UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
Programa de Graduação em Estatística

Samuel de Oliveira

INFERÊNCIA E ANÁLISE DE RESÍDUOS E DE DIAGNÓSTICO EM MODELOS LINEARES GENERALIZADOS.

Juiz de Fora
2013

Samuel de Oliveira

INFERÊNCIA E ANÁLISE DE RESÍDUOS E DE DIAGNÓSTICO EM MODELOS LINEARES GENERALIZADOS.

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, com requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador: Clécio da Silva Ferreira

Juiz de Fora
2013

Oliveira, Samuel – Juiz de Fora, 2013

Inferência e Análise de Resíduos e de Diagnóstico em Modelos Lineares Generalizados / Samuel de Oliveira

62.p

Monografia – Universidade Federal de Juiz de Fora e Instituto de Ciências Exatas

Samuel de Oliveira

INFERÊNCIA E ANÁLISE DE RESÍDUOS E DIAGNÓSTICO EM MODELOS LINEARES GENERALIZADOS.

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, com requisito parcial para obtenção do título de Bacharel em Estatística.

Clécio da Silva Ferreira (Orientador) - UFJF

Alfredo Chaoubah - UFJF

Joaquim Henriques Vianna Neto - UFJF

Juiz de Fora, 03 de Maio de 2013

RESUMO

O tema proposto para a realização da monografia é introduzir a classe dos modelos lineares generalizados juntamente com alguns conceitos básicos de regressão linear múltipla. Em seguida, iremos discutir a estimação dos parâmetros, propriedades assintóticas dos estimadores de máxima verossimilhança e a aplicação de alguns testes estatísticos mais conhecidos para a seleção de variáveis e teste global para o ajuste dos modelos. Selecionado o modelo, serão realizados estudos de análise de resíduos e de diagnóstico, utilizando conceitos de pontos de alavanca, influência global, além de ferramentas de seleção de modelos, dentre outros procedimentos. Por fim, aplicaremos a técnica em algumas bases de dados.

Palavras-Chave: MLG, estimação, análise de resíduos e de diagnóstico.

ABSTRACT

The theme proposed for the realization of the monograph is to introduce the class of generalized linear models with some basic concepts of linear regression. Then we will discuss the estimation, asymptotic properties of maximum likelihood estimators and applying some statistical tests most popular for the selection of variables and test the overall fit of the models. Selected model studies will be carried analysis of residuals and diagnosis, using concepts of leverage points, global influence, and selection tools models, among other procedures. Finally, we apply the technique on some databases.

Keywords: GLM estimation, analysis of residuals and diagnostics.

SUMÁRIO

Capítulo 1

Introdução

O modelo de análise de regressão é uma das técnicas estatísticas mais utilizadas nas aplicações em diferentes áreas do conhecimento. No ajuste dos modelos de regressão linear, para relacionar a variável resposta com as variáveis explicativas da matriz $\boldsymbol{X}$, é muito frequente a violação de pressupostos como as hipóteses de linearidade da relação e homocedasticidade das componentes do vetor $\boldsymbol{y}$. Na Seção (2.9) deste trabalho foi estudada a transformação de Box-Cox que tinha o objetivo de resolver estes dois problemas simultaneamente. Nelder e Wedderburn (1972) apresentaram um exemplo, com dados de tuberculose, onde não é possível encontrar um valor para λ, a constante da transformação, que produza linearidade e homocedasticidade ao mesmo tempo. Eles verificaram inclusive que enquanto a transformação raiz quadrada produzia normalidade do erro, a transformação logarítmica era necessária para obter aditividade dos efeitos sistemáticos. Então neste mesmo trabalho, Nelder e Wedderburn desenvolveram uma classe de modelos, generalizando o modelo clássico de regressão linear, conhecidos como *Modelos Lineares Generalizados* (MLGs), também denominados modelos exponenciais lineares, de acordo com Paula (2010), abrindo assim um leque de opções para a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições, bem como dar maior flexibilidade para ligação entre a média da variável resposta e a parte sistemática do modelo, o preditor linear $\boldsymbol{\eta}$. Em MLG, as suposições básicas, tais como, linearidade e homocedasticidade, não são mais exigidas. A idéia básica dos MLGs é transformar as médias dos dados, no lugar de transformar as observações como a técnica de Box e Cox para se obter um modelo de regressão normal linear.

Os Modelos Lineares Generalizados apresentam-se como ferramentas poderosas na análise de dados onde o interesse é o estudo da relação entre uma variável resposta, medida em escala contínua ou discreta, em função das variáveis preditoras, tanto de natureza quantitativas e ou qualitativas. Ocorre em alguns casos que, para se utilizar determinada metodologia de análise, são requeridas algumas pressuposições que nem sempre são atendidas, portanto, o estatístico não pode se omitir sob consequências graves como, por exemplo, valores elevados dos erros e inferências inconsistentes (viesadas). Com o a criação dos MLGs, os problemas com escalas da variável resposta foram reduzidos. Esta metodologia motiva-se como já foi dito anteriormente no sentido de que os efeitos sistemáticos são

linearizados por uma transformação adequada dos valores esperados, permitindo que os valores ajustados permaneçam na escala original.

O objetivo central desta monografia foi realizar uma síntese, com a intenção de consolidar, as técnicas estatísticas mais indicadas para a modelagem estatística, utilizando a princípio o Modelo Linear Clássico por ser a técnica estatística mais difundida para estabelecer a relação entre as variáveis de um experimento e simplicidade, posteriormente usando os Modelos Lineares Generalizados com o enfoque de modelar o que o modelo normal não foi capaz de ajustar. Apresentar e desenvolver a metodologia dos Modelos Lineares Generalizados com a visão de Gilberto Alvarenga Paula e Gauss Moutinho Cordeiro, bem como uma análise de diagnóstico para os modelos em estudo com uma aplicação prática da metodologia utilizada.

Essa monografia é organizada da seguinte forma: o capítulo 1 introduz o assunto que será trabalhado nesta monografia. Os capítulos 2 e 3 fornecem a fundamentação conceitual para a compreensão do capítulo 4, que é a parte de aplicações. Especificamente no capítulo 2 apresento o método de regressão linear clássica, como forma de revisão dos conceitos básicos da metodologia empregada.

No capítulo 3 apresento os Modelos Lineares Generalizados, as distribuições de probabilidade usadas para a variável resposta, a estrutura formal dos MLGs, como é feita a estimação dos coeficientes, os testes de significância dos coeficientes e por último os gráficos indicados para verificar a adequação do modelo. E no capítulo 5 as considerações finais (Conclusão).

Capítulo 2

Modelo Normal Linear

2.1 Definição

Utiliza-se a seguinte notação matricial para a representação do modelo clássico de regressão que no caso é o modelo *normal linear*:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}. \tag{2.1}$$

O modelo clássico de regressão é definido por:

i) respostas y_i independentes (ou pelo menos não correlacionadas) com $i = 1, \ldots, n,$ cada y_i tendo uma distribuição de média $\mu_i = E(y_i)$ e variância σ^2 constante;

ii) a média μ_i é expressa de forma linear como $\mu_i = x_i^T \beta;$ onde x_i^T é um vetor linha de tamanho p com os valores de p variáveis explicativas relacionadas à i-ésima resposta y_i e β é um vetor coluna de tamanho p de parâmetros a serem estimados.

Portanto, utiliza-se a hipótese de aditividade entre $\boldsymbol{y}$ e $\boldsymbol{\mu};$ isto é, $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon};$ onde $\boldsymbol{\epsilon}$ é o vetor de erros de média zero e variância σ^2 constante. Os erros são considerados independentes ou pelos menos não correlacionados. Os efeitos das variáveis explicativas, que formam as colunas da matriz $\boldsymbol{X}$, sobre a variável resposta $\boldsymbol{y}$ são lineares e aditivos. O número de observações n deve ser superior ao número de covariáveis, p, e não deve existir uma correlação significativa entre quaisquer variáveis explicativas. Na formação da matriz modelo, considera-se a primeira coluna como um vetor de 1s sendo o parâmetro β_0 correspondente denominado *intercepto* e as colunas restantes de $\boldsymbol{X}$ é uma matriz com os vetores que multiplicam $\boldsymbol{\beta}$, portanto, são os vetores $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p-1}$. E o número de colunas de $\boldsymbol{X}$ é igual ao número de elementos de $\boldsymbol{\beta}$ e o número de linhas $\boldsymbol{X}$ é o tamanho da amostra.

A suposição de normalidade dos erros é a mais adotada e considera que os erros aleatórios ϵ_i com $i = 1, \ldots, n$ em $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ são não correlacionados e têm distribuição normal $N(0; \sigma^2)$. Como os erros são não correlacionados, pode-se afirmar sob a hipótese de normalidade que estes são independentes. O modelo (2.1) com estas suposições é denominado *Modelo Normal Linear*.

2.2 Estimação

2.2.1 Estimação por Mínimos Quadrados

O objetivo inicial é estimar $\boldsymbol{\beta}$ a partir do vetor $\boldsymbol{y}$ de dados e da matriz modelo $\boldsymbol{X}$ conhecida, suposta de posto completo p. A estimação pelo *Método de Mínimos Quadrados* não requer qualquer hipótese sobre a distribuição das componentes do vetor $\boldsymbol{y}$. Então considere agora o problema da escolha de uma reta de regressão para representar um conjunto de n pontos com coordenadas $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Gostaríamos de obter uma reta que passasse por todos os pontos, para que todos fossem representados por esta reta, isso só seria possível se os *n* pontos do conjunto fossem colineares, mas em geral isto não ocorre, então cometemos erros ao escolher uma reta de regressão para o conjunto de dados por não conseguirmos uma que passe por todos os pontos. As distâncias verticais dos pontos até a reta são chamadas de erros e quanto maior forem estas distâncias, maior será o somatório destes erros, portanto o método de mínimos quadrados consiste em minimizar $\sum_{i=1}^{n} (y_i - \mu_i)^2$ a soma dos quadrados dos erros (para obtermos sempre distâncias com sinal positivo).

A equação da soma de quadrados dos erros $SQE(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mu_i)^2$ correspondente ao modelo (2.1) é dada, em notação matricial, por

$$SQE(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{y} - \boldsymbol{X\beta}). \tag{2.2}$$

Para estimação de β minimiza-se $SQE(\boldsymbol{\beta})$ em relação ao $\boldsymbol{\beta}$, ou seja, minimiza-se o quadrado da distância entre os vetores $\boldsymbol{y}$ e $\boldsymbol{\mu} = \boldsymbol{X\beta}$. A minimização se dá ao derivar $SQE(\boldsymbol{\beta})$ em relação a β_r e igualar a zero o sistema de p equações lineares dadas por

$$\frac{\partial SQE(\boldsymbol{\beta})}{\partial \beta_r} = -2 \sum_{i=1}^{n} x_{ir}(y_i - \mu_i) = 0, \tag{2.3}$$

para $r = 1, \dots, p$. O sistema (2.3) em notação matricial é expresso por $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X\beta}) = 0$. Estas p equações lineares são conhecidas como *Equações Normais*. Como a matriz modelo $\boldsymbol{X}$ tem posto completo, a matriz $\boldsymbol{X}^T\boldsymbol{X}$ é inversível e, portanto, a solução do sistema de equações normais é única. Esta solução corresponde ao *Estimador de Mínimos Quadrados* (EMQ) de $\boldsymbol{\beta}$ dado por

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{2.4}$$

O EMQ $\widehat{\boldsymbol{\beta}}$ em (2.4), segundo o modelo (2.1), tem as seguintes propriedades: $\widehat{\boldsymbol{\beta}}$ é não viesado ($E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$) e $Var(\widehat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{C}_{jj}$, onde $\boldsymbol{C}_{jj}$ é o j-ésimo elemento da diagonal principal da matriz $\boldsymbol{C} = (\boldsymbol{X}^T \boldsymbol{X})^{-1}$. Por fim, $\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})}{(n-p)}$ é o estimador (não viesado) da variância do erro $\boldsymbol{\epsilon}$.

2.2.2 Estimações por Máxima Verossimilhança

Dada uma amostra aleatória de tamanho *n* do modelo de regressão *Normal Linear* em (2.1), sendo $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ o que leva a pensar que $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$, onde $\boldsymbol{I}_n$ é a matriz identidade de ordem *n*.

A função de verossimilhança é igual à densidade conjunta, porém, fazemos os elementos do vetor $\boldsymbol{y}$ fixos e os parâmetros do vetor $\boldsymbol{\beta}$ e σ^2 sendo argumentos da função

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left\{-\frac{1}{2\sigma^2}[\boldsymbol{y} - \boldsymbol{X\beta}]^T[\boldsymbol{y} - \boldsymbol{X\beta}]\right\}.$$

Para definir estimadores de máxima verossimilhança, devemos achar os valores que maximizam a função $L(\boldsymbol{\beta}, \sigma^2)$. Os valores dos $\boldsymbol{\beta}$ e de σ^2 que maximizam $L(\boldsymbol{\beta}, \sigma^2)$ são os mesmos valores que maximizam $l(\boldsymbol{\beta}, \sigma^2) = \ln L(\boldsymbol{\beta}, \sigma^2)$ então

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}[\boldsymbol{y} - \boldsymbol{X\beta}]^T[\boldsymbol{y} - \boldsymbol{X\beta}].$$

A função acima deve ser derivada em relação a cada um dos parâmetros $\boldsymbol{\beta}$ e σ^2, estas derivadas devem ser igualadas a zero e o sistema resultante desta operação deverá ser resolvido da mesma forma que nos mínimos quadrados, levando ao estimador

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Assim, tem-se que $\widehat{\boldsymbol{\beta}} \sim N_p(\beta, \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1})$. Derivando a função $l(\boldsymbol{\beta}, \sigma^2)$ em relação à σ^2 obtemos o seu estimador $\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n}$, que difere do estimador de mínimos quadrados apenas no denominador, n ao invés de $(n - p)$.

2.3 Soma dos Quadrados dos Resíduos

No ajuste de modelos de regressão linear estão presentes erros (desvios) de aproximação associados a cada elemento da amostra. O que isto quer dizer? Que para cada ponto correspondente aos valores da amostra y_i existe uma estimativa $\hat{y}_i$ que pertence à reta de regressão estimada e um valor fixo $\overline{Y}$ que é a média de toda a amostra $\boldsymbol{Y}$. O valor que mede a diferença entre o vetor de observações $\boldsymbol{y}$ e o vetor dos valores ajustados (ou médias ajustadas) $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ é chamado de soma de quadrados dos resíduos (SQR) e é representado na forma matricial por

$$SQR = SQE(\widehat{\boldsymbol{\beta}}) = \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right).$$

Notemos que $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$ onde $\boldsymbol{H}$ é denominada *Matriz Projeção*. As propriedades da matriz $\boldsymbol{H}$ são as seguintes**:** é simétrica, idempotente e tem posto p. Então o vetor $\widehat{\boldsymbol{\beta}}$ que minimiza a distância (2.2) entre os valores de $\boldsymbol{y}$ e $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, segundo Cordeiro e Lima Neto (2006) "é tal que o vetor $\widehat{\boldsymbol{\mu}}$ dos valores ajustados é a projeção ortogonal do vetor de dados $\boldsymbol{y}$ no espaço gerado pelas colunas da matriz $\boldsymbol{X}$." Daí que se origina a terminologia da matriz $\boldsymbol{H}$, *matriz de projeção***.**

2.4 Análise de Variância

A técnica mais utilizada para a verificação da adequação do ajuste do modelo de regressão é a Análise de Variância $(ANOVA)$, que é baseada na soma dos quadrados das diferenças das observações em relação ao seu valor médio, representando dessa maneira uma medida da variabilidade total dos dados, dada pela fórmula

$$SQT = SQRes + SQReg\ ,$$

que na forma matricial fica

$$\boldsymbol{y}^T\boldsymbol{y} - n\overline{\boldsymbol{y}}^2 = \left(\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} - n\overline{\boldsymbol{y}}^2\right) + \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\ , \tag{2.5}$$

onde o termo $SQRes$ é a soma dos quadrados explicada pelo modelo de regressão, enquanto o termo $SQReg$ é a soma de quadrados residual, que não é explicada pelo modelo de regressão. Portanto quanto melhor o ajuste do modelo, maior será a variabilidade explicada por $SQRes$ em relação à variabilidade total SQT do modelo.

