411,7	411,7	10,419	0,004942
AC	1	1447,7	1447,7	36,640	1,29E-05
BC	1	337,5	337,5	8,542	0,009496
Resíduos	17				

Análise de Resíduos

Na figura 4.11, a distribuição dos pontos no gráfico não mostra nenhum indício de inadequação do modelo. Na figura 4.12, observa-se que os pontos estão dispostos ao longo da reta. Novamente utilizamos o teste de Anderson-Darling, com a estatística do teste igual a 0,2593 e o p-valor de 0,682. Assim, não se pode refutar a hipótese de que os resíduos são normalmente distribuídos.

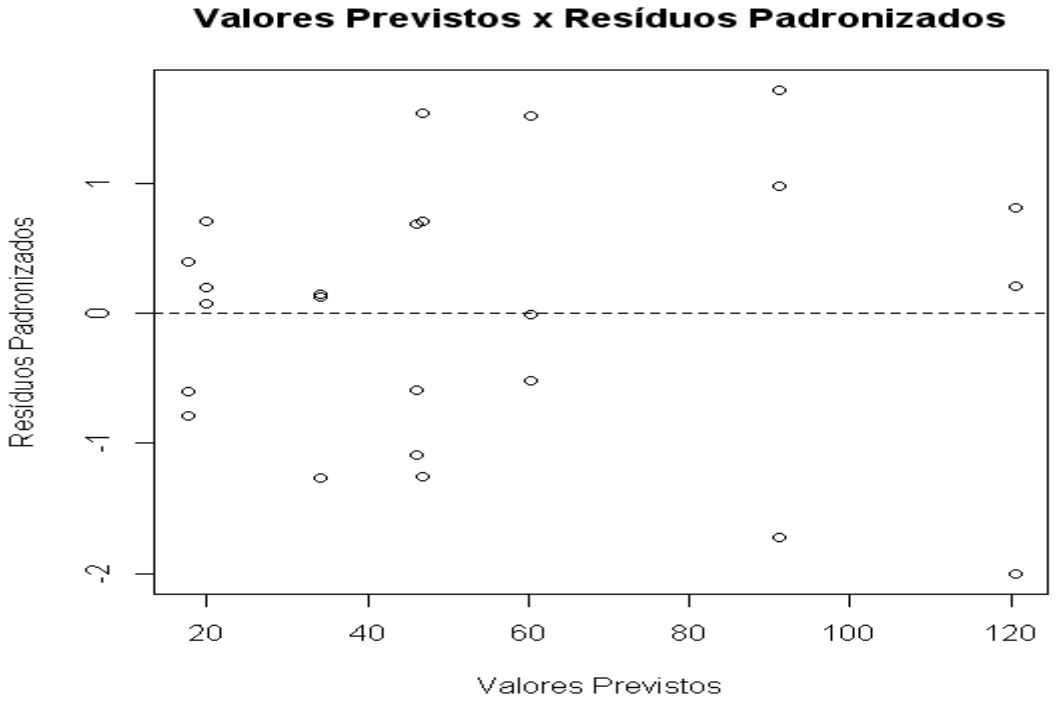

Figura 4.11 - Valores Previstos x Resíduos Padronizados

Figura 4.12 - Gráfico de Probabilidade Normal dos Resíduos

4.2.2. Transformação logarítmica

Essa transformação estabiliza a variância quando, nos dados originais, o desvio padrão varia com a média de modo diretamente proporcional, consequentemente, o coeficiente de variação é aproximadamente constante. Essa transformação é utilizada para contínuos e positivos. Portanto, o modelo, considerando todos os efeitos será

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

Tabela 4.13 - Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	3,7963	0,02601	145,958	<2E-16
A	0,5070	0,02601	19,492	1,42E-12
B	0,0839	0,02601	3,225	0,00529
C	0,3648	0,02601	14,027	2,08E-10
D	-0,1023	0,02601	-3,934	0,00119
AB	0,0629	0,02601	2,418	0,0279
AC	-0,0162	0,02601	-0,622	0,54251
BC	0,0793	0,02601	3,048	0,00767
R^2 Múltiplo=0,9748, R^2 Ajustado=0,9637				
Estatística F=88,3 para 7 e 16 graus de liberdade				
p-valor=1,379E-11				

Tabela 4.14 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
A	1	6,1687	6,1687	379,940	1,42E-12
B	1	0,1689	0,1689	10,404	0,005287
C	1	3,1946	3,1946	196,763	2,08E-10
D	1	0,2513	0,2513	15,477	0,00119
AB	1	0,0949	0,0949	5,846	0,027905
AC	1	0,0063	0,0063	0,387	0,54251
BD	1	0,1508	0,1508	9,290	0,00767
Resíduos	16	0,2598	0,0162		

Através das tabelas 4.13 e 4.14, percebe-se que, se considerarmos um nível de significância igual a 5%, o efeito da interação AC não é significante. Portanto, tentaremos um novo ajuste descartando essa interação.

Tabela 4.15 - Estimativa dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	3,79630	0,02554	148,662	<2E-16
A	0,50698	0,02554	19,853	3,38E-13
B	0,08389	0,02554	3,285	0,004367
C	0,36484	0,02554	14,287	6,69E-11
D	-0,10232	0,02554	-4,007	0,00091
AB	0,06289	0,02554	2,463	0,024764
BC	0,07928	0,02554	3,104	0,006443
R^2 Múltiplo=0,9742, R^2 Ajustado=0,965				
Estatística F=106,8 com 6 e17 graus de liberdade				
p-valor=1,522E-12				

Tabela 4.16 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
A	1	6,1687	6,1687	394,146	3,38E-13
B	1	0,1689	0,1689	10,793	0,004367
C	1	3,1946	3,1946	204,120	6,69E-11
D	1	0,2513	0,2513	16,056	0,00091
AB	1	0,0949	0,0949	6,065	0,024764
BC	1	0,1508	0,1508	9,638	0,006443
Resíduos	17	0,2661	0,0157		

Dessa vez, na tabela 4.15 e 4.16, todos os efeitos contidos no ajuste são significantes. Novamente, R^2 múltiplo e R^2 ajustado indicam boa adequação do ajuste. O p-valor para o teste de significância da regressão é muito pequeno.

Análise de Resíduos

Observando a figura 4.13, os pontos estão distribuídos aleatoriamente, então se conclui que não foi detectada nenhuma inadequação do modelo. Na figura 4.14, os pontos estão razoavelmente distribuídos em torno da reta, realizando o teste de Anderson-Darling para os resíduos, obteve-se estatística igual a 0,2295 com p-valor de 0,7844. Portanto, não podemos rejeitar a hipótese de normalidade dos resíduos.

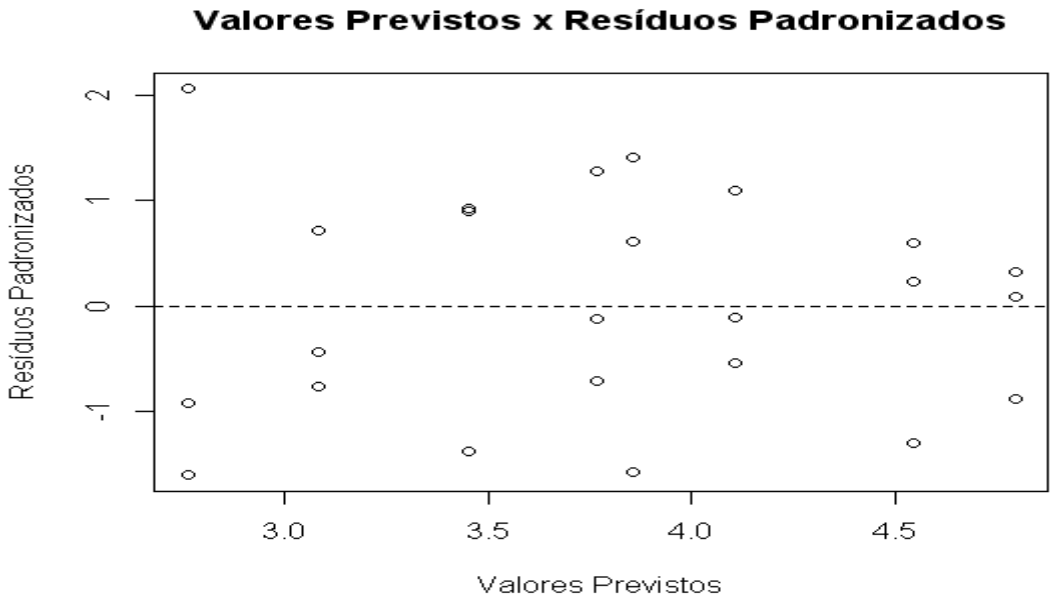

Figura 4.13 - Valores Previstos x Resíduos Padronizados

Figura 4.14 - Gráfico de Probabilidade Normal dos Resíduos

4.2.3. Modelo Linear Generalizado

Distribuição Gama com ligação logarítmica

A função logarítmica não é a ligação canônica para distribuição gama. Mesmo assim, esta função pode ser empregada de maneira bem sucedida, por que, diferentemente da ligação recíproca, a ligação logarítmica não origina valores negativos para estimativa da resposta. Além disso, a função logarítmica está muito relacionada ao modelo linear normal como foi feito neste exemplo anteriormente. Outro ponto importante é que a utilização da ligação recíproca implica em impor restrições aos possíveis valores para os parâmetros β_j.

Tabela 4.17 - Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	3,8017	0,02605	145,958	<2E-16
A	0,5045	0,02605	19,370	1,57E-12
B	0,0853	0,02605	3,274	0,00477
C	0,3644	0,02605	13,992	2,16E-10
D	-0,1012	0,02605	-3,886	0,00131
AB	0,0607	0,02605	2,328	0,0332
AC	-0,0161	0,02605	-0,620	0,54412
BC	0,0779	0,02605	2,992	0,00862

Tabela 4.18 - Análise da Deviance

Termo do Modelo	Graus de Liberdade	Deviance	Graus de Liberdade Restantes	Deviance Residual
Nulo			23	9,6984
A	1	5,6511	22	4,0473
B	1	0,3089	21	3,7384
C	1	2,9989	20	0,7395
D	1	0,2430	19	0,4965
AB	1	0,0846	18	0,4119
AC	1	0,0062	17	0,4056
BC	1	0,1456	16	0,2600

Segundo os dados das tabelas 4.17, o único efeito não significante do modelo é AC. Podemos chegar a esta conclusão observando o p-valor para a estimativa do coeficiente para AC que é superior ao nível de significância de 5%. Podemos chegar a mesma conclusão ao observar a deviance para AC na tabela 4.18, pois sua contribuição é muito pequena comparada aos outros fatores.