Pode-se medir a adequação global do ajuste de um modelo através da comparação de $SQRes$ com SQT, por meio da razão desses dois termos, que é dada por

$$R^2 = \frac{SQRes}{SQT} = \frac{\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} - n\overline{\boldsymbol{y}}^2}{\boldsymbol{y}^T \boldsymbol{y} - n\overline{\boldsymbol{y}}^2}.$$

Esta razão dada por R^2 é denotada de *coeficiente de correlação múltipla de Pearson,* o qual varia entre 0 e 1, e quanto mais próximo de 1 melhor será o ajuste. Porém, tão importante quanto R^2 próximo de 1, é a estimativa de σ^2 ser pequena, por este motivo não devemos escolher o melhor ajuste apenas pelo R^2.

Para a construção da Tabela de Análise de Variância utilizaremos a equação (2.5). Segundo Cordeiro e Lima Neto (2006), para cada soma de quadrados de (2.5) estão associados graus de liberdade, que são obtidos expressando a soma de quadrados correspondente em forma quadrática, cujo posto iguala ao número de graus de liberdade, e a soma dos quadrados $SQRes$ e SQT têm distribuições *Qui-quadrado* com $(p-1)$ e $(n-1)$ graus de liberdade, respectivamente.

A Tabela (2.1) apresenta a Tabela de Análise de Variância usada para testar a hipótese de significância do modelo de regressão, expressado como

$$\begin{cases} H_0: & \boldsymbol{\beta} = \mathbf{0} \\ H_1: & pelo\ menos\ um\ \beta_k \neq 0 \end{cases}$$

Desta forma, se o modelo não for adequado, aceita-se a hipótese nula que consiste em afirmar que o modelo possui todos os parâmetros nulos ($\boldsymbol{\beta} = \mathbf{0}$) e no caso de o modelo ser adequado aceita-se a hipótese alternativa que afirma que pelo menos um parâmetro é não nulo ($\beta_k \neq 0$) . Então testa-se a adequação do modelo ajustado comparando a estatística $F = \frac{MQE}{MQR}$ calculada na Tabela (2.1) com o ponto crítico $F_{(p-1),(n-p),(\alpha)}$ da distribuição *F de Snedecor* com os graus de liberdade $(p-1)\ e\ (n-p)$, ao nível de significância (α). Se o valor do ponto *F* calculado pela tabela for maior que o valor crítico tabelado com os graus de liberdade e nível de significância da distribuição *F* de Snedecor, podemos dizer que ao nível alfa (α) de significância rejeita-se a hipótese nula e aceita-se a hipótese alternativa de que pelo menos umas das variáveis independentes do modelo é significativa para explicar a variabilidade da variável resposta. Caso contrário, não rejeita-se a hipótese nula de que o efeito global destas variáveis para explicar o comportamento da variável dependente não é significativo.

Tabela 2.1:Tabela de Análise de Variância

Efeito	Soma de Quadrados	GL	Média de Quadrados	Estatística
Regressão	$SQRes = \widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} - n\overline{\boldsymbol{y}}^2$	$p-1$	$MQE = SQRes/(p-1)$	$F = MQE/MQR$
Residual	$SQReg = \boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$	$n-p$	$MQR = SQReg/(n-p)$	
Total	$SQT = \boldsymbol{y}^T\boldsymbol{y} - n\overline{\boldsymbol{y}}^2$	$n-1$		

Fonte: Tabela retirada de Paula (2010).

2.5 Seleção das Variáveis Explicativas

Após verificarmos a adequação global dos parâmetros das variáveis explicativas através do teste de hipóteses da $ANOVA$, é fundamental verificar as significâncias de cada variável adicionada ao modelo de regressão, para que este seja o mais parcimonioso contendo apenas variáveis significantes (com real importância para explicar a variabilidade da variável dependente). Portanto, para definirmos quais serão as variáveis explicativas que são significantes, iremos precisar conhecer a distribuição das estimativas dos parâmetros do modelo.

Para o modelo de regressão normal-linear sabemos que $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{C})$ e a estimativa $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ pelo método de mínimos quadrados também possui distribuição normal como visto na seção (2.2). Portanto, como $\widehat{\boldsymbol{\beta}}$ é independente de $\hat{\sigma}^2$, este com distribuição $\frac{(n-p)^{-1}\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$, a estatística de teste T_j com $j = 1, 2, 3, \ldots, p$ tem distribuição t_{n-p} *de Student* com $n-p$ graus de liberdade e é dada pela expressão

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\boldsymbol{C}_{jj}}}. \tag{2.6}$$

Esta estatística permite testar (a hipótese) individualmente para cada variável explicativa, correspondente a cada elemento do vetor $\widehat{\boldsymbol{\beta}}$ que deverá ficar no modelo. Se aplicarmos esta estatística e obtivermos um valor inferior, em módulo, ao valor crítico da distribuição t_{n-p}, não rejeitamos a hipótese nula $\left(H_0: \hat{\beta}_j = 0\right)$. Ou seja, a variável independente não é significativa para explicar a variabilidade da resposta e poderá ser eliminada do modelo. Caso contrário, rejeitamos a hipótese nula e optamos pela hipótese alternativa $\left(H_1: \hat{\beta}_j \neq 0\right)$, isto é, a variável é estatisticamente significante para explicar o comportamento da variável resposta.

2.6 Intervalos de confiança

Considerando a estatística dada por (2.6), um intervalo com $100(1-\alpha)\%$ de confiança para os coeficientes β_j do modelo de regressão é dado por

$$\beta_j \pm t_{\left(\frac{\alpha}{2}, n-p\right)}\sqrt{\hat{\sigma}^2 \boldsymbol{C}_{jj}}\,, \qquad j = 1, \ldots, p.$$

Portanto todos os intervalos de confiança para os coeficientes que conterem o valor zero estes serão considerados estatisticamente não significantes para o modelo, pois este pode assumir o valor zero, assim descartando sua necessidade no modelo.

2.7 Outras técnicas para a seleção e ajuste de variáveis para o modelo

Há uma variedade de procedimentos e critérios para a seleção de um subconjunto de variáveis regressoras para serem incorporadas aos modelos de regressão. Embora nenhum deles seja consistente, e nem sempre métodos diferentes chegam ao mesmo resultado, dado que podemos ter modelos com ajustes equivalentes. Os procedimentos apresentados neste trabalho serão o forward, backward, stepwise e AIC. Alguns desses métodos serão descritos brevemente a seguir.

2.7.1 Método forward

Iniciamos o método pelo modelo $\mu = \beta_0$. Ajustamos então para cada variável explicativa o modelo

$$\boldsymbol{\mu} = \beta_0 + \beta_j \boldsymbol{x}_j\,, (j = 1, \ldots, p-1).$$

Testamos H_0: $\beta_j = 0$ contra H_1: $\beta_j \neq 0$, utilizando a estatística de teste em (2.6). Seja P o menor nível descritivo dentre os $p-1$ testes. Se $P \leq P_E$, a variável correspondente entra no modelo.

Vamos supor que $\boldsymbol{x}_1$ tenha sido escolhido, sem perda de generalidade. Então, no passo seguinte ajustamos os modelos

$$\boldsymbol{\mu} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_j \boldsymbol{x}_j\,, (j = 2, \ldots, p-1).$$

Testamos $H_0: \beta_j = 0$ contra $H_1: \beta_j \neq 0$. Seja P o menor nível descritivo dentre os $(p-2)$ testes. Se $P \leq P_E$, a variável correspondente entra no modelo. Repetimos o procedimento até que ocorra $P > P_E$, então a variável não entrará no modelo (Paula, 2010).

2.7.2 Método backward

Iniciamos o método pelo modelo completo, isto é, com todas as variáveis adicionadas

$$\boldsymbol{\mu} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_{p-1} \boldsymbol{x}_{p-1}.$$

Testamos $H_0: \beta_j = 0$ contra $H_1: \beta_j \neq 0$ para $j = 1, \ldots, p-1$. Seja P o maior nível descritivo dentre os $p-1$ testes. Se $P > P_S$, a variável correspondente sai do modelo. Vamos supor que $\boldsymbol{x}_1$ tenha saído do modelo, sem perda de generalidade. Então, o novo ajuste do modelo fica

$$\boldsymbol{\mu} = \beta_0 + \beta_2 \boldsymbol{x}_2 + \cdots + \beta_{p-1} \boldsymbol{x}_{p-1}.$$

Testamos $H_0: \beta_j = 0$ contra $H_1: \beta_j \neq 0$ para $j = 2, \ldots, p-1$. Seja P o maior nível descritivo dentre os $(p-2)$ testes. Se $P > P_S$, então a variável correspondente sai do modelo. Repetimos o procedimento até que ocorra $P \leq P_S$, então a variável será mantida no modelo (Paula, 2010).

2.7.3 Método stepwise

É a junção dos dois procedimentos anteriores. Iniciamos o processo com o modelo $\mu = \beta_0$. Após duas variáveis terem sido incluídas no modelo, verificamos se a primeira sai ou não do modelo. O processo continua até que nenhuma variável seja retirada, ou seja, incluída no modelo. Geralmente adotamos $0{,}15 \leq P_E, P_S \leq 0{,}25$, outra sugestão seria usar $P_E = P_S = 0{,}20$ (Paula, 2010).

2.7.4 Método de Akaike

Segundo Paula (2010), este método realiza um processo de minimização que não envolve testes estatísticos. A idéia básica é selecionarmos um modelo que seja parcimonioso, ou em outras palavras, que esteja bem ajustado e tenha um número reduzido de parâmetros. Como o logaritmo da função de verossimilhança cresce com o aumento do número de

parâmetros do modelo, uma proposta seria encontrarmos o modelo com menor valor para a função

$$AIC = -L(\widehat{\boldsymbol{\beta}}) + p,$$

em que p denota o número de parâmetros.

No caso do modelo normal linear podemos mostrar que AIC fica expresso, quando σ^2 é desconhecido, na forma

$$AIC = nlog\{D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})/n\} + 2p,$$

em que $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$.

2.8 Análise de Resíduos e Técnicas de Diagnóstico

Quando falamos em técnicas de diagnóstico, logo pensamos em maneiras de descobrirmos problemas relacionados a um indivíduo, neste caso o indivíduo é o modelo de regressão ajustado. Iremos então verificar problemas de ajuste. Segundo Cordeiro e Lima Neto (2006), esses problemas são de três tipos: o primeiro é a presença de pontos mal ajustados, no caso pontos aberrantes; o segundo problema é a violação dos pressupostos para os erros e ou para as estruturas das médias; e por último, o terceiro é a presença de observações influentes.

2.8.1 Matriz de projeção

Voltemos a falar da matriz de projeção $\boldsymbol{H}$ enunciada na seção (2.4) justamente por ser fortemente utilizada nas técnicas de diagnóstico. Os elementos da diagonal principal desta matriz, denotados por h_{ii} mede o quão distante a observação y_i está das demais $n-1$ observações no espaço definido pelas variáveis explicativas do modelo e h_{ii} depende apenas dos valores das variáveis explicativas relacionados à matriz $\boldsymbol{X}$ e não possui relação com os elementos do vetor de observações $\boldsymbol{y}$. Portanto o elemento h_{ii} representa uma *Medida de Alavanca* da i-ésima observação, então se h_{ii} for grande o valor da variável explicativa associado a i-ésima observação será *atípico*, ou seja, estará distante do valor médio da variável explicativa, o que poderá ter influência no cálculo dos coeficientes da regressão.

2.8.2 Resíduos

Uma das técnicas de diagnóstico é a análise de resíduos. O resíduo para a *i*-ésima observação é obtido através da função $r_i = y_i - \hat{\mu}_i$, que mensura a diferença entre o valor observado e o valor ajustado, chamado de resíduo ordinário da variável resposta do modelo. Então podemos afirmar que modelos bem ajustados deverão apresentar pequenos resíduos e caso contrário modelos mal ajustados apresentarão grandes resíduos. De acordo com Cordeiro e Lima Neto (2006), os resíduos ordinários não são muito informativos, por não apresentar variância constante $Var(r_i) = \sigma^2(1 - h_{ii})$, pois depende dos valores de h_{ii}. A solução encontrada é comparar os resíduos de forma padronizada, então obtém-se o *resíduo padronizado* pela expressão

$$r_i^* = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}. \tag{2.7}$$

Caso o modelo de regressão esteja correto todos os resíduos terão a mesma variância e serão adequados para a verificação de normalidade e homocedasticidade (variância constante) dos erros. As observações que possuírem os valores absolutos dos resíduos padronizados maiores que 2 poderão ser considerados pontos aberrantes ou mal ajustados. Segundo Cordeiro e Lima Neto (2006), como o resíduo de cada observação não é independente da variância estimada, não obtemos uma distribuição *t-Student*, como será esperado. O problema da dependência entre r_i e $\hat{\sigma}^2$ pode ser contornado substituindo $\hat{\sigma}^2$ por $\hat{\sigma}_i^2$, o erro quadrático médio correspondente ao modelo sem a *i*-ésima observação. O índice (i) indica que a *i*-ésima observação foi excluída. A expressão do *Resíduo Studentizado* é dada por

$$t_i = \sqrt{\frac{n - p - 1}{n - p - r_i^{*2}}}\, r_i^* \,; \tag{2.8}.$$

t_i tem distribuição *t-Student* com $n - p - 1$ graus de liberdade Cordeiro e Lima Neto (2006). Os resíduos *Studentizados* definidos na equação (2.8) têm a grande vantagem de serem obtidos diretamente da regressão original com todas as observações. Estes resíduos podem ser usados para testar se há diferenças significativas entre os valores ajustados obtidos *com* e *sem* a *i*-ésima observação. E é um teste de hipóteses para verificarmos se o ponto é aberrante, comparando-se o valor absoluto de t_i com o quantil $t_{\left(1-\frac{\alpha}{2}, n-p-1\right)}$. Se observarmos que t_i em módulo é maior, então podemos concluir que o *i*-ésimo ponto é um *outlier*.

2.8.3 Teste para a Hipótese de Normalidade

A validação da hipótese de normalidade pode ser verificada por meio do gráfico dos resíduos ordenados versus os quantis da normal padrão, podendo ser medida pelo cálculo do coeficiente de correlação entre estes que é dado por

$$r_Q = \frac{\sum_{j=i}^{n}(e_j - \bar{e})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^{n}(e_j - \bar{e})^2}\sqrt{\sum_{j=1}^{n}(q_{(j)} - \bar{q})^2}},$$

onde e_j é o *j*-ésimo resíduo padronizado e $q_{(j)}$ é *j*-ésimo quantil da normal. Outra maneira para verificarmos normalidade é aplicando os testes Shapiro e Lilliefors.

2.8.4 Pontos de Alavanca

Segundo Cordeiro e Lima Neto (2006), as propriedades da matriz $\boldsymbol{H}$ citadas na seção (2.4), permitirão fazer afirmações sobre o valor do elemento h_{ii}. Por exemplo, vemos que o seu valor encontra-se no intervalo $\frac{1}{n} \leq h_{ii} \leq 1$. Além disso, pode ser mostrado que podemos dizer que $h_{ii} = \sum_j h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$ e $tr(\boldsymbol{H}) = \sum_i h_{ii} = p$. Se uma observação y_i tem grande alavancagem, o valor de h_{ii} é próximo de 1, implicando que a variância do resíduo correspondente r_i é próxima de zero, pois $Var(r_i) = \sigma^2(1 - h_{ii})$. Logo, o valor médio ajustado $\hat{\mu}_i$ é determinado praticamente pelo valor da observação y_i. Entretanto, como $Var(\hat{\mu}_i) = \hat{\sigma}^2 h_{ii}$, a variabilidade da média ajustada referente à observação y_i é proporcional ao valor de h_{ii}.