Tabela 4.19 - Estimativa dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	3,8019	0,02556	148,755	<2E-16
A	0,5045	0,02556	19,740	3,71E-13
B	0,0853	0,02556	3,336	0,00391
C	0,3644	0,02556	14,259	6,90E-11
D	-0,1012	0,02556	-3,960	0,00101
AB	0,0607	0,02556	2,373	0,0297
BC	0,0779	0,02556	3,049	0,00725

Tabela 4.20 - Análise da Deviance

Fonte de Variação	Graus de Liberdade	Deviance	Graus de Liberdade Restantes	Deviance Residual
Nulo			23	9,6984
A	1	5,6511	22	4,0473
B	1	0,3089	21	3,7384
C	1	2,9989	20	0,7395
D	1	0,2430	19	0,4965
AB	1	0,0846	18	0,4119
BC	1	0,1456	17	0,2663

No segundo ajuste, desconsiderando a interação AC, o p-valor da diferença de deviance entre o modelo completo e o modelo reduzido é muito próximo de 1, considerando uma distribuição qui-quadrado com 9 graus de liberdade. A estatística de Pearson é igual a 0,2665, também possuindo p-valor próximo de 1. Essas informações indicam a boa adequação do ajuste.

Análise de Resíduos

Na figura 4.15, da mesma forma que o modelo anterior, nenhuma inadequação foi observada. Na figura 4.16, quase todos os pontos estão distribuídos ao longo da reta. O p-valor do teste de Anderson-Darling para a normalidade dos resíduos é igual a 0,682. Dessa forma, a hipótese de que os resíduos estão distribuídos normalmente não é rejeitada.

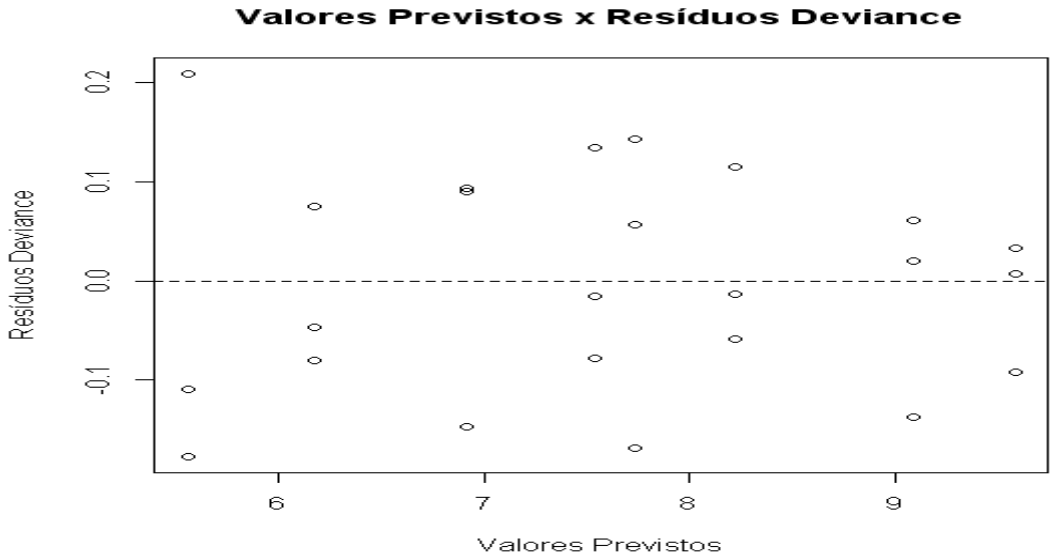

Figura 4.15 - Valores Previstos x Resíduos Deviance

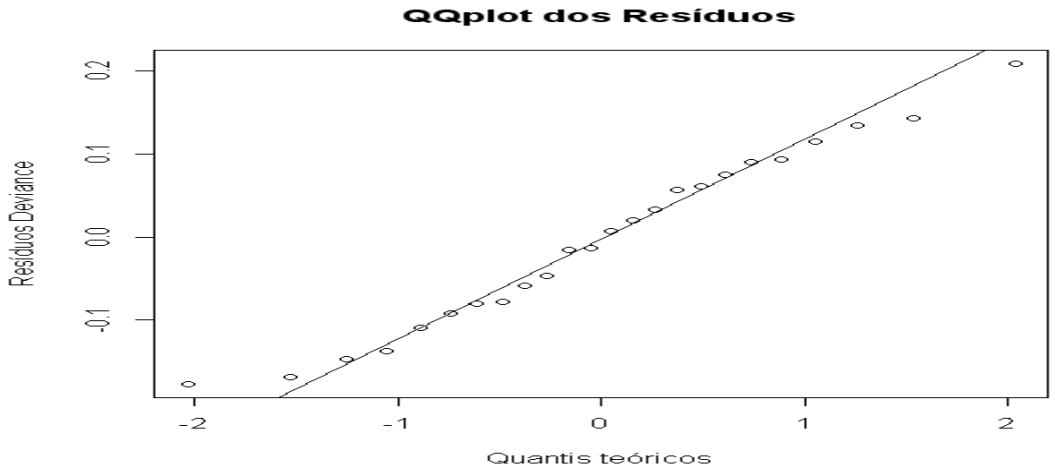

Figura 4.16 - Gráfico de Probabilidade Normal dos Resíduos Deviance

4.2.4. Conclusões

Neste caso, apenas observando os dados é difícil saber qual modelo fornece as melhores previsões. Por isto, neste caso, uso do erro médio quadrático se torna ainda mais útil. Constata-se que o EQM para os ajustes da transformação logarítmica e do MLG são muito próximos. No entanto, quando o MLG é utilizado, verifica-se que intervalo de confiança para a resposta média não é constante, diferentemente dos outros dois casos, a amplitude do intervalo aumenta à medida que a resposta cresce, o que é positivo, porque o intervalo parece se adequar ao valor de cada observação. Além disso, apesar de não apresentar inadequações, deve-se lembrar que o modelo obtido por meio

da transformação da resposta tem sempre uma desvantagem: os dados obtidos têm de ser recolocados na escala original.

Tabela 4.21- Os valores previstos e os limites inferiores e superiores do intervalo de confiança de 95% para a resposta média.

	Modelo Linear Normal			Transf. Log.			MLG		
Resposta	Previsto	LI	LS	Previsto	LI	LS	Previsto	LI	LS
20,9	19,825	12,663	26,987	21,869	18,964	25,220	21,921	19,200	25,027
34,7	34,008	26,846	41,171	31,550	27,358	36,383	31,754	27,813	36,254
13,4	17,575	10,413	24,737	15,861	13,754	18,291	16,093	14,096	18,374
54,9	46,758	39,596	53,921	47,312	41,027	54,561	47,73	41,805	54,494
40,2	45,942	38,779	53,104	43,320	37,565	49,957	43,498	38,099	49,662
100,3	91,192	84,029	98,354	94,103	81,601	108,520	94,458	82,733	107,85
57,5	60,258	53,096	67,421	60,841	52,758	70,163	61,015	53,441	69,662
121,6	120,51	113,35	127,671	120,525	104,513	138,990	120,715	105,731	137,82
20,2	19,825	12,663	26,987	21,869	18,964	25,220	21,921	19,200	25,027
27,3	34,008	26,846	41,171	31,550	27,358	36,383	31,754	27,813	36,254
19,7	17,575	10,413	24,737	15,861	13,754	18,291	16,093	14,096	18,374
50,5	46,758	39,596	53,921	47,312	41,027	54,561	47,73	41,805	54,494
42,8	45,942	38,779	53,104	43,320	37,565	49,957	43,498	38,099	49,662
96,4	91,192	84,029	98,354	94,103	81,601	108,520	94,458	82,733	107,85
68,3	60,258	53,096	67,421	60,841	52,758	70,163	61,015	53,441	69,662
124,8	120,51	113,35	127,671	120,525	104,513	138,990	120,715	105,731	137,82
23,6	19,825	12,663	26,987	21,869	18,964	25,220	21,921	19,200	25,027
34,8	34,008	26,846	41,171	31,550	27,358	36,383	31,754	27,813	36,254
14,4	17,575	10,413	24,737	15,861	13,754	18,291	16,093	14,096	18,374
40,1	46,758	39,596	53,921	47,312	41,027	54,561	47,730	41,805	54,494
49,6	45,942	38,779	53,104	43,320	37,565	49,957	43,498	38,099	49,662
82,1	91,192	84,029	98,354	94,103	81,601	108,520	94,458	82,733	107,85
60,2	60,258	53,096	67,421	60,841	52,758	70,163	61,015	53,441	69,662
109,9	120,51	113,35	127,671	120,525	104,513	138,990	120,715	105,731	137,82

Tabela 4.22 - EQM

Modelo	Modelo Linear	Transf. Log.	MLG
EQM	27,987	26,028	26,033

4.3. “Sobrevivência do Espermatozóide em um banco de esperma’’

Este exemplo foi retirado do texto MONTGOMERY, MYERS,VINING, ROBINSON (2010). O exemplo selecionado é originado de um estudo sobre a sobrevivência de espermatozóide em um banco de esperma. Os espermatozóides são armazenados em citrato de sódio e glicerol, e as quantidades dessas substâncias variaram juntamente com o tempo de equilíbrio em um conjunto fatorial. Cinquenta amostras de material foram usadas em cada combinação de tratamento. O propósito desse experimento é avaliar o efeito dos fatores na proporção de sobrevivência.

Tabela 4.23 – Plano fatorial do estudo de sobrevivência dos espermatozóides.

x_1 (Citrato de Sódio)	x_2 (Glicerol)	x_3 (tempo de equilíbrio)	y(sobreviventes)
-1	-1	-1	34
1	-1	-1	20
-1	1	-1	8
1	1	-1	21
-1	-1	1	30
1	-1	1	20
-1	1	1	10
1	1	1	25

4.3.1. Modelo Linear Normal

Como o plano possui apenas uma replicação, é necessário construir o gráfico de probabilidade normal para identificar os efeitos significantes. Os efeitos B e AB são os únicos localizados distantes da reta, então são os únicos efeitos significantes do modelo.

O fator A será adicionado ao ajuste, mesmo que aparentemente não seja importante para o modelo, a significância da interação AB é um indicio que, de algum modo, a real importância de B pode estar ofuscada devido aos outros fatores. Dessa forma, temos que o ajuste será $\hat{p} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_1 x_1 x_2$, no qual, x_1 e x_2 representam, respectivamente, os fatores A e B.

No entanto, observa-se que na tabela 4.24, o p-valor para fator A ultrapassa o nível de significância. Portanto, descartamos o efeito A deste modelo.

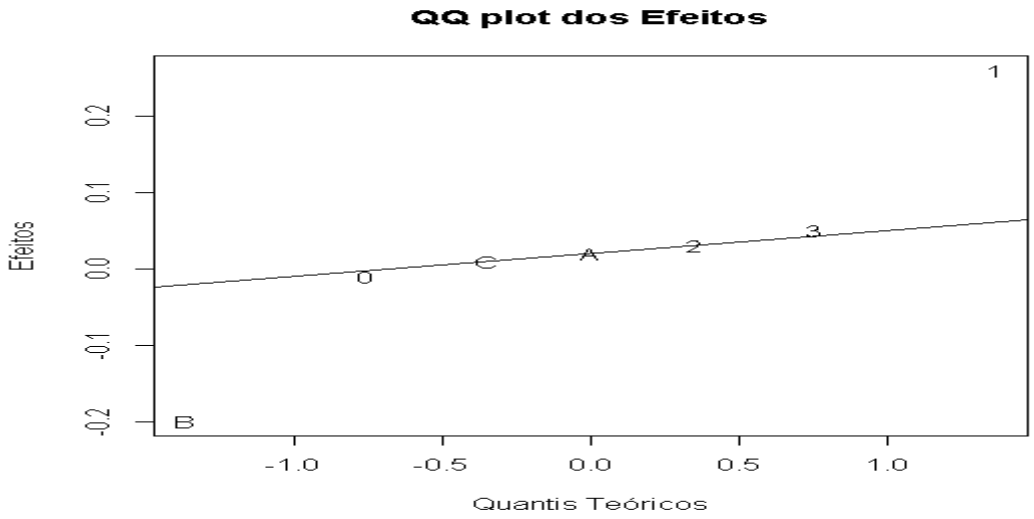

Figura 4.17 - Gráfico de Probabilidade Normal dos Efeitos

Tabela 4.24 - Estimativa dos coeficientes do modelo

Coeficiente	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	0,420	0,015	28	9,68E-06
A	0,010	0,015	0,667	0,54147
B	-0,100	0,015	-6,667	0,00263
AB	0,130	0,015	8,667	0,000975

Tabela 4.25 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
A	1	0,0008	0,0008	0,4444	0,54147
B	1	0,08	0,08	44,4444	0,00263
AB	1	0,1352	0,1352	75,1111	0,00098
Resíduos	4	0,0072	0,0018		

Nesta nova tentativa de ajuste, com as informações nas tabelas 4.26 e 4.27, todos os coeficientes são significantes, a estatística F do modelo possui um p-valor muito baixo, portanto rejeitamos a hipótese nula de que todos os coeficientes são zero. Além disso, o R^2 múltiplo e o R^2 ajustado estão próximos de 1, isso indica que o ajuste é adequado.

Tabela 4.26 - Estimativas dos coeficientes

Coeficiente	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	0,420	0,01414	29,698	8,12E-07
B	-0,100	0,01414	-7,071	0,000875
AB	0,130	0,01414	9,192	0,000256

R^2 múltiplo: 0,9642, R^2 ajustado: 0,9498
Estatística F: 67,25 com 2 e 5 graus de liberdade, p-valor: 0,0002432

Tabela 4.27 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
B	1	0,08	0,08	50,0	0,00088
AB	1	0,1352	0,1352	84,5	0,00026
Residuos	5	0,008	0,0016		

Análise de Resíduos

Neste exemplo, trabalhamos com somente oito observações, isto complica a análise visual, porém o gráfico contido na figura 4.18 não revela nenhum indicio de violação à suposição de homogeneidade da variância, pois os pontos estão distribuídos de modo aleatório. Na figura 4.19, os pontos estão, razoavelmente, próximos a reta. O teste de Anderson-Darling para a normalidade dos resíduos tem um p-valor de 0,379. Dessa forma, não podemos descartar a hipótese de que os resíduos seguem a distribuição normal.

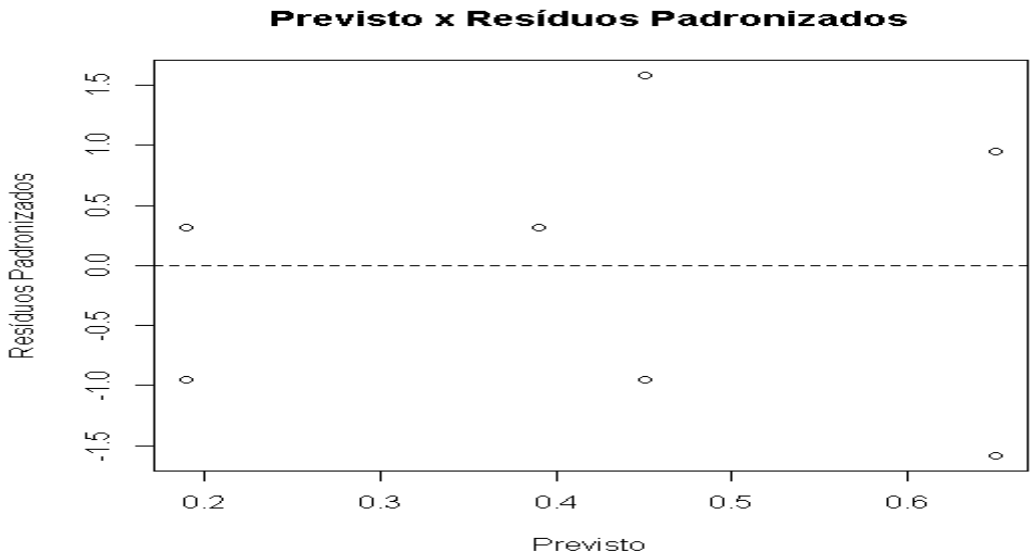

Figura 4.18 - Valores Previstos x Resíduos Padronizados

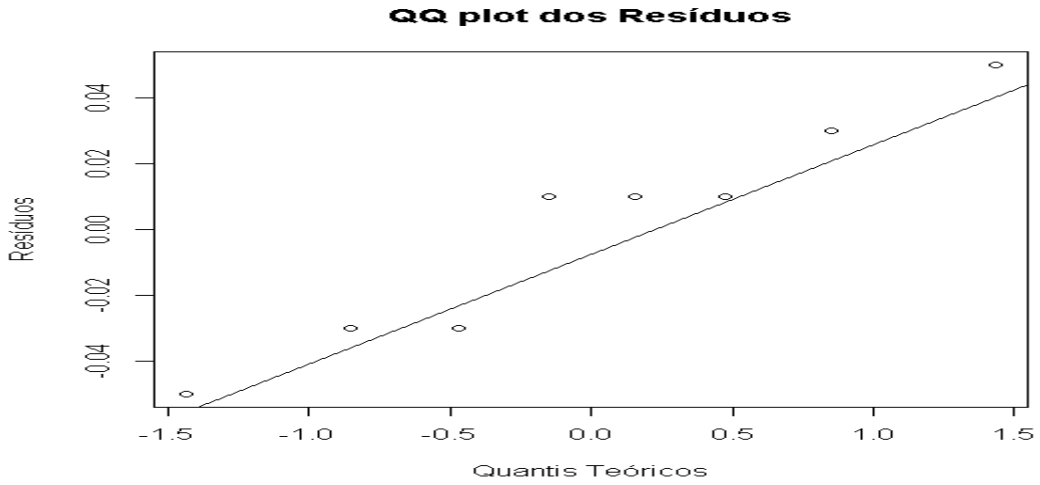

Figura 4.19 – Gráfico de Probabilidade Normal

4.3.2. Transformação Arco Seno da Raiz Quadrada

A transformação $arcsen\sqrt{y}$ é utilizada para estabilizar a variância quando a variável resposta é uma proporção. No gráfico contido na figura 4.20, vemos que os efeitos estão distribuídos de forma quase idêntica ao caso em que usamos os dados originais. Dessa forma, os efeitos significantes são B e AB.

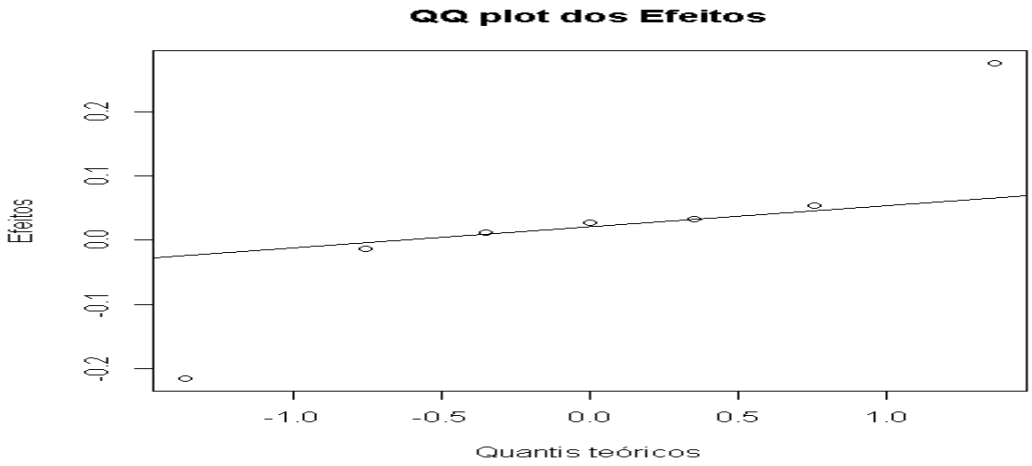

Figura 4.20 – Gráfico de Probabilidade Normal dos Efeitos

O p-valor da estatística F mostra que devemos rejeitar a hipótese nula, que diz que todos os coeficientes são iguais à zero. O R^2 Múltiplo e o R^2 Ajustado possuem valores muito altos, mostrando que o modelo pode ser adequado.

Tabela 4.28 - Estimativas dos Coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	0,69883	0,01593	43,864	1,16E-07
B	-0,10743	0,01593	-6,743	0,001088
AB	0,13768	0,01593	8,642	0,000343

R^2 Múltiplo: 0,960, R^2 Ajustado: 0,9441
Estatística F: 60,08 com 2 e 5 graus de liberdade, p-valor: 0,000319

Tabela 4.29 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
B	1	0,092328	0,092328	45,470	0,00109
AB	1	0,151651	0,151651	74,685	0,00034
Resíduos	5	0,010153	0,002031		

Análise de Resíduos

Na figura 4.21, não se pode observar nenhuma inadequação ao modelo ou violação à suposição de homogeneidade da variância. Na figura 4.22, observa-se que os pontos estão mais próximos da reta em comparação ao modelo obtido com os dados, a estatística do teste de Anderson-Darling para os resíduos é igual a 0,3068 com p-valor de 0,482. Assim, a hipótese de normalidade dos resíduos não é rejeitada.