Como foi visto na seção (2.8.1), é muito razoável utilizar h_{ii} como uma medida da influência da *i*-ésima observação sobre o próprio valor ajustado. Então, supondo que todos os pontos exerçam a mesma influência sobre os valores ajustados, podemos esperar que h_{ii} esteja próximo de $\frac{p}{n}$. Portanto, convém examinar as observações correspondentes aos maiores valores de h_{ii}. Alguns autores sugerem $h_{ii} \geq \frac{2p}{n}$ como um indicador de pontos de alta alavanca. "Esta regra funciona bem na prática, entretanto, geralmente detecta muitas observações que poderão ser pontos de alavanca ou não. Assim, outras medidas de diagnóstico serão sempre necessárias para confirmar esse primeiro diagnóstico" (Cordeiro e Lima Neto, 2006).

2.8.5 Influência Global

Apresentam-se agora algumas medidas de diagnóstico mais utilizadas na avaliação do grau de dependência entre $\widehat{\boldsymbol{\beta}}$ e cada uma das observações. Começaremos pela distância de *Cook*, obtida através da equação dada por

$$D_i = \frac{h_{ii}}{p(1-h_{ii})} r_i^{*2}; \tag{2.9}$$

Como podemos observar, D_i será grande em duas situações, quando tivermos a medida de alavancagem h_{ii} próxima do valor 1 e quando a medida da discrepância da i-ésima observação dada por r_i^{*2} for grande. Então D_i é chamado de medida de influência de *Cook* para o modelo de regressão.

Cordeiro e Lima Neto (2006) comentam que a medida D_i poderá não ser adequada para os resíduos padronizados grandes e quando h_{ii} for próximo de zero. Neste caso, a variância estimada pode estar inflacionada e não havendo nenhuma compensação por parte de h_{ii}, portanto D_i pode ser pequeno. As observações serão consideradas influentes quando $D_i = F_{p,n-p}\,(0.50)$ e, portanto, é recomendado examinar os efeitos da retirada dessas observações no ajuste do modelo. Como para a maioria das distribuições F, o quantil de 50% é próximo de 1, sugere-se na prática que se o maior valor de D_i for muito inferior a um, então a eliminação de qualquer observação do modelo não irá alterar muito as estimativas dos parâmetros (Cordeiro e Lima Neto, 2006). Para investigar detalhadamente a influência das observações para valores maiores de D_i, o pesquisador terá que eliminar estas observações e recalcular as estimativas dos parâmetros.

Quando a i-ésima observação for identificada como um ponto atípico (baseando-se em r_i^*) ou então como um ponto que exerça forte alavanca (baseando-se em h_{ii}), usa-se o valor de D_i para verificar se esta observação é influente, ou seja, se quando removida do vetor $\boldsymbol{y}$ causará mudanças consideráveis nas estimativas de $\boldsymbol{\beta}$.

2.8.6 Técnicas Gráficas para Diagnóstico

Para detectarmos os três tipos de problemas de diagnóstico citados no inicio da seção (2.8) podem ser utilizados técnicas gráficas. O problema de pontos aberrantes pode ser diagnosticado através do gráfico dos resíduos padronizados r_i^* dados pela equação (2.7) versus a ordem das observações, para detectar as observações mais atípicas; o segundo

problema (é violação dos pressupostos para os erros e ou para as estrutura das médias então segundo Cordeiro e Lima Neto (2006)) pode ser analisado através de um gráfico dos resíduos padronizados r_i^* versus os valores ajustados $\hat{\mu}_i$ e um gráfico de probabilidade dos resíduos padronizados ordenados versus os quantis da distribuição da normal padrão. Então no primeiro gráfico dos resíduos padronizados, os pontos devem estar aleatoriamente distribuídos entre as duas retas $y = -2$ e $y = 2$ paralelas ao eixo horizontal, sem exibir qualquer tendência ou forma definida. Se neste gráfico os pontos exibirem algum padrão, isto poderá ser um indicativo de heterocedasticidade da variância dos erros ou da não-linearidade dos efeitos das variáveis explicativas nas médias das observações. No segundo gráfico, se os pontos ficarem praticamente dispostos sobre uma reta, as observações podem ser consideradas como tendo, aproximadamente, distribuição normal; e por ultimo o terceiro tipo de problema, (presença de observações de alavanca e influência), utilizam os gráficos de h_{ii} e D_i versus a ordem das observações para detectar as possíveis observações influentes.

2.9 Transformação de Box-Cox

Segundo Demétrio e Zocchi (2008), Box e Cox (1964) propuseram um método para a família de transformações potência, fornecendo:

(a) estrutura linear simples;
(b) constância da variância do erro;
(c) independência entre as observações;
(d) normalidade.

Quando a distribuição normal não se adéqua aos dados, muitas vezes é útil aplicar a transformação de Box-Cox para obtermos a normalidade. Considerando $y_1, \ldots, y_n$ os dados originais, a transformação de Box-Cox consiste em encontrar um valor λ tal que os dados transformados $Y_1, \ldots, Y_n$ se aproximem de uma distribuição normal. A transformação potência é modificada para que a variável transformada seja contínua. A expressão obtida é

$$Y_i(\lambda) = \begin{cases} \ln y_i, & para\ \lambda = 0 \\ \dfrac{y_i^{\lambda} - 1}{\lambda}, & para\ \lambda \neq 0 \end{cases}$$

Após aplicarmos essa transformação aos dados, as especificações e os parâmetros do processo (média, variabilidade inerente e total) são obtidos para os dados transformados, aplicando a análise via dados normais. Da mesma forma, os índices são calculados para os

dados transformados com a distribuição normal. Então, $\boldsymbol{Y}(\lambda) = (Y_1(\lambda), \ldots, Y_n(\lambda))$ é um vetor de dimensão $n \times 1$ e $\boldsymbol{Y}(\lambda) \sim N(\boldsymbol{X\beta}, \hat{\sigma}^2 \boldsymbol{I}_n)$ podendo-se ajustar o modelo $\boldsymbol{Y}(\lambda) = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ aos dados transformados.

O método do perfil de máxima verossimilhança de estimação de λ é constituído de três etapas:

1- Arbitram-se valores para λ. Os valores de λ são escolhidos num determinado intervalo. Inicialmente, o intervalo pode ser $\lambda = \{-2, 2\}$;
2- Calcula-se, para cada valor de λ, o máximo da log-verossimilhança, dada pela expressão: (para mais detalhes vide Demétrio e Zocchi (2008)).

$$l_{max}(\lambda) = -\frac{1}{2} n \log(\hat{\sigma}^2) + (\lambda - 1) \sum_{i=1}^{n} \ln Y_i + constante,$$

onde

$$\hat{\sigma}^2 = \frac{1}{n} [\boldsymbol{Y}(\lambda) - \boldsymbol{X\beta}]^T [\boldsymbol{Y}(\lambda) - \boldsymbol{X\beta}].$$

3- Depois de calcular $l_{max}(\lambda)$ para os valores do intervalo, verifica-se se o gráfico de $l_{max}(\lambda)$ versus λ contém o ponto de máximo da curva. Se isto ocorrer, o procedimento está terminado e o valor de λ correspondente ao ponto de máximo é o estimador de máxima verossimilhança de λ. Caso contrário, é necessário ampliar o intervalo de variação dos valores para λ.

Segundo Demétrio e Zocchi (2008) há dois motivos para se obter o intervalo de confiança de $100(1-\alpha)\%$ para λ. Primeiro motivo para verificar se o intervalo contém o valor $\lambda = 1$, indicando não haver necessidade de transformação. O segundo, para identificar se o intervalo cobre algum valor de λ, cuja interpretação seja mais simples. Portanto o intervalo é dado por

$$\left\{\lambda: 2\left[l_{max}(\hat{\lambda}) - l_{max}(\lambda) \leq \chi_1^2(\alpha)\right]\right\},$$

onde $l_{max}(\hat{\lambda})$ é a ordenada correspondente ao ponto de máximo da curva $l_{max}(\lambda)$ versus λ.

Para verificarmos se a transformação foi eficiente basta analisarmos a normalidade dos dados transformados via histograma, papel de probabilidade normal ou teste de normalidade de Kolmogorov-Smirnov ou Shapiro, para mais detalhes vide Demétrio e Zocchi (2008).

Capítulo 3

Modelos Lineares Generalizados

3.1 Modelagem Estatística

Segundo o estatístico George E. P. Box “todos os modelos são errados, mas alguns são úteis”, deixando claro que não podemos aceitar a idéia da existência de apenas um modelo, pois os modelos estatísticos são uma representação simplificada da realidade, no sentido de que o erro sempre existirá, mas a questão é como minimizar estes erros?

A classe de modelos em maior destaque nos últimos anos são os Modelos Lineares Generalizados, que apresentam uma variedade de distribuições para a variável resposta, além da distribuição normal, onde se observa transformações da média através do que é chamado de função de ligação, que faz a conexão da parte regressora à média de uma das distribuições da família exponencial. Na próxima seção as definições para estes modelos serão expostas.

3.2 O Modelo Linear Generalizado e suas componentes

É definido por uma distribuição de probabilidade, pertencente à família exponencial, para a variável resposta, um conjunto de variáveis independentes descrevendo a estrutura linear do modelo e uma função de ligação entre a média da variável resposta e a estrutura linear Cordeiro e Lima Neto (2006).

A formulação de um MLG consiste na escolha de uma distribuição de probabilidade para a variável resposta que deve ser única e pertencer à família exponencial, das variáveis quantitativas e/ou qualitativas para representar a estrutura linear do modelo e de uma função de ligação. Para a melhor escolha desta distribuição de probabilidade é aconselhável realizar uma análise exploratória de dados para observarmos algumas características, tais como: assimetria, natureza discreta ou contínua, intervalo de variação e etc. Os termos que compõem a estrutura linear do modelo podem ser de natureza contínua, qualitativa ou mista, e devem contribuir significativamente na explicação da variável resposta.

Uma importante característica dos MLGs é a suposição de independência, ou pelo menos de não-correlação, entre as observações. Como consequência disso, dados exibindo autocorrelação no tempo, por exemplo, não devem fazer parte do contexto dos MLGs.
O modelo linear generalizado (MLG) é definido a partir das seguintes componentes:

3.2.1 Componente Aleatória

Considere um vetor de observações $\boldsymbol{y} = (y_1, \dots, y_n)^T$ referente às realizações das variáveis aleatórias $Y = (Y_1, \dots, Y_n)^T$ independentes e identicamente distribuídas, com médias $\mu = (\mu_1, \dots, \mu_n)^T$ e pertencentes à família exponencial de distribuições com função de probabilidade dada por

$$f_Y(y; \theta, \phi) = exp\{\phi[y\theta - b(\theta)] + c(y; \phi)\} \tag{3.1}$$

onde $b(\cdot)$ e $c(\cdot)$ são funções conhecidas para cada observação; $\phi > 0$ é denominado *parâmetro de dispersão* e θ é denominado *parâmetro canônico* que caracteriza a distribuição em (3.1). Se ϕ é conhecido, a equação (3.1) representa a família exponencial uni paramétrica indexada por θ.

Escrevendo a log-verossimilhança para uma única observação temos

$$l(\theta, \phi; y) = log\ f_Y(y; \theta, \phi). \tag{3.2}$$

A média e a variância de Y podem ser calculadas respectivamente, das relações abaixo

$$E\left(\frac{\partial l(\theta, \phi; y)}{\partial \theta}\right) = 0 \tag{3.3}$$

e

$$E\left(\frac{\partial^2 l(\theta, \phi; y)}{\partial \theta^2}\right) - E\left[\left(\frac{\partial l(\theta, \phi; y)}{\partial \theta}\right)^2\right] = 0, \tag{3.4}$$

obtendo-se

$$E(Y) = \mu = b'(\theta)\ e\ Var(Y) = b''(\theta)/\phi. \tag{3.5}$$

A variância de Y depende do *parâmetro canônico* θ e pode ser escrita como função de μ, sendo chamada de *função de variância* $V = V(\mu)$. Então iremos chamar $b''(\theta)$ de *função de variância* e denotá-la por

$$V(\mu) = b''(\theta) = \frac{d\mu}{d\theta}. \tag{3.6}$$

A distribuição escolhida da família exponencial depende dos dados em questão, que podem ser discretos, contínuos, assimétricos ou proporções.

Na tabela abaixo é apresentada a associação usual entre a distribuição e o tipo de dado estudado:

Tabela 3.1 - Distribuições e Tipo de Dados

Distribuição	Tipos de Dados
Poisson	Contagens
Binomial Negativa	Contagens
Normal	Contínuos
Gama	Contínuos Positivos
Normal Inversa	Contínuos Positivos

Fonte: Dados retirados de Paula (2010).

A tabela abaixo apresenta as principais características das distribuições da Tabela 3.1:

Tabela 3.2 - Características das principais distribuições utilizadas nos MLGs

Modelos	Normal	Poisson	Binomial	Binomial Negativa	Gama	Normal Inversa
Notação	$N(\mu, \sigma^2)$	$P(\mu)$	$B(n, \mu)$	$BN(\mu, \mathrm{k})$	$\mathrm{G}(\mu, \nu)$	$N^{-}(\mu, \sigma^2)$
Variação de Y	$(-\infty, +\infty)$	0,1,2,..	$0,1, \ldots, n$	1,2..	$(0, \infty)$	$(0, \infty)$
θ	μ	$log\mu$	$log\left(\frac{\mu}{n-\mu}\right)$	$log\left(\frac{\mu}{\mu+k}\right)$	-1/ μ	-1/ $2\mu^2$
$b(\theta)$	$\theta^2/2$	e^{θ}	$nlog(1+e^{\theta})$	$klog\left(\frac{k}{1-e^{\theta}}\right)$	$-log(-\theta)$	$-\sqrt{-2\theta}$
Parâmetro de dispersão ϕ	σ^{-2}	1	1	1	ν^{-1}	ϕ
$\mu(\theta) = \mathrm{E}(\mathrm{Y};\ \theta)$	θ	e^{θ}	$\frac{ne^{\theta}}{1+e^{\theta}}$	$\phi\frac{e^{\theta}}{1-e^{\theta}}$	$-1/\theta$	$(-2\theta)^{\frac{1}{2}}$
Função de Variância $V(\mu)$	1	μ	$\frac{\mu}{n}(n-\mu)$	$\mu\left(\frac{\mu}{\phi}+1\right)$	μ^2	μ^3

* Para a binomial negativa, $\pi = \frac{\mu}{\mu+k}$ é a probabilidade de sucesso em cada tentativa. Para ver $c(y;\phi)$, consulte Cordeiro e Demétrio (2008).

3.2.2 A Componente Sistemática e a Função de Ligação

No MLG a componente sistemática, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$, também chamada de preditor linear, é uma função linear dos parâmetros desconhecidos $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T$ representada por

$$\boldsymbol{\eta} = \boldsymbol{X\beta},$$

onde $\boldsymbol{X}$ é uma matriz modelo $n \times p$ com $(p < n)$ conhecida de posto p. Além disso, outra característica da componente sistemática de um MLG é que a média μ do vetor $\boldsymbol{y}$ é expressa por uma função conhecida (monótona e diferenciável) de $\boldsymbol{\eta}$,

$$\mu_i = \mathrm{g}^{-1}(\eta_i),$$

onde $i = 1, \ldots, n$ e denominando-se $\mathrm{g}(\cdot)$ função de ligação.

O papel da função de ligação é garantir que $\boldsymbol{\beta}$ seja estimado em $\mathbb{R}^p$, ou seja,

$$g: D\mu_i \longmapsto \mathbb{R}, \tag{3.7}$$

onde $D\mu_i$ é o domínio de μ_i.

3.2.3 Funções de Ligação Canônica

Como já foi dito anteriormente, a função de ligação conecta o preditor linear $\boldsymbol{\eta}$ à média $\boldsymbol{\mu}$ do vetor $\boldsymbol{y}$. Para uma determinada distribuição, se a função de ligação é $\theta = \eta$, onde θ é o *parâmetro canônico* definido na Seção (3.2.1), então está garantida a existência de uma estatística suficiente de dimensão igual a $\boldsymbol{\beta}$. Esta função é chamada de *ligação canônica* e tem a vantagem de tornar aditivos os efeitos sistemáticos.