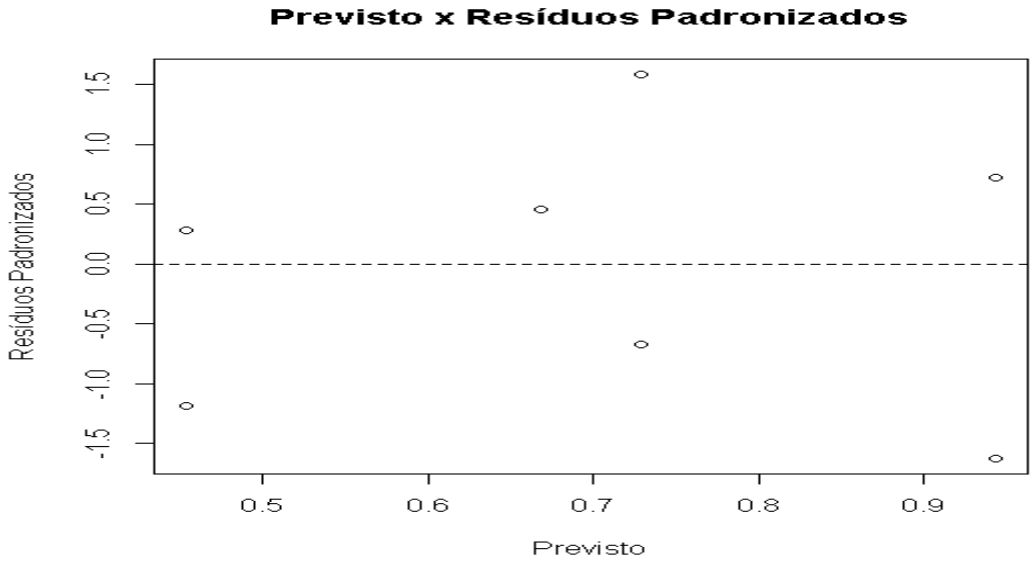

Figura 4.21 - Valores Previstos x Resíduos Padronizados

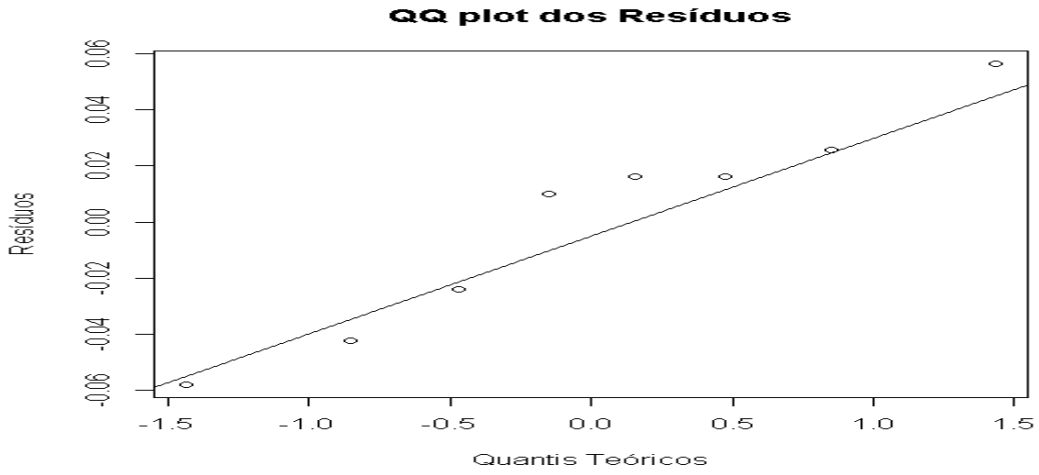

Figura 4.22 – Gráfico de Probabilidade Normal

4.3.3. Modelos Lineares Generalizados

Considerando uma distribuição binomial com função de ligação logit, tentaremos obter um ajuste do modelo. Neste caso, também o gráfico de probabilidade normal para selecionar os efeitos significantes.

No gráfico a seguir, percebemos que o comportamento dos efeitos é o mesmo em relação aos dois casos anteriores, apenas B e AB são significantes.

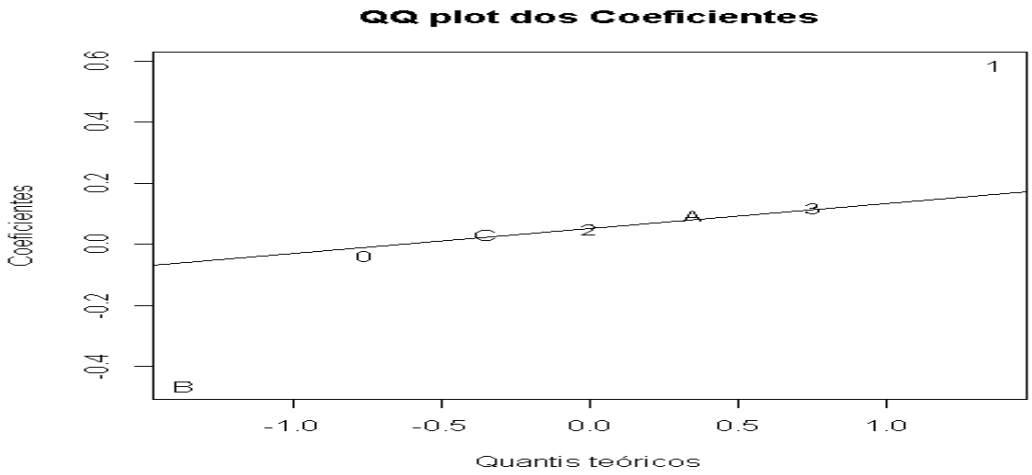

Figura 4.23 – Gráfico de Probabilidade Normal dos Coeficientes

Os dados das tabelas a seguir, comprovam que B e AB realmente são significantes. A diferença de deviance comparada ao modelo é 2,345, considerando uma distribuição qui-quadrado com 5 graus de liberdade, p-valor para esta estatística é igual a 0,799. Portanto, não podemos rejeitar a hipótese nula, o modelo reduzido é adequado. Além disso, a estatística do teste qui-quadrado de Pearson é igual a 2,342, que fornece um p-valor de 0,80, também indicando um bom ajuste do modelo.

Tabela 4.29 - Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor z	Pr(>|z|)
Intercepto	-0,3637	0,1081	-3,363	0,00077
B	-0,4505	0,1084	-4,155	3,25E-05
AB	0,5747	0,1086	5,291	1,22E-07

Tabela 4.30 - Análise da Deviance

Termo do Modelo	Graus de Liberdade	Deviance	Graus de Liberdade Restantes	Diferença de Deviance
Nulo	7	48,292		
B	1	16,547	6	31,745
AB	1	29,4	5	2,345

Análise de Resíduos

Na figura a seguir, não foi observado qualquer padrão na distribuição dos resíduos. Com isso, não há evidências de inadequação do modelo ou violação da hipótese de variância constante. Através do gráfico da figura 4.25, não é possível obter uma conclusão sobre a distribuição dos resíduos, pois análise visual continua difícil.
O p-valor do teste de Anderson-Darling para os resíduos deviance é igual a 0,405, portanto podemos considerar que resíduos deviance seguem a distribuição normal.

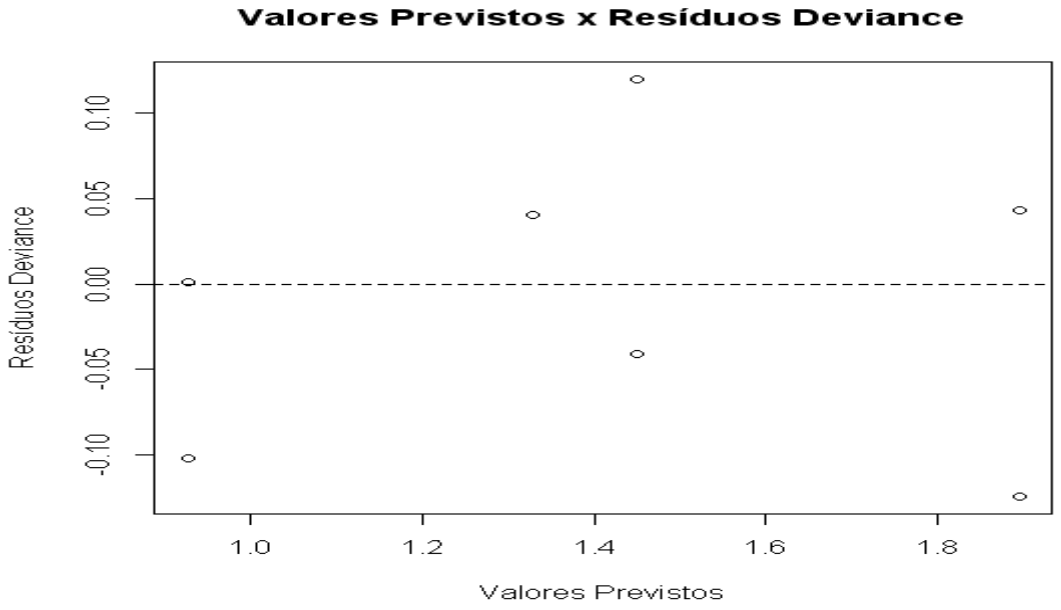

Figura 4.24 - Valores Previstos x Resíduos Deviance

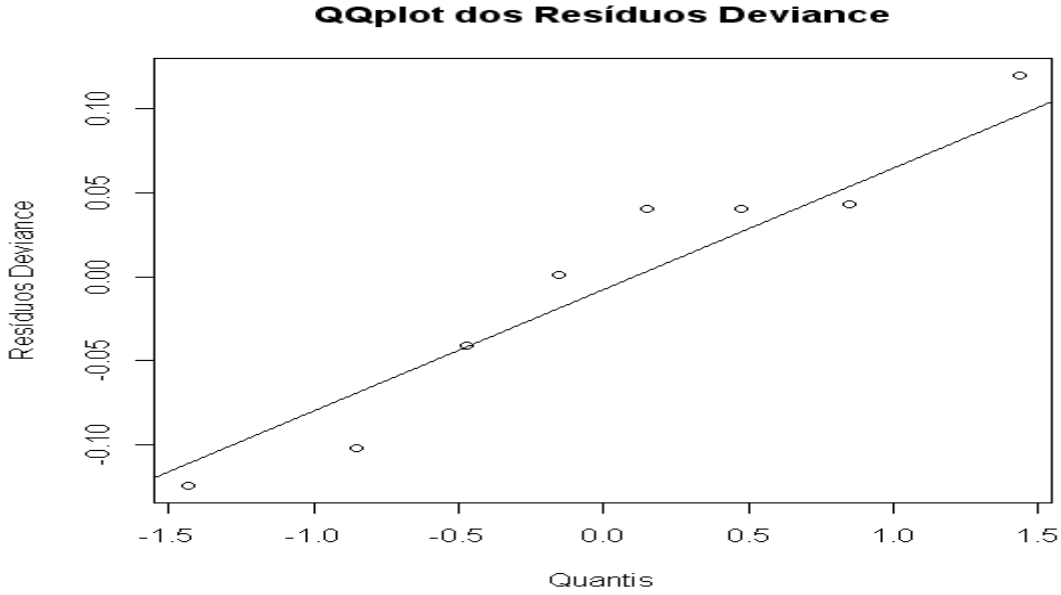

Figura 4.25 – Gráfico de Probabilidade Normal

4.3.4. Conclusões

Todos os modelos apresentam previsões próximas aos valores reais da variável resposta. Isto se comprova por meio dos resultados para o erro quadrático médio, nas três situações, os valores são pequenos e próximos. No entanto, as amplitudes dos intervalos de confiança para a resposta média foram maiores quando foi utilizado o MLG. Portanto, neste caso, o modelo linear sob má-especificação não foi inferior aos outros.

Tabela 4.31 - Os valores previstos, limites inferiores e limites superiores do intervalo de confiança de 95% para a resposta média dos três casos.

	Modelo Linear Normal			Transformação			MLG		
Resposta	Prev	LI	LS	Prev	LI	LS	Prev	LI	LS
0,68	0,650	0,587	0,713	0,656	0,587	0,722	0,666	0,574	0,736
0,4	0,390	0,327	0,453	0,384	0,317	0,454	0,380	0,301	0,467
0,16	0,190	0,127	0,253	0,192	0,140	0,251	0,200	0,143	0,271
0,42	0,450	0,387	0,513	0,444	0,374	0,515	0,440	0,357	0,528
0,6	0,650	0,587	0,713	0,656	0,587	0,722	0,666	0,574	0,736
0,4	0,390	0,327	0,453	0,384	0,317	0,454	0,380	0,301	0,467
0,2	0,190	0,127	0,253	0,192	0,140	0,251	0,200	0,143	0,271
0,5	0,450	0,387	0,513	0,444	0,374	0,515	0,440	0,357	0,528

Tabela 4.32 - EQM

Modelo	Modelo Linear	Transf. Arco Seno	MLG
EQM	0,0010	0,0011	0,0012

4.4. “Simulação de dados binomiais II”

Esse exemplo é muito semelhante ao 4.1, a única diferença é que o número de observações para cada combinação de tratamento é igual a 20. No exemplo anterior, temos uma variável de 100 observações com distribuição de Bernoulli. Segundo o Teorema Central do Limite, esta variável pode ter uma distribuição aproximadamente normal, mas se utilizamos um número de observações bem inferior tal afirmação não pode ser feita.

Tabela 4.33 – Simulação dos dados binomiais II

Observação	A	B	C	D	*p*	Resposta
1	-1	-1	-1	-1	0,65	0,60
2	1	-1	-1	-1	0,98	0,95
3	-1	1	-1	-1	0,03	0,05
4	1	1	-1	-1	0,45	0,40
5	-1	-1	1	-1	0,6	0,70
6	1	-1	1	-1	0,35	0,30
7	-1	1	1	-1	0,02	0,00
8	1	1	1	-1	0,01	0,00
9	-1	-1	-1	1	0,07	0,05
10	1	-1	-1	1	0,99	1,00
11	-1	1	-1	1	0,02	0,05
12	1	1	-1	1	0,95	0,95
13	-1	-1	1	1	0,06	0,55
14	1	-1	1	1	0,45	0,45
15	-1	1	1	1	0,01	0,00
16	1	1	1	1	0,17	0,20

4.4.1. Modelo Linear Normal

Da mesma forma que o exemplo 4.1, o gráfico de probabilidade normal dos efeitos indica que os efeitos significantes são A, B, C, AC, AD e BD.

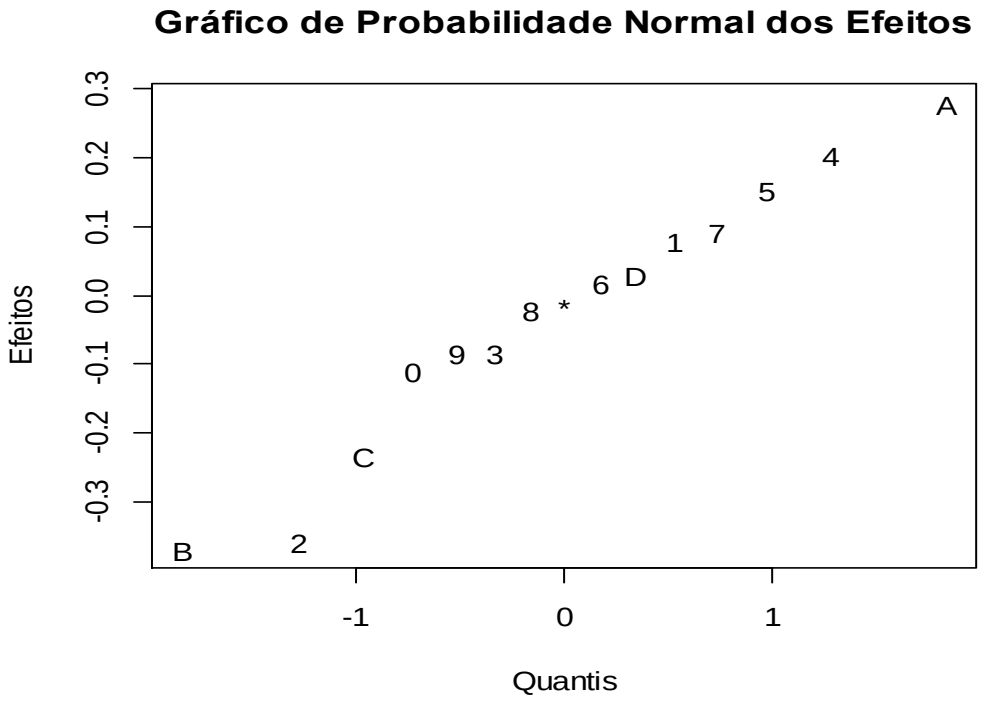

Figura 4.26 – Gráfico de Probabilidade Normal dos Efeitos

A tabela 4.33 mostra que todos os fatores selecionados são realmente significantes, e todas estatísticas indicam uma boa adequação do ajuste. Na tabela 4.34, temos as estimativas dos coeficientes de todos os fatores.

Tabela 4.33 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
A	1	0,31641	0,31641	17,113	2,53E-03
B	1	0,54391	0,54391	29,417	4,20E-04
C	1	0,21391	0,21391	11,569	7,86E-03
AC	1	0,50766	0,50766	27,456	5,35E-04
AD	1	0,17016	0,17016	9,203	0,01416
BD	1	0,09766	0,09766	5,282	0,04714
Resíduos	9	0,16641	0,01849		

R^2 Múltiplo: 0,9175, R^2 Ajustado: 0,8624
Estatística F: 16,67 com 6 e 16 graus de liberdade, p-valor: 0,002073

Tabela 4.34 -Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	0,39062	0,03399	11,491	1,11E-06
A	0,14062	0,03399	4,137	0,002534
B	-0,18438	0,03399	-5,424	0,00042
C	-0,11562	0,03399	-3,401	0,007857
AC	-0,17812	0,03399	-5,240	0,000535
AD	0,10312	0,03399	3,034	0,014164
BD	0,07812	0,03399	2,298	0,047139

Análise de Resíduos

Na figura 4.27, não é possível identificar nenhuma inadequação do modelo. Na figura 4.28, pontos estão dispostos ao longo da reta. O p-valor do teste de Anderson-Darling para a normalidade dos resíduos é igual a 0,790. Assim, podemos considerar que os resíduos estão normalmente distribuídos.

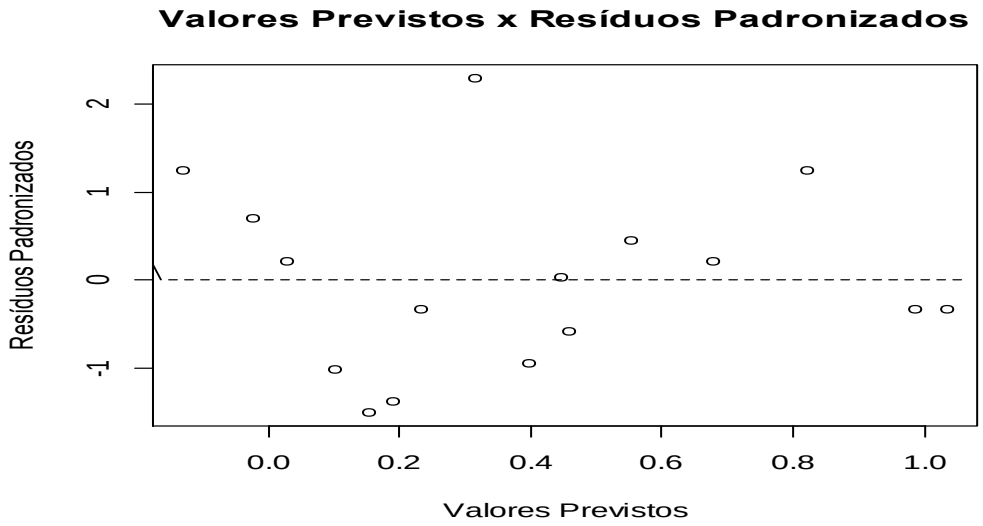

Figura 4.27 – Valores Previstos x Resíduos Padronizados

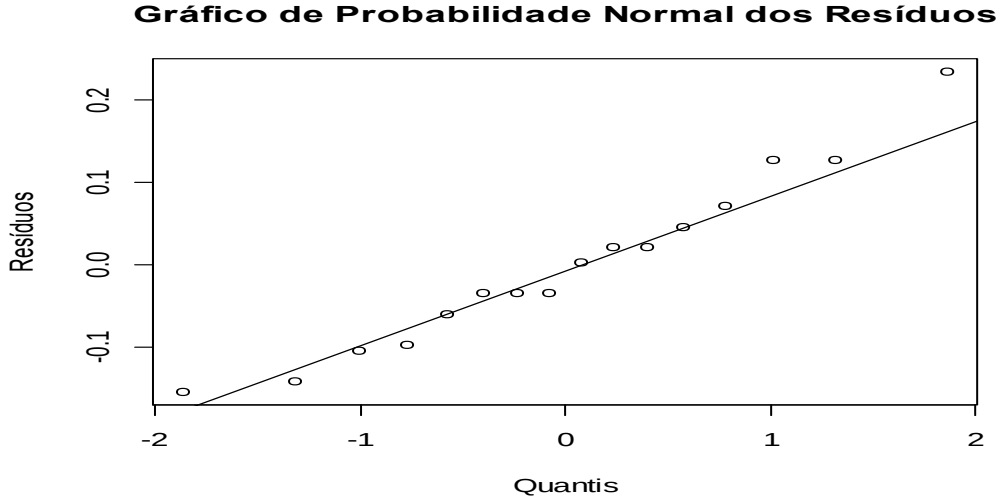

Figura 4.28 – Gráfico de Probabilidade Normal dos Resíduos

4.4.2. Transformação Arco Seno da Raiz Quadrada

Esta transformação é muito utilizada para homogeneizar a variância de dados que são proporções, especialmente quando estas cobrem uma amplitude grande de valores. O gráfico de probabilidade normal é novamente utilizado para selecionar os efeitos significantes e se comporta de modo muito parecido com gráfico para o caso dos dados originais. Dessa forma, os fatores considerados como significantes serão os mesmos.