No modelo de regressão clássico usa-se a ligação identidade, porque $\boldsymbol{\eta}$ e $\boldsymbol{\mu}$ assumem valores em toda a reta $(-\infty, +\infty)$. Já no caso da distribuição binomial $(0 < \mu < 1)$ o domínio da função de ligação é o intervalo (0,1) e sua imagem tem que ser o intervalo $(-\infty, +\infty)$. Logo, deve ser utilizada a função logit (ou logística)

$$\eta = log\{\, \mu/(1 - \mu)\}.$$

Se $\boldsymbol{y}$ tem distribuição de Poisson, como $\mu > 0$, a função de ligação adequada é a logarítmica, porque esta tem o domínio positivo e o contradomínio na reta real. Sua expressão é $\eta = log\ \mu$.

Porém, como podem ser observadas na Tabela (3.2), as funções de ligação canônicas para as distribuições gama e normal inversa são funções de contra domínio positivo. Portanto é necessário criar funções que tenham a característica da equação (3.7), que no caso é assumir valores pertencentes ao conjunto dos reais. Para verificarmos outras ligações além das canônicas, implementadas no software livre R vide Apêndice.

Tabela 3.3 – Ligações Canônicas das principais distribuições utilizadas nos MLGs

Modelos	Normal	Poisson	Binomial Negativa	Gama	Normal Inversa
Ligação Canônica $\theta(\mu)$	$\eta = \mu$	$\eta = \log \mu$	$\eta = \log\{\mu/(\mu - k)\}$	$\eta = \mu^{-1}$	$\eta = \mu^{-2}$

Fonte: Dados retirados de Cordeiro e Lima Neto (2006).

3.4 Algoritmo de Estimação dos Parâmetros do MLG

Apesar de existirem outros métodos de estimação para $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$, aqui será apresentado apenas o método da máxima verossimilhança, por ser este um método que apresenta muitas propriedades ótimas, tais como, consistência e eficiência assintótica, sendo este mais preferido e mais utilizado pelos softwares estatísticos.

O algoritmo para o cálculo das estimativas de máxima verossimilhança dos parâmetros $\boldsymbol{\beta}$ foi desenvolvido por Nelder e Wedderburn (1972). A principal diferença em relação aos modelos de regressão é que as equações de máxima verossimilhança são não-lineares. Assim, o estimador é encontrado utilizando um método semelhante ao de *Newton-Raphson* que é o *Método de Escore de Fisher*.

O método consiste em resolver o sistema $U(\beta) = 0$, em que $U(\boldsymbol{\beta})$ é conhecido como função escore ou função suporte e $l(\beta)$ a log-verossimilhança como função de β

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta},$$

além de utilizar a matriz de informação de Fisher

$$K = \left\{-E\left(\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_s}\right)\right\} = -E\left(\frac{\partial U(\beta)}{\partial \beta}\right)$$

Expandindo a função escore em série de Taylor até termos de primeira ordem, obtém-se

$$U\left((\beta)^{(m+1)}\right) = U((\beta)^m) + \frac{\partial U\left((\beta)^{(m)}\right)}{\partial \beta}\left(\beta^{(m+1)} - \beta^{(m)}\right),$$

ou ainda

$$\beta^{(m+1)} = \beta^{(m)} - \left(\frac{\partial U\left((\beta)^{(m)}\right)}{\partial \beta}\right)^{-1} U(\beta)^{(m)},$$

onde o expoente (m) significa o valor do termo na *m*-ésima iteração. Este é o método de *Newton-Raphson* para o cálculo da estimativa de máxima verossimilhança.

Substituindo-se

$$-\frac{\partial U\left((\beta)^{(m)}\right)}{\partial \beta},$$

pelo seu valor esperado K, obtém-se então o método de estimação de Fisher (Fisher, 1925).

Para desenvolver a expressão do algoritmo considera-se o MLG

$$\eta_i = \mathrm{g}(\mu_i) = \sum_{\mathrm{r}=1}^{\mathrm{p}} x_{ir}\beta_r = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

onde $\boldsymbol{x}_i^T$ é a i-ésima linha de $\boldsymbol{X}$ e a log-verossimilhança é dada por

$$l(\beta) = \frac{1}{\phi}\sum_{i=1}^{n}\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i;\ \phi).$$

Derivando $l(\beta)$ em relação ao vetor $\boldsymbol{\beta}$, têm-se

$$U(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\{y_i - b'(\theta_i)\}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}}.$$

Calculando-se

$$\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}},$$

e usando os resultados de (3.5) e (3.6), têm-se, respectivamente,

$$\mu_i = b'(\theta_i)\ e\ \ V(\mu_i) = b''(\theta_i) = \frac{d\mu_i}{d\theta_i}$$

Como $\boldsymbol{x}_i^T$ é a *i*-ésima linha de $\boldsymbol{X}$ e $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ temos

$$\frac{\partial \eta_i}{\partial \beta} = \boldsymbol{x}_i \quad e \quad \frac{\partial \mu_i}{\partial \eta_i} = \left(\mathrm{g}'(\mu_i)\right)^{-1},$$

onde $\boldsymbol{x}_i$ é um vetor coluna $p \times 1$.

E finalmente, a função escore é expressa por

$$U(\beta) = \frac{1}{\phi}\sum_{i=1}^{n}\{y_i - b'(\theta_i)\}\ \frac{1}{V(\mu_i)\mathrm{g}(\mu_i)}\boldsymbol{x}_i.$$

A matriz de informação para $\boldsymbol{\beta}$ é dada por

$$\boldsymbol{K} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X},$$

onde $\boldsymbol{W}$ é uma matriz diagonal de pesos, em cada elemento da diagonal é dado por

$$w_i = \frac{1}{\phi}V_i^{-1}\mathrm{g}'(\mu_i)^{-2}.$$

Então, a função escore usando esta matriz de pesos é dada por

$$U(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}^*,$$

onde $\boldsymbol{z}^*$ é um vetor $n \times 1$ com elementos

$$z_i^* = (y_i - \mu_i)\frac{\partial \mathrm{g}'(\mu_i)}{\partial \mu_i}.$$

Usando estes dois últimos resultados o algoritmo pode ser então expresso por

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left(\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{z}^{*(m)}.$$

Colocando-se $\left(\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\right)^{-1}$ em evidência têm-se

$$\boldsymbol{\beta}^{(m+1)} = \left(\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{y}^{*(m)},$$

onde $\boldsymbol{y}^{*(m)}$ é uma variável resposta modificada dada por

$$\boldsymbol{y}^{*(m)} = \boldsymbol{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{z}^{*(m)}.$$

Assim, conclui-se que o método escore equivale a calcular repetidamente uma regressão linear ponderada entre $\boldsymbol{y}^*$ e $\boldsymbol{X}$ usando mínimos quadrados reponderados, com matriz de pesos $\boldsymbol{W}$ (Paula, 2010). Dessa forma, quanto maior for a variância das observações, menor será seu peso no cálculo das estimativas dos parâmetros. Podemos obter um resultado semelhante com o método de *Newton-Raphson*.

Uma observação segundo Cordeiro e Lima Neto (2006) é que no modelo Binomial com ligação logística, Poisson com ligação logarítmica e Gama com ligação inversa, os dois métodos são idênticos. Entretanto, para estes modelos, os erros padrão das estimativas dos parâmetros são diferentes.

Os programas computacionais de ajustamento do MLG usam o método *Escore de Fisher* para o cálculo da estimativa do $\boldsymbol{\beta}$, pois no método de *Newton-Raphson* existe uma possibilidade de não convergência do algoritmo.

3.5 Adequação do Modelo

3.5.1 A Função Desvio

O objetivo principal é analisar a adequação do modelo como um todo e a realização de uma investigação detalhada quanto às discrepâncias locais que, no caso de serem significativas, podem levar a uma nova escolha do modelo inicialmente proposto. Existem algumas medidas para verificarmos a bondade do ajuste. Uma destas medidas é denominada *Desvio* e equivale à diferença de log-verossimilhanças maximizadas.

Ajustar um modelo estatístico a um determinado conjunto de dados é resumir razoavelmente a informação de n observações para p parâmetros, ou seja, é substituir um conjunto de valores observados y por um conjunto de valores ajustados μ, com um número menor de parâmetros. Porém, o modelo mais simples, chamado de modelo nulo, contém apenas um parâmetro que representa a média μ comum a todas as observações do vetor y. Por outro lado, o modelo saturado contém n parâmetros, um para cada observação. Assim, um modelo adequado tem que resumir os dados parcimoniosamente, de forma que a informação perdida seja não significante. Em termos estatísticos, isto equivale a comparar o modelo ajustado com o saturado e verificar se a discrepância é significativa.

Seja $l(\widehat{\boldsymbol{\mu}};\ \boldsymbol{y})$ o máximo da log-verossimilhança para o modelo em estudo com p parâmetros e $l(y;\ y)$ o modelo saturado com n parâmetros. Segundo Cordeiro e Lima Neto (2006), a estatística definida por Nelder e Wedderburn (1972), com ϕ constante, dada por

$$D^*(\boldsymbol{y};\ \widehat{\boldsymbol{\mu}}) = \phi\, D(\boldsymbol{y};\ \widehat{\boldsymbol{\mu}}) = 2[\, l(\boldsymbol{y};\ \boldsymbol{y}) - l(\widehat{\boldsymbol{\mu}};\ \boldsymbol{y})],$$

é chamada de *desvio* do modelo em investigação, sendo a palavra desvio uma tradução de "*deviance*" feita por Cordeiro (1986), que serve para medir a distância dos valores ajustados

aos dados. Como a distribuição do desvio é desconhecida, na prática, como uma etapa preliminar de verificação, compara-se seu valor $D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ com o valor crítico da $\chi^2_{n-p}(\alpha)$ (qui-quadrado com $n-p$ graus de liberdade e nível de significância α).

Na Tabela 3.4 apresentam-se as formas da função desvio para as principais distribuições da família exponencial.

Tabela 3.4 – Função Desvio para as principais distribuições da família exponencial.

Modelo	Desvio
Normal	$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$
Binomial Negativa	$2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (y_i + k)\log\left(\frac{\hat{\mu}_i + k}{y_i + k}\right)\right\}$
Binomial	$2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right\}$
Poisson	$2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}$
Gama	$2\sum_{i=1}^{n}\left\{\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}\right\}$
Normal Inversa	$\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{(\hat{\mu}_i^2 y_i)}$

Fonte: Dados retirados de Cordeiro e Lima Neto (2006).

3.5.2 Estatística de Pearson Generalizada

A estatística de Pearson generalizada é outra medida importante, definida da seguinte forma

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

onde $V(\hat{\mu}_i)$ é a função de variância estimada para a distribuição de interesse. As duas medidas de Pearson Generalizada e o Desvio têm, considerando modelo normal linear, distribuição χ^2 exata. Resultados assintóticos são possíveis para outras distribuições. Segundo Cordeiro e Lima Neto (2006), a vantagem da função desvio é que ela é aditiva e que acrescentando variáveis explicativas ao modelo, o desvio deve decrescer, diferente da estatística de Pearson generalizada.

3.5.3 Análise de Desvio

Também conhecida como *ANODEV*, é uma generalização da análise de variância para os MLGs, com o propósito de testar modelos encaixados, isto é, cada modelo possui mais termos que os anteriores, tendo a mesma função de ligação e distribuição, com o objetivo de obter os efeitos de fatores, covariáveis e suas possíveis interações. São considerados modelos encaixados ($M_{pr} < M_{ps}$) quando os termos que formam M_{ps} incluem todos os termos que compõem M_{pr} mais outros termos que não estão em M_{pr}.

Então obtendo-se uma sequência de r modelos encaixados $M_{p1} \subset M_{p2} \subset \cdots \subset M_{pr}$, com as seguintes dimensões $p_1 < p_2 < \cdots < p_r$, matrizes $\boldsymbol{X}_{p1}, \boldsymbol{X}_{p2}, \ldots, \boldsymbol{X}_{pr}$ e desvios, $D_{p2}, \ldots, D_{pr}$, todos eles com a mesma distribuição e função de ligação. É bom ressaltar que as desigualdades entre os desvios não são válidas para a estatística de Pearson generalizada. Portanto a comparação de modelos encaixados é realizada, exclusivamente, pela função desvio.

As diferenças entre os desvios $D_{p_i} - D_{p_j}$, $p_i < p_j$ devem ser interpretadas como uma medida de variação dos dados, explicada pelos termos que estão em M_{p_j} e não estão em M_{p_i}. Se a diferença for dada por

$$D_{p_i} - D_{p_j} > \chi^2_{p_j - p_i,}(\alpha), \tag{3.8}$$

consideramos que os termos que estão em M_{p_j} e não estão em M_{p_i} são significativos. Paula (2010) mostra um procedimento para a compreensão da análise de resíduo através de um exemplo de planejamento com dois fatores A e B, com a e b níveis, respectivamente. Ajustam-se sucessivamente, os modelos: primeiro o modelo nulo, e depois o modelo saturado (com todos os fatores e interações possíveis).

Tabela 3.5 – Exemplo de Análise de Desvio com dois fatores A e B.

Modelo	g. l.	Desvio	Diferença	g. l.	Termo
$Constante$	$ab-1$	D_1			
A	$a(b-1)$	D_A	$D_1 - D_A$	$a-1$	$A\ ignorando\ B$
$A+B$	$(a-1)(b-1)$	D_{A+B}	$D_A - D_{A+B}$	$b-1$	$B\ incluído\ A$
$A+B+AB$	0	0	D_{A+B}	$(a-1)(b-1)$	$interação\ AB$ $incluídos\ A\ e\ B$

Fonte: Tabela retirada de Paula (2010).

Paula (2010) apresenta a seguinte ilustração, para o uso das diferenças de desvios para testar hipóteses em modelos encaixados, supondo um MLG com dois fatores, A e B, sendo o fator A com a níveis e o fator B com b níveis. Descrevemos na Tabela 3.5 os possíveis testes envolvendo os dois fatores. Note que, se o interesse é testar a inclusão do fator B dado que o fator A já está no modelo, devemos comparar a diferença $\phi\{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}_A) - D(\boldsymbol{y};\ \widehat{\boldsymbol{\mu}}_{A+B})\}$ com os níveis críticos da distribuição *qui-quadrado* com $(b-1)$ graus de liberdade. Podemos também comparar o valor observado da estatística F correspondente com os níveis da distribuição $F_{(b-1),(n-a-b+1)}$ com $(b-1)$ e $(n-a-b+1)$ graus de liberdade. Para calcular os níveis descritivos das diferenças apresentadas na Tabela 3.5, usamos a inequação (3.8) vista anteriormente para calcular se os termos que estão em M_{p_j} e não estão em M_{p_i} são significativos.

3.5.4 Seleção de Modelos

Os métodos de seleção de modelos descritos no segundo capítulo na Seção (2.7) podem ser estendidos diretamente para os MLGs. Porém, segundo Paula (2010), algumas observações, são necessárias, como nos casos dos modelos de regressão logística e de Poisson, o teste da razão de verossimilhanças, pelo fato de ser obtido pela diferença de duas funções desvio, aparece como o mais indicado. Para os casos de modelagem com regressão normal, normal inversa e gama o teste F, por não exigir a estimativa de máxima verossimilhança do parâmetro de dispersão, é o mais indicado. Isso não impede de utilizar os outros testes. Já o método de Akaike pode ser expresso numa forma mais simples em função do desvio do modelo. Nesse caso, o critério consiste em encontrarmos o modelo tal que a quantidade abaixo seja minimizada

$$AIC\ =\ D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}})\ +\ 2p,$$

em que $D^*(\boldsymbol{y};\widehat{\boldsymbol{\mu}})$ denota o desvio do modelo e p o número de parâmetros.