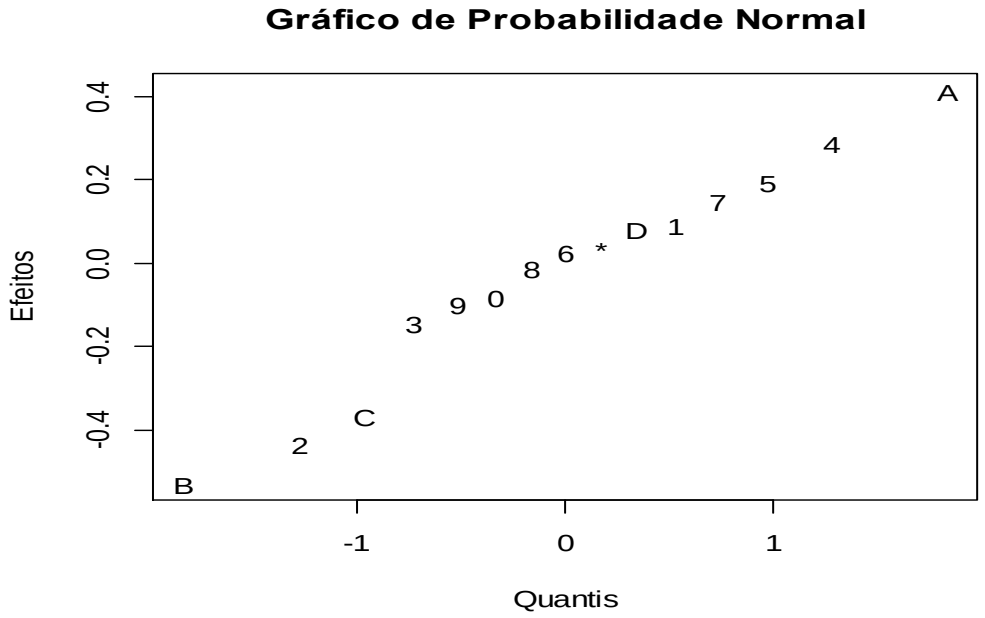

Figura 4.29 – Gráfico de Probabilidade Normal dos Efeitos

Entretanto, a tabela 4.35 revela que, considerando um nível de significância de 5%, a interação BD não é significante. Portanto, este efeito é descartado do ajuste.

Tabela 4.35 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
A	1	0,69532	0,69532	20,192	0,001502
B	1	1,11545	1,11545	32,393	0,000297
C	1	0,52688	0,52688	15,301	0,003556
AC	1	0,73595	0,73595	21,372	0,000125
AD	1	0,33688	0,33688	9,783	0,012165
BD	1	0,15196	0,15196	4,413	0,065047
Resíduos	9	0,30992	0,03444		

Neste novo ajuste, com as informações nas tabelas 4.33 e 4.34, todos os fatores são significantes e as estatísticas indicam uma boa adequação do modelo.

Tabela 4.33 - ANOVA

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Valor F	Pr(>F)
A	1	0,69532	0,69532	15,0543	0,003059
B	1	1,11545	1,11545	24,1505	0,00061
C	1	0,52688	0,52688	11,4074	0,007032
AC	1	0,73595	0,73595	15,934	0,002552
AD	1	0,33688	0,33688	7,2937	0,022287
Resíduos	9	0,46187	0,04619		
R^2 Múltiplo: 0,8807, R^2 Ajustado:0,08211					
Estatística F: 14,77 com 5 e 10 graus de liberdade, p-valor: 0,000242					

Tabela 4.34 - Estimativas dos coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	0,63212	0,03399	11,491	1,11E-06
A	0,20846	0,03399	4,137	0,002534
B	-0,26404	0,03399	-5,424	0,00042
C	-0,18147	0,03399	-3,401	0,007857
AC	-0,21447	0,03399	-5,240	0,000535
AD	0,14510	0,03399	3,034	0,014164

Análise de Resíduos

Na figura 4.31, os pontos estão distribuídos aleatoriamente no gráfico. Então, não é possível identificar nenhuma inadequação do modelo. Na figura 4.30, os pontos estão dispostos ao longo da reta. O p-valor da estatística do teste de Anderson-Darling para a normalidade dos resíduos é igual a 0,782. Com isso, podemos considerar que os resíduos estão normalmente distribuídos.

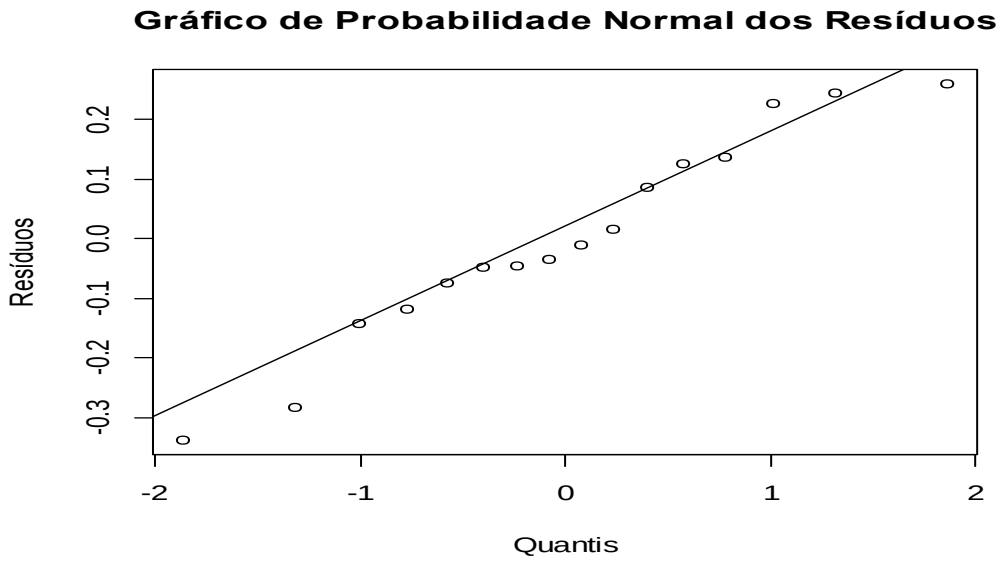

Figura 4.30 – Gráfico de Probabilidade Normal dos Resíduos.

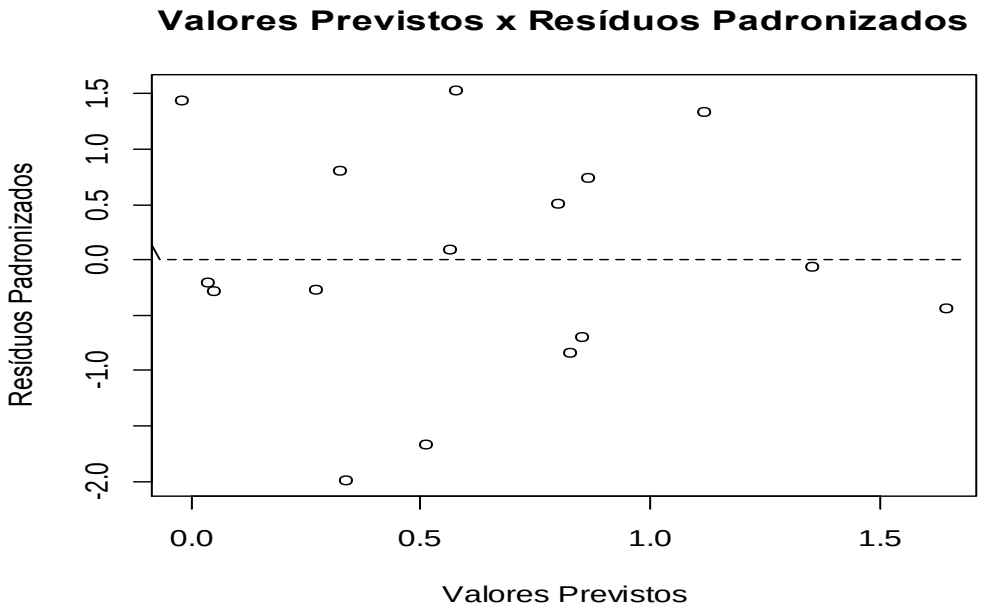

Figura 4.31 – Valores Previstos x Resíduos Padronizados

4.4.3. Modelos Lineares Generalizados

Novamente, o modelo é obtido por meio da regressão logística e espera-se que este se mostre o modelo mais adequado. O gráfico de probabilidade normal dos coeficientes é usado para a seleção dos fatores significativos.

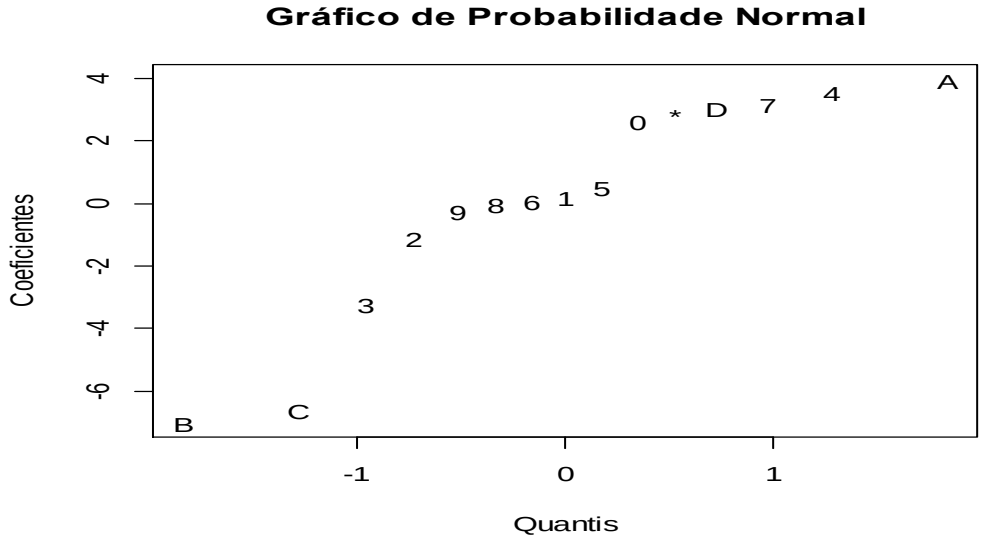

Figura 4.32 – Gráfico de Probabilidade Normal dos Coeficientes

Após observar a figura 4.32, os fatores selecionados foram A, B, C, D, AB, BC, AD, BD e ABC. Entretanto, as informações contidas nas tabelas 4.35 mostram que o fator D não é significante, pois sua contribuição para a deviance é muito pequena em comparação aos outros fatores. Repare que, surpreendentemente, alguns fatores considerados significantes neste modelo não estão presentes na equação para simulação dos dados.