3.6 Testes de Hipóteses

Os métodos de inferência nos Modelos Lineares Generalizados baseiam-se, na máxima verossimilhança. De acordo com isto, existem três estatísticas para testar hipóteses relativas aos parâmetros do vetor $\boldsymbol{\beta}$, que são deduzidas de distribuições assintóticas de funções adequadas das estimativas dos $\boldsymbol{\beta}$. E estas estatísticas são: Razão de Verossimilhança,

Wald e Escore. Assintoticamente equivalentes e, sob H_0 e para ϕ conhecido, convergem para uma variável com distribuição χ_p^2, sendo, porém, a razão de verossimilhanças, o teste uniformemente mais poderoso.

A razão de verossimilhanças para testar componentes do vetor pode ser obtida como uma diferença de desvios entre modelos encaixados. A estatística de Wald é baseada na distribuição normal assintótica de $\widehat{\boldsymbol{\beta}}$. A estatística escore é obtida da função escore.

Segundo Cordeiro (2006), dependendo da hipótese a ser testada, em particular, qualquer uma dessas três estatísticas pode ser a mais apropriada. Para hipóteses relativas a um único coeficiente β_j, a estatística de Wald é a mais utilizada. Para hipóteses relativas a vários coeficientes, a razão de verossimilhanças é, geralmente, preferida.

3.6.1 Hipóteses simples

As generalizações para os MLGs serão apresentadas a seguir. Vamos supor a seguinte situação de hipóteses simples: $H_0: \boldsymbol{\beta} = \boldsymbol{\beta^0}$ contra $H_1: \boldsymbol{\beta} \neq \boldsymbol{\beta^0}$, em que $\boldsymbol{\beta^0}$ é um vetor de tamanho p conhecido e parâmetro de dispersão também conhecido. Então temos os seguintes testes.

Teste de Wald

Para a situação de hipótese dada anteriormente, a estatística de Wald é definido por

$$\xi_w = \left[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right]^T \phi\left(\boldsymbol{X}^T \widehat{\boldsymbol{W}} \boldsymbol{X}\right)^T \left[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right],$$

onde $\phi\left(\boldsymbol{X}^T \widehat{\boldsymbol{W}} \boldsymbol{X}\right)$ e a $\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right)$. Podemos observar que se o número de parâmetros do vetor $\boldsymbol{\beta}$ for igual a 1, então pode-se escrever o teste de Wald como sendo equivalente ao teste t^2 que é utilizado normalmente no modelo clássico de regressão

$$\xi_w = \frac{\left(\hat{\beta} - \beta^0\right)^T \left(\hat{\beta} - \beta^0\right)}{\widehat{Var}\left(\hat{\beta}\right)}.$$

Segundo (Paula, 2004), um problema com a estatística de Wald é a dependência de ξw com a parametrização usada, quando $\boldsymbol{\eta}\left(\boldsymbol{\beta}\right)$ é não-linear em $\boldsymbol{\beta}$, ou seja, duas formas diferentes e equivalentes para $\boldsymbol{\eta}\left(\boldsymbol{\beta}\right)$, podem levar a diferentes valores de ξ_w.

Teste da razão de verossimilhanças

A estatística de teste, no caso das hipóteses simples, é definida por

$$\xi_{RV} = \phi\{D(\boldsymbol{y};\boldsymbol{\mu}^0) - D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})\},$$

a entra duas funções desvio, em que $\boldsymbol{\mu}^0 = \mathrm{g}^{-1}(\boldsymbol{\eta}^0)$, $\boldsymbol{\eta}^0 = \boldsymbol{X}\boldsymbol{\beta}^0$.

Teste de escore

Conhecido também como teste de Rao, é definido quando $U(\widehat{\boldsymbol{\beta}}) = 0$, por

$$\xi_{SR} = \phi^{-1}U(\boldsymbol{\beta}^0)^T(\boldsymbol{X}^T\boldsymbol{W}_0\boldsymbol{X})^{-1}U(\boldsymbol{\beta}^0),$$

onde $\boldsymbol{W}_0$ é calculado sob a hipótese nula. Segundo Paula (2010), a estatística de escore pode ser muito conveniente em situações em que a hipótese alternativa é bem mais complicada do que a hipótese nula. Nesses casos, somente seria necessário estimar os parâmetros sob H_1 quando o modelo em H_0 fosse rejeitado. Para o modelo de regressão clássico, temos que as estatísticas ξ_{RV} e ξ_w coincidem com ξ_{SR}. Assintoticamente e sob a hipótese nula, tem-se que as estatísticas $\xi_{RV}, \xi_w \; e \; \xi_{SR} \sim \chi^2_p$.

Teste F

A estatística F é dada para o caso de hipóteses simples por

$$F = \frac{\{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}^0) - D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})\}/p}{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})/(n-p)},$$

em que para $\phi \to \infty$ e sob H_0 segue uma distribuição $F_{p,(n-p)}$. Segundo (Paula, 2004) Esse resultado vale também para $n \to \infty$ quando colocamos no denominador da estatística F uma estimativa consistente para ϕ^{-1}. Uma propriedade interessante das estatísticas ξ_{RV}, ξ_{SR} e F é o fato de serem invariantes com reparametrizações, sendo muito útil na construção de intervalos de confiança para os parâmetros. A estatística F não depende do parâmetro de dispersão ϕ^{-1}. Como essa estatística é obtida diretamente de funções desvio, é uma das mais utilizadas na prática.

3.6.2 Modelos encaixados

São considerados modelos encaixados ($M_{pr}\ e\ M_{ps}$) quando os termos que formam M_{ps} incluem todos os termos que compõem M_{pr} e mais outros termos que não estão em M_{pr}. Cada modelo incluindo mais termos que os anteriores, os efeitos de fatores, covariáveis e suas possíveis interações.

Suponha que a seguinte partição $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ e as hipóteses H_0: $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^0$ contra H_1: $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^0$, então teremos

$$\xi_{RV} = \phi\{D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}^0) - D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})\}$$

onde $\widehat{\boldsymbol{\mu}}^0$ é a estimativa de máxima verossimilhança dos MLGs com parte sistemática $\hat{\eta} = \hat{\eta}_1^0 + \hat{\eta}_2$ em que $\hat{\eta}_1^0 = \sum_{j=1}^{q} x_j \beta_j^0$ e $\hat{\eta}_2 = \sum_{j=q+1}^{p} x_j \hat{\beta}_j$. A quantidade de $\hat{\eta}_1^0$ representa um *offset* (parte conhecida no preditor linear). Maiores detalhes vide Paula (2010).

Teste de Wald

Sob a hipótese nula, a estatística de Wald é dada por

$$\xi_w = \left[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0\right]^T \widehat{Var}^{-1}(\widehat{\boldsymbol{\beta}}_1)\left[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0\right],$$

onde $\widehat{\boldsymbol{\beta}}_1$é um vetor de parâmetros da partição $\widehat{\boldsymbol{\beta}}$ e $\widehat{Var}^{-1}(\widehat{\boldsymbol{\beta}}_1) = \phi^{-1}\left[\boldsymbol{X}_1^T \boldsymbol{W}^{1/2} \boldsymbol{M}_2 \boldsymbol{W}^{1/2} \boldsymbol{X}_1\right]^{-1}$. Mais detalhes sobre este teste vide Paula (2010).

Teste escore

A função escore tem a forma $U_\beta = \phi^{1/2} \boldsymbol{X}^T \boldsymbol{W}^{1/2} \boldsymbol{r}_P$, onde $\boldsymbol{r}_P = \phi^{1/2} \boldsymbol{V}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})$ é conhecido como resíduo de Pearson. Segundo Paula (2010), $\boldsymbol{r}_P$ tem a mesma distribuição de $\boldsymbol{Y}$, porém, o valor esperado de $\boldsymbol{r}_P$ é igual a zero e $Var(\boldsymbol{r}_P) = \boldsymbol{I}_p$. O teste de escore é definido por

$$\xi_{SR} = U_{\beta_1}(\widehat{\boldsymbol{\beta}}^0)^T \widehat{Var}_0(\widehat{\boldsymbol{\beta}}_1) U_{\beta_1}(\widehat{\boldsymbol{\beta}}^0),$$

onde $U_{\beta 1}(\boldsymbol{\beta}) = \partial L(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_1 = \phi \boldsymbol{X}_1^T \boldsymbol{W}^{1/2} \boldsymbol{V}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})$, $\widehat{\boldsymbol{\beta}}^0 = (\boldsymbol{\beta}_1^{0T}, \widehat{\boldsymbol{\beta}}_2^{0T})^T$ e $\widehat{\boldsymbol{\beta}}_2^0$ é a estimativa de máxima verossimilhança de $\boldsymbol{\beta}_2$ sob o modelo com parte sistemática $\boldsymbol{\eta} = \boldsymbol{\eta}_1^0 + \widehat{\boldsymbol{\eta}}_2$, isto é, sob H_0, em que $\boldsymbol{\eta}_1^0 = \boldsymbol{X}_1 \boldsymbol{\beta}_1^0$ e $\widehat{\boldsymbol{\eta}}_2 = \boldsymbol{X}_2 \widehat{\boldsymbol{\beta}}_2$.

Segundo Paula (2010), a expressão para $Var(\widehat{\boldsymbol{\beta}}_1)$, realizando algumas álgebras é dada por:

$$Var(\widehat{\boldsymbol{\beta}}_1) = \phi^{-1}(\boldsymbol{R}^T\boldsymbol{W}\boldsymbol{R})^{-1},$$

onde $\boldsymbol{R} = \boldsymbol{X}_1 - \boldsymbol{X}_1\boldsymbol{C}_2$ e $\boldsymbol{C}_2 = (\boldsymbol{X}_2^T\boldsymbol{W}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T\boldsymbol{W}\boldsymbol{X}_1$. Aqui $\boldsymbol{C}_2$ é uma matriz $n \times q$ cuja j-ésima coluna é o vetor de coeficientes da regressão linear (com pesos $\boldsymbol{W}$) da j-ésima coluna de $\boldsymbol{X}_1$ sobre $\boldsymbol{X}_2$. De acordo com (Paula, 2010), $\boldsymbol{R}$ pode ser interpretado como uma matriz $n \times q$ de resíduos e a j-ésima coluna de $\boldsymbol{R}$ corresponde aos resíduos ordinários da regressão linear (com pesos $\boldsymbol{W}$) da j-ésima coluna de $\boldsymbol{X}_1$ sobre $\boldsymbol{X}_2$.

Assim, o teste de escore definido acima fica reescrito na forma

$$\xi_{SR} = \hat{\boldsymbol{r}}_{P_0}^T\widehat{\boldsymbol{W}}_0^{1/2}\,\boldsymbol{X}_1(\widehat{\boldsymbol{R}}_0^T\widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{R}}_0)^{-1}\boldsymbol{X}_1^T\boldsymbol{W}_0^{1/2}\,\hat{\boldsymbol{r}}_{P_0},$$

onde as quantidades $\hat{\boldsymbol{r}}_{P_0}^T$, $\widehat{\boldsymbol{W}}_0$ e $\widehat{\boldsymbol{R}}_0$ são avaliadas em $\widehat{\boldsymbol{\beta}}^0$. (Vide exemplo em Paula, 2010).

3.8 Análise de Resíduos e Técnicas de Diagnóstico

Ao ajustarmos um modelo a um conjunto de dados, uma etapa muito importante é a verificação de possíveis afastamentos das suposições do modelo, levando-se em consideração a parte aleatória e sistemática do modelo, da mesma forma que verificamos a presença de observações com alguma influência fora de padrão nos resultados do ajuste.

Inicialmente, realizamos a análise de resíduos para detectar possíveis pontos extremos e avaliar a adequação da distribuição proposta para a variável resposta. Assim como no modelo clássico de regressão, as técnicas usadas para análise de resíduos e diagnóstico para os modelos lineares generalizados são semelhantes, com algumas adaptações, devido à estrutura dos MLGs.

3.8.1 Resíduos

Os resíduos da modelagem estatística têm um papel muito importante que está relacionada com a qualidade do ajuste, constituindo uma das etapas mais importantes no processo de escolha do modelo adequado. Nos MLGs, segundo Cordeiro e Lima Neto (2006), os resíduos são usados para explorar a adequação do modelo ajustado com respeito à escolha da função de variância, da função de ligação e de termos no preditor linear. Além disso, eles

também são úteis na identificação de pontos aberrantes, que poderão ser influentes ou não. Os resíduos medem discrepâncias entre os valores observados y_i e seus valores ajustados $\hat{\mu}_i$.

3.8.2 Resíduo de Pearson

O resíduo de Pearson tem a seguinte expressão:

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}.$$

A desvantagem deste resíduo é que sua distribuição é, geralmente, bastante assimétrica para modelos não-normais Cordeiro e Lima Neto (2006).

3.8.3 Desvio Residual

O desvio da seção (3.5.1) é usado como uma medida de discrepância de um MLG, obtida através da diferença de log-verossimilhanças maximizadas dos modelos $\tilde{l}_n$ e $\hat{l}_p$, respectivamente, o saturado e o restrito.

Então, cada unidade de *D* contribui com certa quantidade

$$d_i = 2[\, l(\boldsymbol{y};\, \boldsymbol{y}) - l(\hat{\boldsymbol{\mu}};\, \boldsymbol{y})],$$

abrindo a equação acima temos

$$d_i = 2\big(\tilde{l}_n - \hat{l}_p\big) = 2\lambda_i\{y_i\big(\tilde{\theta}_i - \hat{\theta}_i\big) - b\big(\tilde{\theta}_i\big) + \big(\hat{\theta}_i\big)\},$$

tal que $\sum_{i=1}^{n} d_i = D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ e $\lambda_i = 1$ caso mais comum para as principais distribuições da família exponencial. Cordeiro e Lima Neto (2006) afirmam que dessa maneira, surge uma nova definição de resíduo, a partir das componentes d_i que formam o desvio, conhecido como *Desvio Residual*.

Segundo Cordeiro e Lima Neto (2006), o desvio residual é definido como

$$r_{D_i} = sinal(y_i - \hat{\mu}_i)\sqrt{d_i},$$

ao invés de d_i pois, se existe uma transformação que venha a normalizar a distribuição do resíduo, então as raízes quadradas das componentes do desvio são resíduos que possuem as mesmas propriedades impostas por esta transformação Cordeiro e Lima Neto (2006). Desta

forma, os resíduos r_{Di} podem ser considerados como variáveis aleatórias tendo aproximadamente distribuição normal padrão e, consequentemente, $r_{D_i}^2 = d_i$ têm aproximadamente distribuição χ_1^2.

3.8.4 Resíduos Padronizados

Contudo, os resíduos mais utilizados em modelos lineares generalizados são definidos a partir dos componentes da função desvio. A versão padronizada fica dada por

$$t_{D_i} = \frac{d^*(y_i;\hat{\mu}_i)}{\sqrt{1-h_{ii}}} = \frac{\phi^{1/2} d(y_i;\hat{\mu}_i)}{\sqrt{1-h_{ii}}}$$

em que $d(y_i;\hat{\mu}_i) = \pm\sqrt{2\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + (\hat{\theta}_i)\}}^{1/2}$. O sinal de $d(y_i;\hat{\mu}_i)$ é o mesmo de $(y_i - \hat{\mu}_i)$. Segundo Paula (2010), (Williams,1984) verificou através de simulações que a distribuição de t_{D_i} tende a estar mais próxima da normalidade do que as distribuições dos demais resíduos.

3.9 Verificando a Função de Ligação

Segundo Cordeiro e Lima Neto (2006), um procedimento informal para tal verificação consiste na construção de um gráfico entre a variável dependente ajustada e o preditor linear. Então se os dados plotados no gráfico for aproximadamente linear, a função de ligação estará correta. Para dados binários este gráfico é não informativo, sendo necessário o uso de métodos formais. O procedimento adotado por Paula (2010) utiliza técnica gráfica para verificar a adequação da função de ligação. Essa técnica consiste na construção de um gráfico entre a variável z e o preditor linear. Os valores z são dados pela soma do preditor linear e mais os resíduos de Pearson divididos pela raiz quadrada da matriz estimada de pesos (**W**).

Um dos procedimentos formais segundo Cordeiro e Lima Neto (2006) é o método proposto por Hinkley (1985), que consiste em adicionar $\hat{\eta}^2$ como uma nova covariável na matriz modelo. Se isto causar uma redução significativa no desvio, a função de ligação não é adequada. Para verificar se a redução é estatisticamente significante, pode-se utilizar o teste $ANODEV$.

3.10 Verificando a Função de Variância

Segundo Cordeiro e Lima Neto (2006), uma estratégia informal para verificar a adequação da função de variância seria construir um gráfico dos resíduos absolutos versus os valores ajustados. Caso os pontos estejam dispersos sem uma tendência (local ou global) definida, podemos considerar a função de variância adequada. Entretanto, uma tendência positiva indica que a variância está crescendo de acordo com a média. Com isso, segundo Cordeiro e Lima Neto (2006) a escolha inicial de $V(\mu)\ proporcional\ a\ \mu$ pode ser substituída por $V(\mu)\ proporcional\ a\ \mu^2$. Uma tendência negativa indica o efeito contrário.

3.11 Medida de alavancagem

Cordeiro e Lima Neto (2006) dizem que a idéia sobre os pontos de influência e de alavancagem consistem em verificar a dependência do modelo estatístico sobre as várias observações que foram coletadas e ajustadas. Estes pontos exercem um papel fundamental no ajuste final dos parâmetros de um modelo estatístico, ou seja, sua exclusão pode implicar em grandes mudanças dentro das análises estatísticas.

No modelo clássico de regressão uma medida de alavancagem é dada pelos elementos da diagonal da matriz

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T,$$

conhecida como matriz de projeção ou matriz *hat.*

No contexto dos MLGs, as observações conhecidas como pontos de alavancagem podem ser detectadas pelos elementos h_{ii} da matriz *hat generalizada*, definida por

$$\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{W}}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^T\,\widehat{\boldsymbol{W}}\,\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\widehat{\boldsymbol{W}}^{\frac{1}{2}},$$

onde $\widehat{\boldsymbol{W}}$ é o valor de $\boldsymbol{W}$ em $\widehat{\boldsymbol{\beta}}$.

Espera-se que as observações distantes do espaço formado pelas variáveis explicativas apresentem valores apreciáveis de h_{ii}. Como *H* é matriz de projeção, e h_{ii} encontra-se no intervalo $0 \leq h_{ii} \leq 1$*;* Hoalgin e Welsh (1978) sugerem usar $h > 2p/n$ para indicar os pontos de alavancagem. Então uma ferramenta informal, porém muito eficaz para visualizar tais observações, consiste em usar um gráfico indexado dos h_{ii} versus *i* com limite $h = 2p/n$.

3.12 Medidas de influência

Cordeiro e Lima Neto (2006) afirmam que a informação de alavancagem contida em h_{ii} reflete *parcialmente* a influência de uma observação. Para verificarmos a total influência da i-ésima observação, levando-se em consideração aspectos como estimativas dos parâmetros, valores ajustados, estatísticas de bondade de ajuste, etc., torna-se necessário a comparação entre as estimativas $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\beta}}_{(i)}$, esta última obtida quando a i-ésima observação é excluída. Cordeiro e Lima Neto (2006) dizem que a estatística, conhecida como *Distância entre verossimilhanças*, para verificar estas determinadas observações é dada por

$$LD_i = \frac{2}{p}\left[l(\widehat{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}}_{(i)})\right],$$

onde $l(.)$ é a função de log-verossimilhança.

Os autores mostram que, expandindo LD_i em série de Taylor, obtém-se

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \widehat{w}_i^{\frac{1}{2}}(1-h_{ii})^{\frac{1}{2}}\, r_{Pi}\, (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}x_i.$$

Assim, a equação acima pode ser aproximada pela *distância generalizada de Cook*, dada por

$$D_i = \frac{h_{ii}}{p(1-h_{ii})} r_{P_i}^{*2}, \quad com \quad r_{P_i}^{*} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{(V(\hat{\mu}_i)(1-h_{ii}))}},$$

onde p é o posto da matriz modelo $\boldsymbol{X}$ e $r_{P_i}^{*}$ é o resíduo de Pearson padronizado.

Lee (1987) propõe julgar os pontos $D_i > \frac{\chi_p^2(\alpha)}{p}$ como influentes. Uma ferramenta informal para visualizar tais observações é usar um gráfico indexado dos D_i versus i com limite $\frac{\chi_p^2(\alpha)}{p}$. Entretanto, McCullagh e Nelder (1989) propõem medir a influência de uma observação através da *estatística modificada de Cook*, expressa no contexto dos MLGs, por

$$T_i = \left\{\frac{n-p}{p}\frac{h_{ii}}{1-h_{ii}}\right\}^{\frac{1}{2}} \left|r_{D(i)}^2\right|,$$

onde $r_{D_{(i)}}$ é aproximadamente o desvio residual deletado. Aqui, $r_{D(i)}^2$ é definido pela variação no desvio residual causada pela omissão da i-ésima observação. Atkinson (1981) propôs julgar os pontos em que $T_i > 2\sqrt{\frac{p}{n}}$ como influentes.

3.13 Técnicas gráficas

As técnicas gráficas são basicamente iguais às que foram descritas para o modelo linear clássico, sendo os gráficos mais recomendadas para os MLGs segundo Paula (2010): (i) gráficos de t_{Di} contra a ordem das observações, contra os valores ajustados e contra as variáveis explicativas, ou contra o tempo ou alguma ordem em que há suspeita de correlação entre as observações; (ii) gráfico normal de probabilidades para t_{D_i} com envelope, (iii) gráfico de $\hat{z}_i$ contra $\hat{\eta}_i$ para verificarmos a adequação da função de ligação (uma tendência linear indica adequação da ligação) e (iv) gráficos de LD_i, contra a ordem das observações.

Os gráficos normais de probabilidades com envelope destacam-se em dois aspectos: a identificação da distribuição originária dos dados e a identificação de valores que se destacam no conjunto de observações. Os envelopes, no caso dos MLGs com distribuições diferentes da normal, são construídos com os resíduos sendo gerados a partir do modelo ajustado Paula (2010).

Capítulo 4

Aplicações

Neste capítulo aplicarei a metodologia dos Modelos Lineares Generalizados visto nos Capítulos 2 e 3 deste trabalho em duas situações caracterizadas pela natureza dos dados, primeiro analisando modelos para dados discretos na seção 4.1 e depois modelos para dados contínuos serão ajustados na seção 4.2.

Análise de Dados de Contagem.

4.1 Dados de ocorrência de infecções no ouvido.

Esta parte do trabalho tem como objetivo ajustar modelos lineares generalizados para os dados de ocorrência de infecções no ouvido de recrutas. A análise deste conjunto de dados foi proposta como exercício por Paula (2012), para o programa de pós-graduação para a obtenção do título de Mestre e Doutor em Estatística pela Universidade de São Paulo - USP.

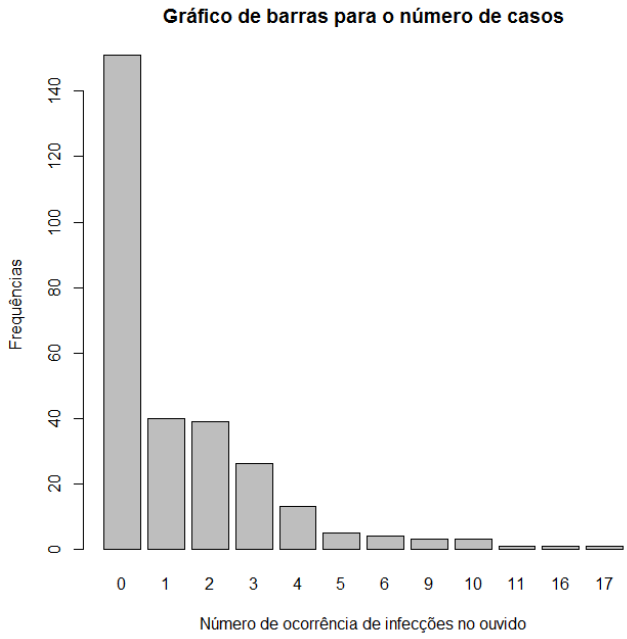
No arquivo **recrutas.txt** são descritos os resultados de um estudo desenvolvido em 1990 com recrutas americanos referente à associação entre o número de infecções de ouvido e alguns fatores. Os dados são apresentados na seguinte ordem: hábito de nadar (ocasional ou frequente), local onde costuma nadar (piscina ou praia), faixa-etária (15-19, 20-25 ou 25-29), sexo (masculino ou feminino) e número de infecções de ouvido diagnosticadas pelo próprio recruta. Verificarei qual dos modelos, normal linear com a variável resposta transformada, Poisson ou com resposta Binomial Negativa, se ajusta melhor aos dados. Serão consideradas apenas interações de 1ª ordem e o método $stepAIC$ do software R que utiliza AIC como critério para selecionar um modelo. Serão realizadas análises de resíduos e diagnóstico por meio de técnicas utilizando gráficos, e por último interpretação dos resultados do modelo selecionado.

4.1.1 Analise Exploratória dos Dados

Segundo Lauretto (2011), o principal papel da Análise Exploratória de Dados (AED) é examinar os dados previamente à aplicação de qualquer técnica estatística. Desta forma o pesquisador consegue um entendimento básico dos dados e das relações existentes entre as variáveis analisadas. A AED extrai informações de um conjunto de dados sem o peso das

suposições de um modelo probabilístico. As técnicas gráficas desempenham um importante papel para esta forma de abordagem.

Gráfico 1 - Gráfico de Barras da variável nºcasos de infecçoes no ouvido em recrutas.

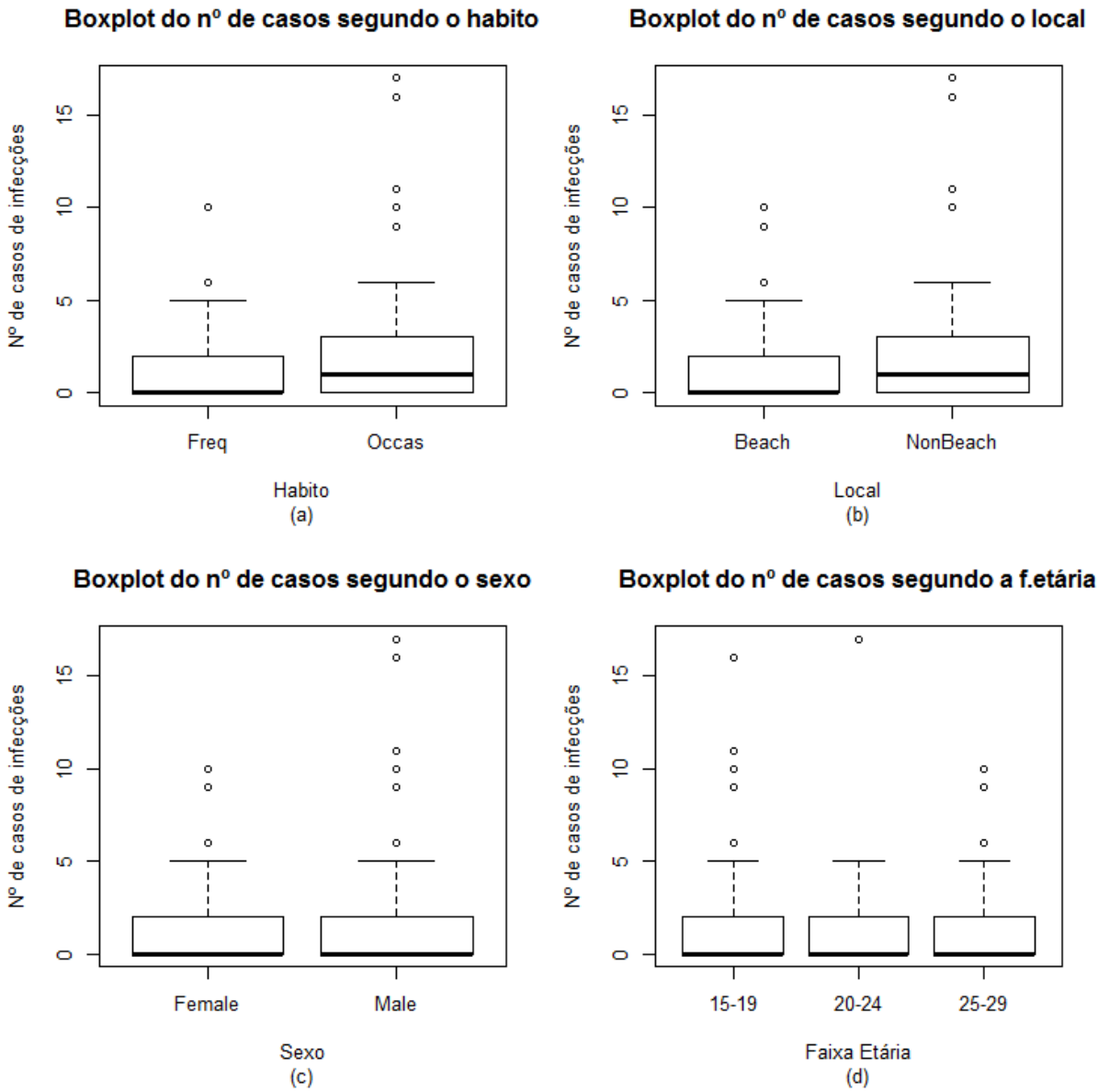


Fonte: Gráfico dos dados, criado pelo autor.

Observando o gráfico de barras da variável resposta número de casos, podemos notar forte assimetria à direita com média igual a 1,38 casos de infecções por recruta e mediana igual a zero, ou seja, a distribuição da variável número de casos pode caracterizar-se por uma distribuição de Poisson.

A seguir serão apresentados alguns gráficos das variáveis que compõem o problema de ocorrência de infecções no ouvido. Começarei analisando o Gráfico Boxplot (2.a) entre a variável resposta segundo o hábito do recruta. Como podemos notar existe uma relação de maior ocorrência de infecções em recrutas que ocasionalmente têm o habito de nadar, presumindo que o hábito frequente de nadar seja um fator de adaptação para as infecções. Observando o gráfico Boxplot (3.b) nota-se que o local onde se nada pode ser responsável por uma maior ocorrência de infecções, pois recrutas que nadam em piscina tendem a ter um número maior de infecções nos ouvidos. Já os gráficos Boxplot (4.c) e Boxplot (5.d) não apresentam uma tendência maior ou menor no número de casos de infecções para o sexo feminino e masculino, e nem as três faixas etárias. Observa-se apenas maior quantidade de pontos discrepantes no número de infecções no sexo masculino e para pessoas com faixa etária entre 15-19 anos.

Gráfico 2 - Gráficos Boxplot da varrável nºcasos de infecçoes no ouvido em recrutas segundo as variáveis de habito, local, sexo e faixa etária.



Fonte: Gráficos Boxplot (a), (b), (c) e (d), utilizando o software R.

4.1.2 Ajuste pelo Modelo Normal Linear

Primeiramente os dados serão ajustados pelo modelo normal, para que posteriormente seja comparado aos modelos lineares generalizados que serão utilizados para ajustar os mesmo dados, observando as vantagens e desvantagens de cada tipo de modelagem.

Como foi apresentada na seção anterior a variável resposta é quantitativa discreta (contagem), o número de casos de infecções no ouvido observado pelos recrutas. Para que possamos ajustar os dados por um modelo normal linear, devemos transformar a variável

resposta, a transformação usada foi o logaritmo da variável *casos* acrescida de $+1$, ou seja, $log(y+1)$.