Tabela 4.35 - Análise da Deviance

Fonte de Variação	Graus de Liberdade	Deviance	Graus de Liberdade Restantes	Deviance Residual
Nulo			15	210,697
A	1	27,048	14	183,649
B	1	52,058	13	131,591
C	1	23,98	12	107,611
D	1	0,464	11	107,147
AB	1	11,923	10	95,224
BC	1	12,911	9	82,312
AD	1	21,544	8	60,769
BD	1	9,986	7	50,783
ABC	1	43,978	6	6,805

No novo ajuste, com todos os fatores significativos, a diferença de deviance entre o modelo completo e o modelo reduzido é igual a 9,788. Considerando uma distribuição qui-quadrado com 7 graus de liberdade, o p-valor para a diferença de deviance é 0,20. A estatística do teste de Pearson é igual a 9,38 com um p-valor de 0,22. Todas informações indicam que o modelo é adequado. Para confirmação, é necessário analisar os resíduos.

Tabela 4.35 - Análise da Deviance

Fonte de Variação	Graus de Liberdade	Deviance	Graus de Liberdade Restantes	Deviance Residual
Nulo			15	210,697
A	1	27,048	14	183,649
B	1	52,058	13	131,591
C	1	23,98	12	107,611
AB	1	11,916	11	95,695
AC	1	12,890	10	82,805
AD	1	22,036	9	60,769
BD	1	6,704	8	54,065
ABC	1	44,276	7	9,788

Tabela 4.36 - Estimativas dos Coeficientes

Coeficientes	Estimativa	Erro Padrão	Valor z	Pr(>|z|)
Intercepto	-1,7751	0,2689	-6,603	4,04R-11
A	2,1524	0,3453	6,233	4,58E-10
B	-2,5029	0,3602	-6,948	3,70E-12
C	-1,9556	0,3218	-6,078	1,22E-09
AB	1,2849	0,2446	5,253	1,50E-07
BC	-1,3010	0,237	-5,490	4,02E-08
AD	0,7113	0,1846	3,853	0,00012
BD	0,3796	0,1846	2,057	0,03973
ABC	1,4654	0,3018	4,855	1,20E-06

Análise de Resíduos

Na figura 4.33, os pontos estão distribuídos de modo aleatório. Assim, não podemos afirmar que existe alguma inadequação do modelo. Na figura 4.34, os pontos estão dispostos ao longo da reta. O teste de Anderson-Darling para a normalidade dos resíduos deviance tem um p-valor igual a 0,9084. Assim, podemos considerar que os resíduos deviance estão normalmente distribuídos.

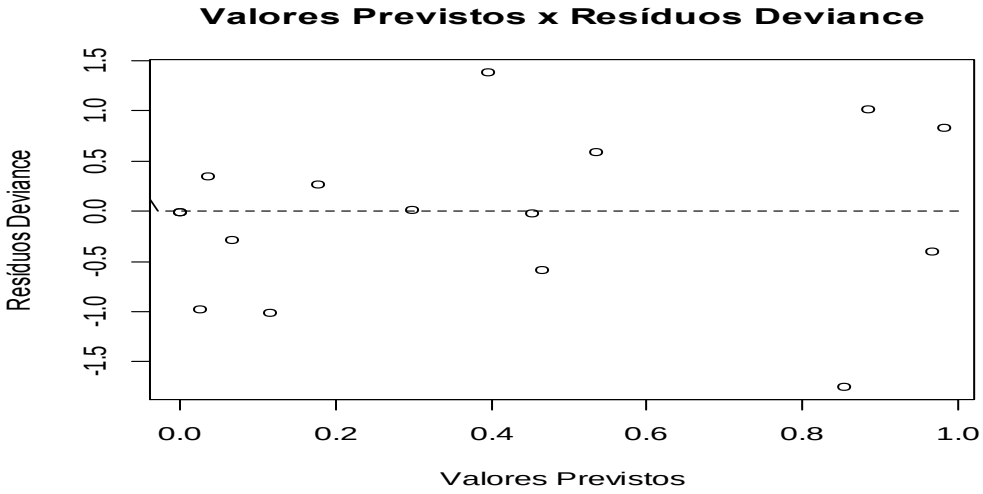

Figura 4.33 – Valores Previstos x Resíduos Deviance

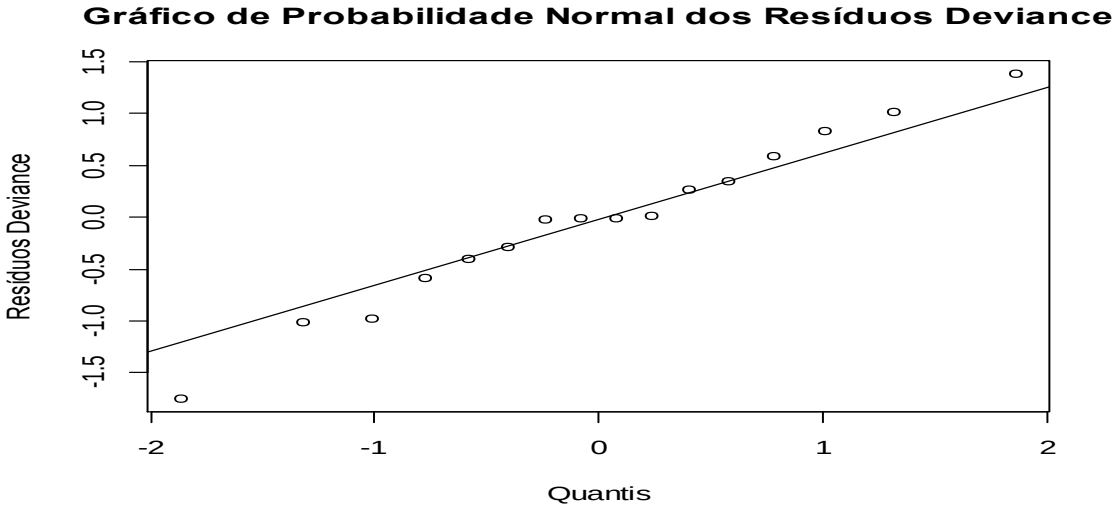

Figura 4.34 – Gráfico de Probabilidade Normal dos Resíduos Deviance

4.4.4. Conclusões

O modelo normal apresenta os mesmos problemas do exemplo 4.1, alguns valores previstos estão fora do intervalo [0,1]. As previsões e os intervalos de confiança para a resposta estimada no modelo obtido pela transformação arco seno não apresenta qualquer incoerência, o mesmo também ocorre no modelo de regressão logística. No entanto, em comparação o exemplo 4.1, os três modelos possuem intervalos de confiança mais amplos e os valores previstos para a resposta não estão tão próximos dos

valores originais, pois o erro quadrático médio aumentou para os três casos. Isso pode ter acontecido devido ao menor número de combinação para cada tratamento.

Neste exemplo, o EQM para o modelo de regressão logística também é o menor, ou seja, em média, este modelo apresentou as previsões mais próximas dos valores originais da resposta.

Tabela 4.37 - Os valores previstos, limites inferiores e limites superiores do intervalo de confiança de 95% para a resposta média dos três casos.

	Modelo Linear Normal			Transf. Arco Seno			MLG		
Resposta	Prev	LI	LS	Prev	LI	LS	Prev	LI	LS
0,6	0,553	0,350	0,757	0,514	0,235	0,789	0,535	0,345	0,716
0,95	0,984	0,781	1,188	0,954	0,763	0,994	0,967	0,795	0,996
0,05	0,028	-0,175	0,232	0,072	0,000	0,287	0,065	0,016	0,235
0,4	0,459	0,256	0,663	0,542	0,259	0,811	0,465	0,284	0,655
0,7	0,678	0,475	0,882	0,580	0,294	0,840	0,853	0,697	0,937
0,3	0,397	0,193	0,600	0,285	0,071	0,571	0,298	0,147	0,510
0	0,153	-0,050	0,357	0,110	0,002	0,348	0,000	0,000	0,000
0	-0,128	-0,332	0,075	0,001	0,065	0,104	0,024	0,006	0,082
0,05	0,191	-0,013	0,394	0,238	0,046	0,517	0,115	0,047	0,255
1	1,034	0,831	1,238	0,994	0,953	0,871	0,983	0,872	0,998
0,05	-0,022	-0,225	0,182	0,000	0,094	0,074	0,035	0,007	0,152
0,95	0,822	0,618	1,025	0,808	0,539	0,975	0,885	0,745	0,953
0,55	0,316	0,112	0,519	0,296	0,078	0,583	0,397	0,230	0,591
0,45	0,447	0,243	0,650	0,568	0,283	0,831	0,452	0,264	0,655
0	0,103	-0,100	0,307	0,002	0,059	0,112	0,000	0,000	0,000
0,2	0,234	0,031	0,438	0,102	0,001	0,337	0,176	0,068	0,386

Tabela 4.38 – EQM

Modelo	Linear Normal	Transf. Arco Seno	MLG
EQM	0,010	0,012	0,004

CAPÍTULO 5 – Considerações Finais

Nesse trabalho, fizemos uma revisão dos conteúdos de Planejamento de Experimentos, especialmente, os modelos fatoriais 2^k, e uma breve introdução aos Modelos Lineares Generalizados. Além disso, abordamos situações em que a teoria dos MLG's é aplicada em planos fatoriais 2^k.

Em cada exemplo trabalhado, foram utilizados três métodos: primeiramente, o modelo linear normal, segundo, uma transformação dos dados, e finalmente, o MLG. Isso foi feito com intuito de comparar os métodos e verificar se os MLG's eram realmente adequados aos problemas propostos. Concluímos que o uso dos MLG's se mostrou eficiente em todos os casos.

Houve situações em que apesar da variável resposta não ser normalmente distribuída, o modelo normal obteve resultados bem próximos aos que foram produzidos pelo MLG. No entanto, em outras situações, observamos incoerências, no exemplo 4.1: valores previstos maiores que 1 ou menores que zero para dados que representam proporções, quando se utilizou um modelo obtido através das transformações dos dados, não foi possível obter o intervalo de confiança da resposta média para alguns valores da variável resposta.

Outras possibilidades de abordagem que podem ser consideradas em trabalhos futuros são estudos de modelos fatoriais com variável resposta que provem de outras distribuições da família exponencial, já que nos exemplos de aplicações, nos mantemos restritos aos modelos: binomial e gama. Além disso, seria interessante utilizar os MLG's em situações que é necessário usar a blocagem.