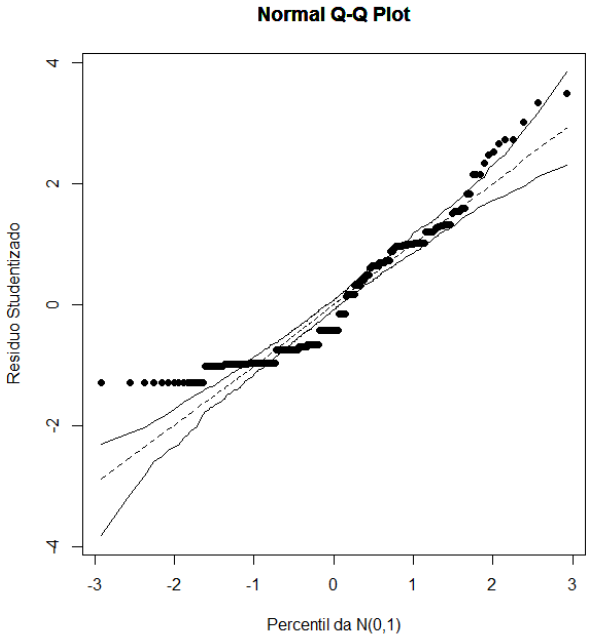
Ajustando o modelo com interações de até 1ª ordem e aplicando o método de seleção de modelos $stepAIC$, obtive um modelo com medida de qualidade de ajuste AIC igual a $600{,}11$, e os seus coeficientes estimados estão dispostos na tabela abaixo.

Tabelas 4.1 – Estimativas dos parâmetros referentes ao modelo normal linear final ajustado com AIC igual a 600,11

Coeficientes	Estimativa	E.Padrão	t value	Pr(>|t|)
Constante	0,44213	0,09580	4,615	**5,97E-06**
habitoOccas	0,21705	0,08031	2,703	**0,0073**
localNonBeach	0,02407	0,14128	0,170	0,8648
sexoMale	-0,16353	0,11358	-1,440	0,1510
localNonBeach:sexoMale	0,34774	0,17289	2,011	**0,0452**

Fonte: Modelagem feita pelo autor.

Gráfico 3 - Gráfico normal de probabilidades para o modelo Normal Linear ajustado.



Fonte: Criado pelo autor através de um script, criado por (Paula, 2010) através do software R.

O gráfico de probabilidade normal com envelopes simulados para um ajuste da distribuição normal representado pelo Gráfico (3) mostra que o modelo não se ajustou bem aos dados, pois há muitos pontos (resíduos) fora das bandas de confiança. Portanto há fortes indícios de que a distribuição normal não é adequada para ajustarmos os dados. Na próxima

seção irei ajustar os dados aos modelos lineares generalizados considerando distribuições mais apropriadas para a variável resposta.

4.1.3 Ajuste pelos Modelos Lineares Generalizados

A partir da análise exploratória, podemos verificar que a variável resposta número de casos de infecções no ouvido é uma variável quantitativa discreta, tratando-se de uma variável de contagem, para iniciarmos a escolha de um modelo linear generalizado adequado. Começamos com a seguinte pergunta: Qual distribuição da família exponencial a variável resposta pertence? Podemos supor que esta tenha distribuição de Poisson ou de uma Binomial Negativa. São as duas principais distribuições para variáveis de contagem.

Começamos ajustando os dados supondo que a variável resposta segue uma distribuição de Poisson. Denotarei por Y_{ijklm} o número de casos de infecções no ouvido num determinado período de tempo do *m*-ésimo recruta, com o *i*-ésimo habito de nadar, pertencente ao *j*-ésimo sexo, no *k*-ésimo local de natação e com a *l*-ésima faixa etária, em que i, j, k = 1, 2, l =1, 2, 3 e m = 1,... , 287. Vamos supor que $Y_{ijklm} \sim \mathrm{P}(\mu_{ijkl})$ com parte sistemática dada por:

$$\mu_{ijkl} = \alpha + \beta_i + \gamma_j + \delta_k + \rho_l + \beta_i * \gamma_j + \beta_i * \delta_k + \beta_i * \rho_l + \gamma_j * \delta_k + \gamma_j * \rho_l + \delta_k * \rho_l$$

com $\beta_1 = 0, \gamma_1 = 0, \delta_1 = 0$ e $\rho_1 = 0$. Assim temos um elemento casela de referência em β_2, γ_2 é a diferença entre os efeitos do sexo masculino e do sexo feminino, δ_2 é a diferença entre os efeitos do local de natação na praia com relação à piscina e por ultimo ρ_2 e ρ_3 denotam os incrementos da faixa etária de 20-24 anos e de 25-29 anos, respectivamente em relação à faixa de 15-19 anos.

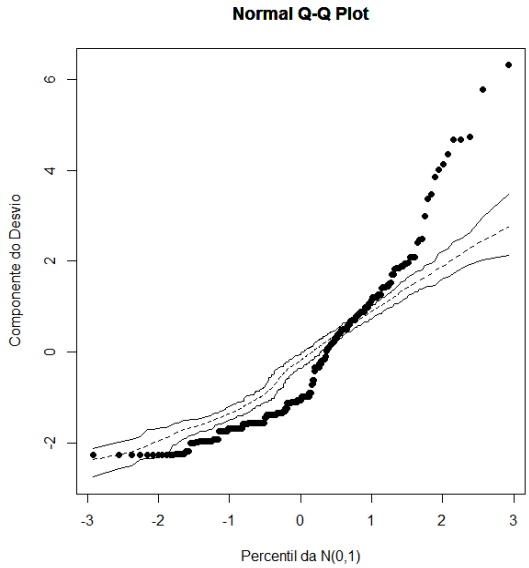
Após ajustar o modelo de Poisson com todas as funções de ligação implementadas no software R, e preditores com interações de até 1ª ordem, foi aplicando o método de seleção de modelos $stepAIC$, obtendo o modelo com ligação logarítmica com a menor medida de ajuste AIC igual a $(1128{,}6)$, e os seus coeficientes estimados estão dispostos na tabela abaixo.

Tabela 4.2 – Estimativas dos parâmetros referentes ao modelo log-linear Poisson final ajustado com AIC igual 1128,6

Coeficientes	Estimativa	E.Padrão	z	value Pr(>|z|)
Constante	0.38000	0.16883	2.251	**0.024400**
habitoOccas	0.23238	0.18856	1.232	0.217802
localNonBeach	-0.37990	0.25783	-1.473	0.140634
fetaria20-24	-0.88047	0.28876	-3.049	**0.002295**
fetaria25-29	-0.69793	0.27725	-2.517	**0.011825**
sexoMale	-0.45759	0.16279	-2.811	**0.004941**
habitoOccas:localNonBeach	0.36745	0.22166	1.658	0.097375
habitoOccas:fetaria20-24	0.09116	0.26124	0.349	0.727122
habitoOccas:fetaria25-29	0.70656	0.29294	2.412	**0.015867**
localNonBeach:fetaria20 -24	0.74090	0.29358	2.524	**0.011612**
localNonBeach:fetaria25-29	0.15594	0.26471	0.589	0.555802
localNonBeach:sexoMale	0.77640	0.23443	3.312	**0.000927**

Fonte: Modelagem feita pelo autor.

Gráfico 4 - Gráfico normal de probabilidades para o modelo log-linear Poisson ajustado aos dados sobre as características dos recrutas.



Fonte: Criado pelo autor através de um script, criado por (Paula, 2010) através do software R.

O gráfico de probabilidade normal com envelopes simulados para um ajuste da distribuição Poisson com ligação logarítmica representado pelo Gráfico (4) mostra que o modelo não se ajustou bem aos dados, pois há muitos pontos (resíduos) fora das bandas de confiança. Portanto há fortes indícios de que a distribuição de Poisson não é apropriada para ajustarmos os dados.

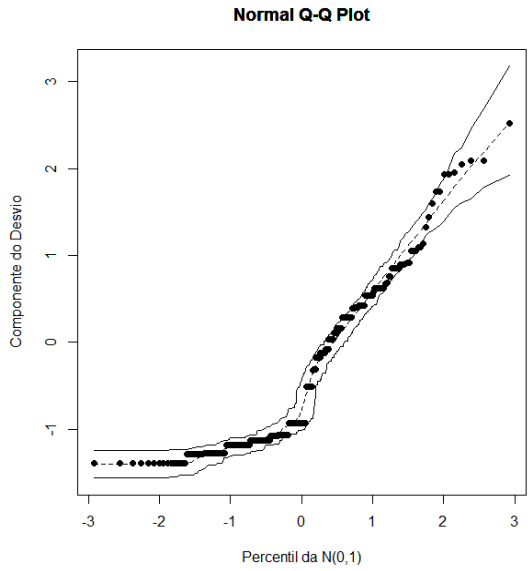
Após ajustar o modelo de Poisson com todas as funções de ligação, irei agora ajustar o modelo com distribuição Binomial Negativa e preditores com interações de até 1ª ordem e aplicando o método de seleção de modelos $stepAIC$, obtendo o modelo com ligação logarítmica com a menor medida de ajuste AIC igual a $(903{,}1)$, e os seus coeficientes estimados dispostos na tabela abaixo.

Tabela 4.3 – Estimativas dos parâmetros referentes ao modelo log-linear Binomial Negativa, final ajustado com AIC igual 903,1

Coeficientes	Estimativa	E.Padrão	z	Pr(>|z|)
Constante	-0.064376	0.228550	-0.282	0.77819
habitoOccas	0.593365	0.189500	3.131	**0.00174**
localNonBeach	0.007495	0.330514	0.023	0.98191
sexoMale	-0.407473	0.274556	-1.484	0.13778
localNonBeach:sexoMale	0.745367	0.407776	1.828	**0.06757**

Fonte: Dados análise do autor.

Gráfico 5 - Gráfico normal de probabilidades referente ao modelo log-linear Binomial Negativa ajustado aos dados sobre as características dos recrutas.



Fonte: Criado pelo autor através de um script, criado por (Paula, 2010) através do software R.

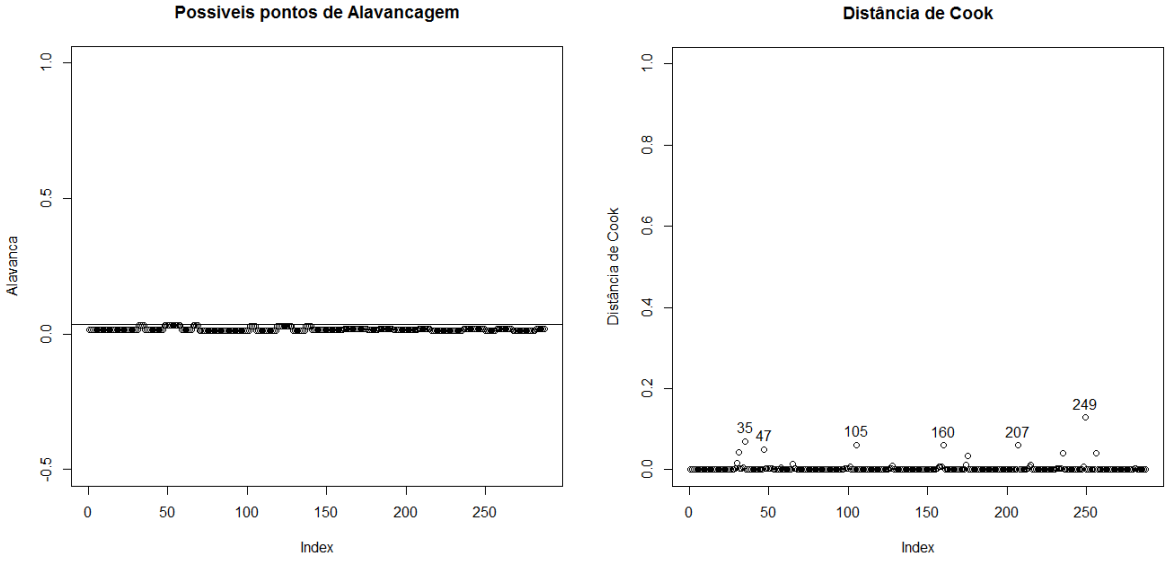
O gráfico de probabilidade normal com envelopes simulados para um ajuste da distribuição Binomial Negativa representado pelo Gráfico (5) mostra que o modelo se ajustou muito bem aos dados, pois todos os pontos (resíduos) dentro ou sobre as bandas de confiança. Portanto há fortes indícios de que a distribuição binomial negativa é adequada para ajustarmos os dados.

Nota-se neste exemplo a superioridade do modelo log-linear Binomial Negativa quando comparado aos outros dois modelos, Normal Linear e o modelo log-linear Poisson. Essa vantagem se reflete não somente pela qualidade do ajuste que pode ser verificada pelo gráfico de envelope, mas também na interpretação dos parâmetros em relação ao modelo normal linear, uma vez que a escala da variável resposta foi preservada.

Verificamos também, neste estudo, como fica o ajuste através de um modelo log-linear de Poisson. Observando os gráficos normais de probabilidades para os dois ajustes nos Gráficos (4 e 5) e notamos uma clara superioridade do modelo log-linear com resposta binomial negativa. O modelo log-linear de Poisson apresenta fortes indícios de sobredispersão com os resíduos cruzando o envelope gerado. Isso é justificado pelo valor do desvio

$$D^*(\boldsymbol{y};\, \widehat{\boldsymbol{\mu}}) \;=\; 732{,}16\ (275\ \mathit{graus\ de\ liberdade}).$$

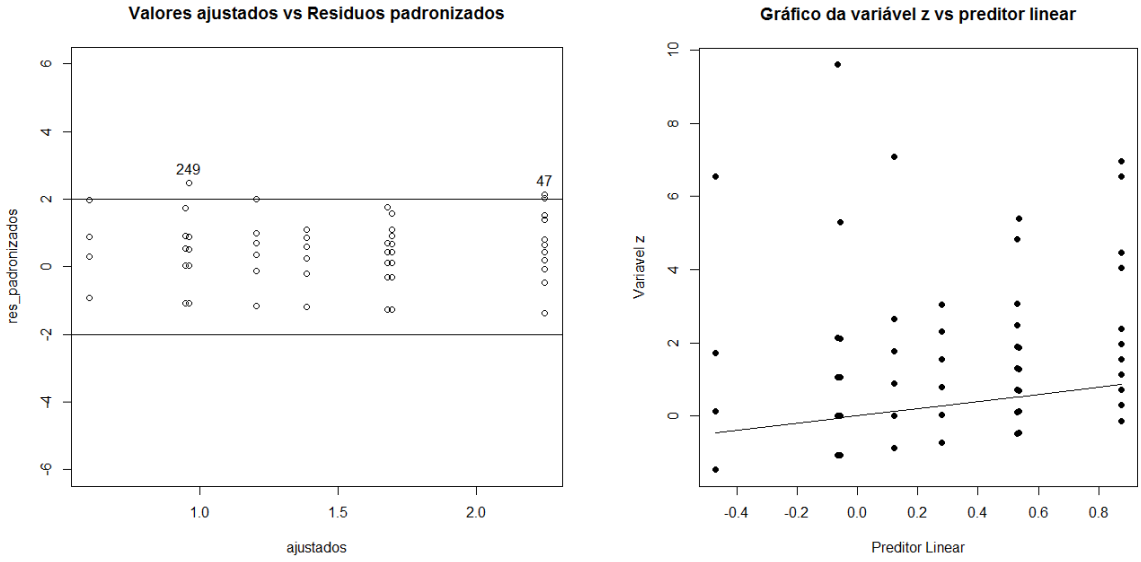
Gráficos 6 – Gráfico para verificação de pontos de alavanca (a) e Gráfico para verificação de pontos de influência (b).



(a) (b)

Fonte: Criado pelo autor através de um script, criado por (Ferreira, 2010) notas de aula de MLG.

Gráficos 7 – Gráfico para verificação de pontos aberrantes e Gráfico para verificação de adequação da função de ligação.



(a) (b)

Fonte: Criado pelo autor através de um script, criado por Ferreira (2010) notas de aula de MLG.