Referências Bibliográficas

[1] MYERS et al. *Generalized Linear Models: with applications in engineering and the sciences.* 2. Ed. New Jersey: Wiley, 2010.
[2] MONTGOMERY, Douglas C. *Design and Analysis of Experiments* 7. Ed. New York: Wiley, 2009.
[3] AMARAL TURKMAN, M.A; Silva, G.L. *Modelos Lineares Generalizados – da Teoria à Pratica.* Lisboa: SPE, 2000.
[4] MYERS, Raymond H.; MONTGOMERY, Douglas C. A tutorial on Generalized Linear Models. *Journal of Quality Technology*, v. 29, n. 3, p. 274-291, jul. 1997.
[5] COLEMAN, David E.; MONTGOMERY, Douglas C. A systematic approach to planning for a designed industrial experiment. *Technometrics*, v.35, n.1, p.1-12, fev. 1993.
[6] PIERCE, D.A.; SCHAFER, D.W. Residuals in Generalized Linear Models. *Journal of the American Statistical Association*, p. 977-986, 1986.
[7] NELDER, John A; WEDDERBURN, Robert W. Generalized linear models. *Journal of the Royal Statistical Society*, v. 135, n. 3, p. 370–384, 1972.
[8] MCCULLAGH, P.; NELDER, J.A. 1989. *Generalized Linear Models*. Chapman & Hall, Londres, 1989.
[9] BOX, G. E. P.; WILSON, K. B., On the Experimental Attainment of Optimum Conditions, *Journal of the Royal Statistical Society*, Series B, v. 13, p. 1-45, 1951.

Apêndice A

Modelo Linear Normal

O modelo de regressão linear explica a relação entre uma variável resposta *y* e um conjunto de variáveis regressoras. Nesse caso, para *n* observações, o modelo pode ser expresso da seguinte forma.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \text{, ou}$$

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$$

Portanto, a esperança do modelo é

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

Os coeficientes β_i's são encontrados através do método de mínimos quadrados, que minimiza a soma do quadrado dos erros.

$$S = \sum_{i=1}^{n} \varepsilon_i^{\,2}$$

$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) = 0$$

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) x_{ij} = 0$$

$$n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^{\,2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} y_i$$

$$\vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^{\,2} = \sum_{i=1}^{n} x_{ik} y_i$$

Essas são as equações normais dos mínimos quadrados. Pode ser mais fácil de resolvê-la usando a notação matricial.

y= Xβ + ε

onde,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},$$

Estimação dos parâmetros

Observe que X é uma matriz *(n x (k+1))*, β é um vetor de tamanho *k (k x 1)*, *y* e ε também são vetores tamanho n (n x 1).

$$S = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \mathbf{(y - X\beta)'(y - X\beta)}$$

$$S = \mathbf{y'y - \beta'X'y - y'X\beta + \beta'X'X\beta}$$

$$= \mathbf{y'y - 2\beta'X'y + \beta'X'X\beta}$$

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = \mathbf{-2X'y + 2X'X\hat{\beta}}$$

$$\mathbf{X'X\hat{\beta} = X'y \quad (1)}$$

Portanto, o estimador de mínimos quadrados é $\mathbf{\hat{\beta} = (X'X)^{-1}X'y}$.

Expandindo **(1),** temos

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^{\,2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik}^{\,2} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

O ajuste do modelo de regressão será $\mathbf{\hat{y} = X\hat{\beta}}$.

Ou, em notação escalar $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$

A diferença entre os valores observados e os valores ajustado é o resíduo.

$$e_i = y_i - \hat{y}_i$$

$$\mathbf{e = y - \hat{y}}$$

Podemos mostrar que o estimador de mínimos quadrados é não viciado.

$$\mathbf{E(\hat{\beta}) = E[(X'X)^{-1}X'y]}$$

$$\mathbf{= E[(X'X)^{-1}X'(X\beta + \varepsilon)]}$$

$$\mathbf{= E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon]}$$

$$\mathbf{= \beta}$$

Como $\mathbf{E(\varepsilon)=0}$ e $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k-1)}$, $\mathbf{E(\hat{\beta}) = \beta}$

A variância de **β**:

$$\mathbf{Var(\hat{\beta}) = Var[(X'X)^{-1}X'y]}$$
$$\mathbf{= (X'X)^{-1}X'Var(y)X(X'X)^{-1}}$$
$$= \sigma^2 \mathbf{(X'X)^{-1}X'X(X'X)^{-1}}$$
$$= \sigma^2 \mathbf{(X'X)^{-1}}$$

Além de não viciado, o estimador tem de possuir variância mínima. Por isso, precisamos de um estimador para σ^2. A soma dos quadrados dos resíduos é

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \mathbf{(y - X\hat{\beta})'(y - X\hat{\beta})}$$
$$\mathbf{= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}X'X\hat{\beta}}$$
$$SS_{res} = \mathbf{y'y - \hat{\beta}'X'y}$$

O estimador de σ^2 é $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k-1)}$ ou $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k-1)}$

$\mathbf{y'y}$ é a soma de quadrados total e $\mathbf{\hat{\beta}'X'y}$ é a soma de quadrados de regressão.
Reescrevendo a equação, temos $SS_T = SS_R + SS_{res}$

Ao invés de trabalhar simplesmente com o valor dos resíduos, muitos preferem utilizar transformações, como por exemplo, o resíduo padronizado, que possuem média igual a zero e variância aproximadamente igual a 1. As observações fora do intervalo $-3 \le d_i \le 3$ podem ser discrepantes.

$d_i = \dfrac{e_i}{\hat{\sigma}}$, $i = 1,2,...,n$

Onde, $\hat{\sigma} = \sqrt{MS_{res}}$

Como já foi mencionado, $\hat{y} = X\hat{\beta}$. Portanto, $\hat{y} = X(X'X)^{-1}X'y$

$\hat{y} = Hy$

Onde, a matriz $H = X(X'X)^{-1}X'$

Sabemos que $e = y - \hat{y}$

Usando o resultado anterior, temos $e = y - Hy$

Conseqüentemente, $e = (I - H)y$

$$Var(e) = Var[(I - H)y]$$
$$= (I - H)Var(y)(I - H)'$$
$$= \sigma^2 (I - H)$$

Por isso, podemos escrever que $Var(e_i) = \sigma^2(1 - h_{ii})$, h_{ii} é o *i-ésimo* elemento da diagonal principal.

Os resíduos studentizados são $r_i = \dfrac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$

Apêndice B

Detalhes do MLG para ligações canônicas.

Nos casos de ligação canônica, como já foi mostrado no capítulo 3, as equações escore são

$$\frac{1}{a(\phi)}\sum_{i=1}^{n}(y_i-\mu_i)\mathbf{x_i}=0 \quad \text{(I)}$$

As equações escore podem ser solucionadas pelo Método de Mínimos Quadrados Reponderados (IRLS), no qual se utiliza uma aproximação de primeira ordem da série de Taylor.

$$y_i-\mu_i \approx \frac{d\mu_i}{d\eta_i}(\eta_i^*-\eta_i)$$

η_i^* é uma aproximação é uma estimativa inicial do preditor linear. Para ligações canônicas, temos que $\eta_i=\theta_i$, então

$$y_i-\mu_i \approx \frac{d\mu_i}{d\theta_i}(\eta_i^*-\eta_i)$$

$$\eta_i^*-\eta_i \approx (y_i-\mu_i)\frac{d\theta_i}{d\mu_i} \quad \text{(II)}$$

Sendo $Var_\mu=\frac{d\mu_i}{d\theta_i}$, $\eta_i^*-\eta_i \approx \frac{(y_i-\mu_i)}{Var_\mu}$

Consequentemente, $y_i-\mu_i \approx Var_\mu(\eta_i^*-\eta_i)$ (III)

Substituindo (III) em (I), temos que $\frac{1}{a(\phi)}\sum_{i=1}^{n}(\eta_i^*-\eta_i)Var_\mu\mathbf{x_i}=0$ (IV)

Sendo $\mathbf{V}=diag\{Var_\mu\}$, a equação (IV) em notação matricial é

$$\mathbf{y}-\boldsymbol{\mu} \approx \frac{1}{a(\phi)}\mathbf{V}^{-1}.(\boldsymbol{\eta}^*-\boldsymbol{\eta})$$

Se $a(\phi)$ é constante, podemos reescrever as equações escore da seguinte forma

$$\mathbf{X}'(\mathbf{y}-\boldsymbol{\mu})=\mathbf{0}$$
$$\mathbf{X}'\mathbf{V}^{-1}.(\boldsymbol{\eta}^*-\boldsymbol{\eta})=\mathbf{0}$$
$$\mathbf{X}'\mathbf{V}^{-1}.(\boldsymbol{\eta}^*-\mathbf{X}\boldsymbol{\beta})=\mathbf{0}$$

Assim, o estimador de máxima verossimilhança de $\boldsymbol{\beta}$ é

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\eta}^*$$

Não sabemos quem é $\boldsymbol{\eta}^*$, então aplicamos um esquema iterativo baseado em

$$z_i=\hat{\eta}_i+(y_i-\hat{\mu}_i)\frac{d\eta_i}{d\mu_i}$$

Considerando o exemplo da regressão logística, temos:

$$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

$$\frac{d\eta_i}{d\mu_i} = \frac{d\eta_i}{d\pi_i} = \frac{d\ln[\pi_i/(1-\pi_i)]}{d\pi_i}$$

$$= \frac{1}{\pi_i(1-\pi_i)}$$

Assim, $z_i = \hat{\eta}_i + (p_i - \pi_i)\frac{d\eta_i}{d\pi_i} = \hat{\eta}_i + \frac{p_i - \pi_i}{\pi_i(1-\pi_i)}$ e $\mathbf{V} = diag\left(\frac{1}{n_i\pi_i(1-\pi_i)}\right)$

Portanto, o Método de Mínimos Quadrados Reponderados pode ser descrito da seguinte forma.

1. Obtenha uma estimativa inicial de $\boldsymbol{\beta}$, $\boldsymbol{\beta}_0$.
2. Use $\boldsymbol{\beta}_0$ para estimar $\mathbf{V}$ e $\boldsymbol{\mu}$.
3. Faça $\boldsymbol{\eta}_0 = \mathbf{X}\boldsymbol{\beta}_0$.
4. Encontre $\boldsymbol{z}_1$ baseado em $\boldsymbol{\eta}_0$.
5. Obtenha uma nova estimativa $\boldsymbol{\beta}_1$, e continue as iterações até atingir um critério de convergência adequado.

Apêndice C

Teste de Anderson-Darling

Este teste avalia se um determinado conjunto de dados provém de uma determinada distribuição de probabilidade. A estatística do teste é igual a

$$A^2 = -n - S$$

Sendo *n* o número de observações, e $S = \sum_{i=1}^{n} \frac{2i-1}{n}\left[\log F(Y_i) + \log(1 - F(Y_{n+1-i}))\right]$

F é a função de distribuição acumulada. Os valores críticos para o teste são dependentes da distribuição que está sendo testada.

As hipóteses do teste são descritas como:

H_0: Os dados seguem uma determinada distribuição.

H_1: Os dados não seguem uma determinada distribuição.