4.1.4 Diagnóstico do Modelo Selecionado.

São apresentados alguns gráficos de diagnóstico (Gráficos 6 e 7). No Gráfico (6.a) em que são apresentados os valores de h_{ii}, nenhum dos pontos são destacados como alavanca. Já pelo Gráfico (6.b), notamos pelo menos um ponto com mais destaque como influente em $\widehat{\boldsymbol{\beta}}$, os recrutas #35, #47, #105, #160, #207 e #249. Os seis recrutas têm vários casos de ocorrência de infecção no ouvido maior ou igual a nove vezes e ocasionalmente têm o hábito de nadar, apenas o recruta #249 tem hábito frequente de nadar, mas apresenta como sendo o ponto com maior influência devido ao fato de ter registrado 17 infecções no ouvido. Pelo Gráfico (7.a), notamos dois pontos com mais destaque como aberrantes, #47 e #249. Esses recrutas tiveram um número alto de ocorrências de infecções, um é mulher (#47) e o outro (#249) é homem. Em geral os pontos aberrantes desse exemplo referem-se a recrutas com número elevado de infecções no ouvido. Finalmente, o Gráfico (7.b) indica que a escolha da ligação logarítmica parece não ser inadequada.

4.1.5 Interpretação do modelo final

Para os Modelos Lineares Generalizados com função de ligação logarítmica ou genericamente chamados de modelos log-lineares neste trabalho, apresentam a mesma forma

de interpretação dos coeficientes estimados, independente da distribuição adotada para a variável resposta. De forma geral temos um modelo com ligação logarítmica dado por

$$log(\hat{\mu}_{x_i}) = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$$
$$\hat{y}_{x_i} = e^{\hat{\alpha}+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}+\cdots+\hat{\beta}_p x_{ip}}$$

Então, se adicionarmos uma unidade a variável x_{i1} dado que as demais variáveis estão fixadas temos

$$\hat{y}_{(x_{i1}+1)} = e^{\hat{\alpha} + \hat{\beta}_1(x_{i1}+1) + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}}$$
$$= e^{\hat{\beta}_1} e^{\hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}}$$

Teremos o efeito dessa adição dada por

$$\frac{\hat{y}_{(x_{i1}+1)}}{\hat{y}_{x_i}} = e^{\hat{\beta}_1}.$$

Portanto, o modelo final ajustado log-linear Binomial Negativa fica

$$\hat{\boldsymbol{y}} = e^{-0{,}064376+0{,}593365*\boldsymbol{habito}+0{,}007495*\boldsymbol{local}-0{,}407473*\boldsymbol{sexo}+0{,}745367*\boldsymbol{local}*\boldsymbol{sexo}}$$

Desse modelo podemos extrair a seguinte interpretação: $e^{\hat{\beta}_2} = e^{0{,}593365} = 1{,}81$ (81%) é o aumento relativo esperado do número de casos de infecções no ouvido se for observado que o recruta tem o hábito ocasional de nadar. Observando a interação entre local e sexo pode se concluir que recrutas do sexo masculino que utilizam piscina como local de natação estão propensos a um aumento de $e^{0{,}745367} = 2{,}10$ (110%) no número de casos de infecções.

Análise de Dados Contínuos

4.2 Dados de experimento com filme para máquinas fotográficas.

A análise deste conjunto de dados, como na Seção **4.1**, foi proposta como exercício por Paula (2010). A fim de avaliar a qualidade de um determinado filme utilizado em máquinas fotográficas, o tempo de duração do filme (em horas) é relacionado com a densidade do filme sob três condições experimentais descritos na tabela 4.4 e contidos no arquivo **dfilme.txt**.

Tabela 4.4 – Dados de experimentação com filme de máquinas fotográficas.

Dados do Experimento					
Tempo	Dmax (72°C)	Tempo	Dmax (82°C)	Tempo	Dmax (92°C)
72	3,55	48	3,52	24	3,46
144	3,27	96	3,35	48	2,91
216	2,89	144	2,50	72	2,27
288	2,55	192	2,10	96	1,49
360	2,34	240	1,90	120	1,20
432	2,14	288	1,47	144	1,04
504	1,77	336	1,19	168	0,65

Fonte: Criado pelo autor com dados extraídos de (Paula, 2010, p. 163).

4.2.1 Analise Exploratória dos Dados

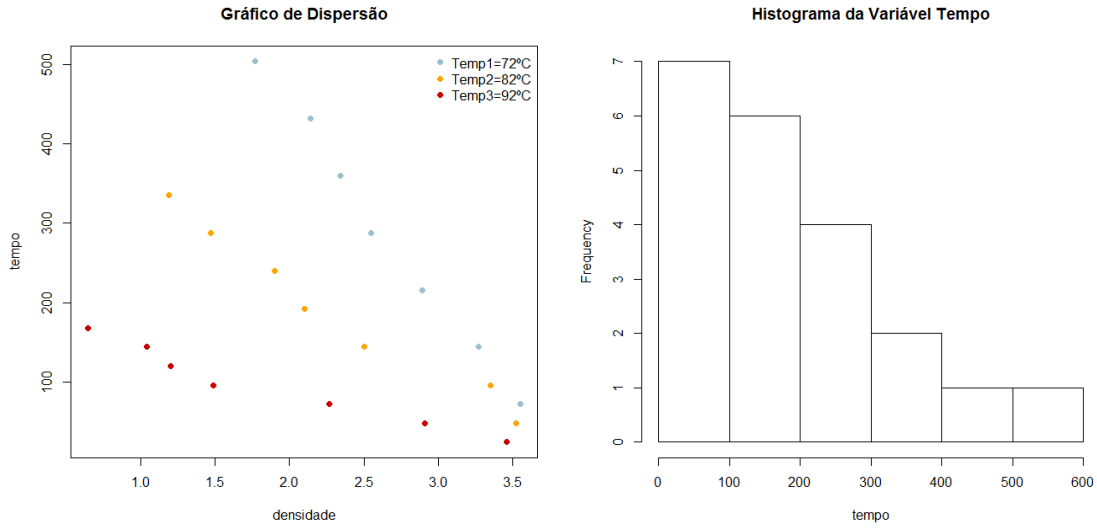
Segundo Cordeiro e Lima Neto (2006), a formulação de um MLG compreende a escolha de uma distribuição de probabilidade para a variável resposta que deve pertencer a uma família exponencial, variáveis explicativas quantitativas e/ou qualitativas para representar a estrutura linear do modelo e de uma função de ligação. Para a melhor escolha da distribuição de probabilidade será necessário realizar uma analise exploratória dos dados para observar características, tais como: assimetria, tipo de dado discreto ou contínuo, variabilidade, tendência central e etc. E as variáveis explicativas que compõem a estrutura linear do modelo devem dar uma contribuição significativa na explicação da variável resposta.

A seguir serão apresentados alguns gráficos das variáveis que compõem o problema de avaliação da qualidade de filmes para máquinas fotográficas. Começarei analisando o Gráfico de Dispersão entre a variável resposta e a densidade máxima do filme sob as três condições experimentais de temperatura (Gráfico 8.a). Como podemos notar existe, uma relação negativa entre densidade e temperatura para com a variável resposta tempo de duração do filme. Observamos que quanto maior a densidade máxima do filme menor é o tempo de

duração deste, levando a temperatura em consideração, observa-se que quanto maior é a temperatura, menor é o tempo de duração do filme. Portanto ao realizarmos um modelo de regressão, os coeficientes relacionados à densidade e temperatura, ambos deverão apresentar sinal negativo.

Observando o histograma da variável resposta tempo (Gráfico 8.b), podemos notar que existe assimetria à direita e assumi apenas valores positivos, ou seja, a distribuição da variável tempo pode caracterizar-se por uma distribuição Gama.

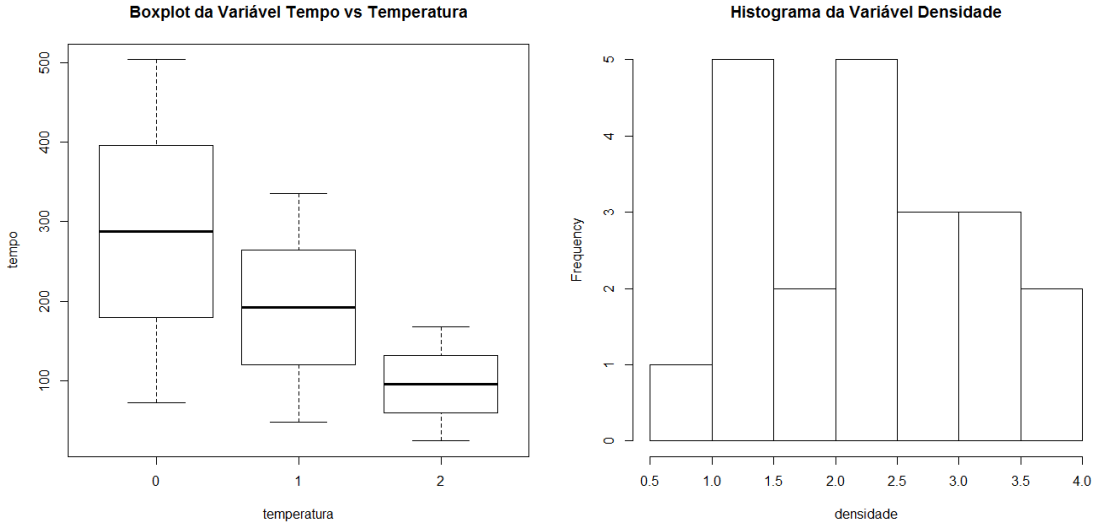
Gráficos 8 - Gráfico de Dispersão da variável tempo por densidade máxima sob três condiçoes de temperatura (a) e o Histograma da variável resposta tempo (b).



(a) (b)

Fonte: Gráficos de dispersão e o histograma criados pelo autor, utilizando o software R.

Gráficos 9 - Gráfico Boxplot da variável resposta tempo pelas três condições de temperatura (a) e o Histograma da variavel densidade máxima (b).



(a) (b)

Fonte: Gráficos boxplot e histograma criados pelo autor, utilizando o software R.

Observando o gráfico de boxplot da variável tempo em função da temperatura, (Gráfico 9.a), onde (0, 1, 2) correspondem às temperaturas (72ºC, 82ºC, 92ºC) respectivamente, pode se concluir que quanto menor a temperatura, maior a duração do filme, conclusão tomada com base na leitura da mediana apresentada em cada boxplot. Por último, o histograma da variável densidade (Gráfico 9.b) apresenta certa simetria, destacando-se apenas dois intervalos o primeiro com densidade entre (1,0 – 1,5) e outro entre (2,0 – 2,5) com mais de 3 observações para cada um.

Através da tabela abaixo observar-se que as suposições feitas com base nos gráficos se confirmam.

Tabela 4.5 – Estatísticas Descritivas das Variáveis Tempo e Densidade

Estatísticas	Tempo	Densidade
Mínimo	24	0,65
Máximo	504	3,55
1st Quartil	96	1,49
Mediana	144	2,27
Média	192	2,265
3rd Quartil	288	2,91
Des. Padrão	133,19	0,89
Coef.Variação	69,37%	39,29%

Fonte: Dados da análise do autor

4.2.2 Ajuste pelo Modelo Normal Linear

Proponho inicialmente um modelo normal linear em que Y denota o tempo e X a densidade máxima do filme sob as condições de experimentais de temperatura. O modelo fica, portanto dado por

$$y_{ij} = \alpha + \beta x_i + \gamma_j + \epsilon_i$$

com $i = 1, \ldots, 21$ e $j = 0, 1, 2$ e suposição de que $\epsilon_i \sim N(0, \sigma^2)$. Como estaremos assumindo parametrização casela de referência teremos a restrição $\gamma_0 = 0$. Ajustando o modelo aos dados, obtém-se os seguintes coeficientes, apresentado na Tabela 4.6.

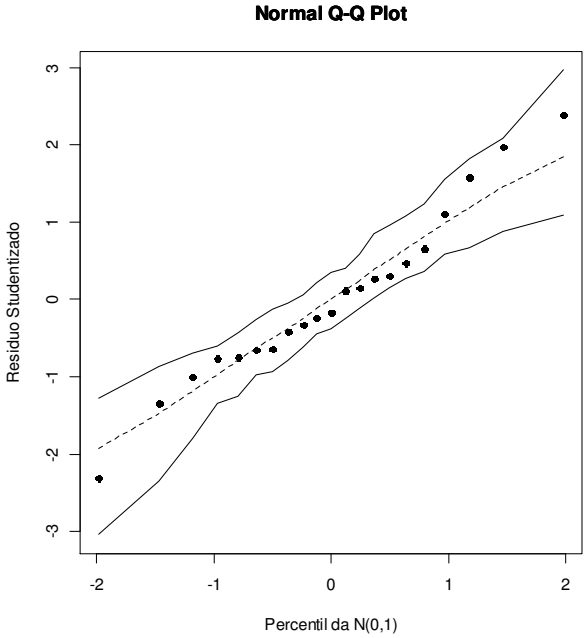
Tabela 4.6 - Estimativas dos coeficientes do modelo Normal Linear

Modelo		Normal	
Efeito	Estimativa	E.Padrão	p-valor
Constante	568,09	52,72	**5,13E-09**
Densidade	-105,92	17,63	**1,41E-05**
Temperatura1	-133,53	35,41	**0,00152**
Temperatura2	-275,07	37,49	**1,16E-06**

Fonte: Dados da análise do autor.

Todos os coeficientes apresentaram bons níveis de significância ($< 0{,}001$) para o modelo.

Gráfico 10 - Gráfico normal de probabilidades para o modelo Normal Linear ajustado.



Fonte: Gráfico criado com o script de (Paula, 2010)

Analisando o Gráfico (10), podemos notar que o modelo ajusta-se muito bem aos dados. O modelo obteve uma medida de ajuste AIC igual a $240{,}61$. Poderíamos aceitar este modelo como sendo um bom modelo para modelar os dados, porém existe um problema com a distribuição adotada. Ela admite valor de $(-\infty, +\infty)$. Nota-se que, no caso da densidade máxima do filme for maior ou igual a 3, e a temperatura for igual a 92ºC, iremos encontrar um valor de tempo negativo. Por causa deste problema, iremos ajustar os modelos log-linear Normal, Gama com todas as ligações que o pacote $stats$ do software livre R oferece e também para o modelo com distribuição Normal Inversa para corrigir este problema, pois estas distribuições estão definidas no intervalo $(0, +\infty)$.

4.2.3 Ajuste pelos modelos log-linear Normal e log-linear Gama

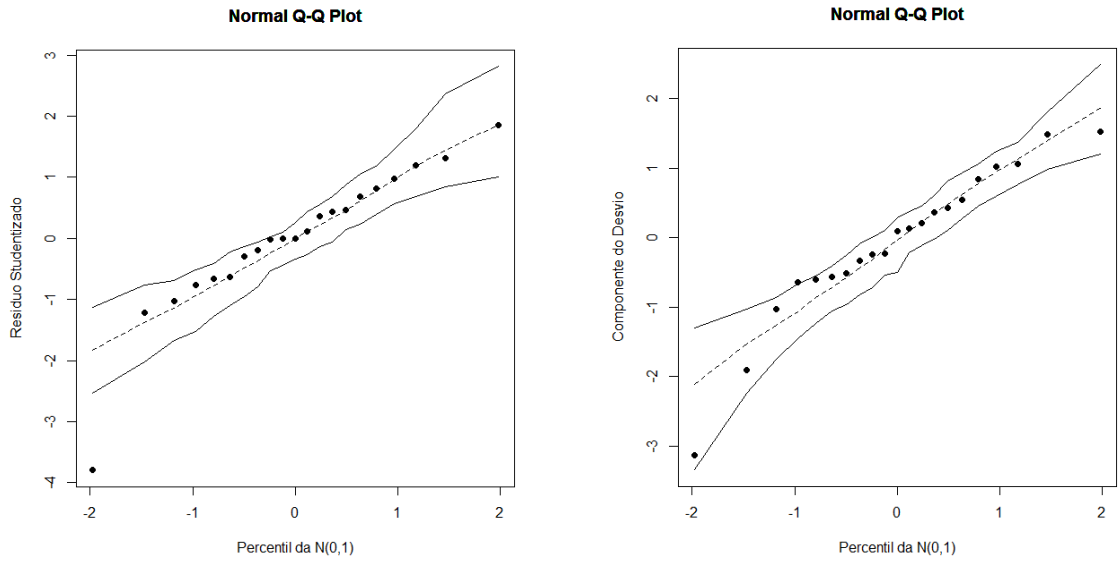
Ajustamos todos os modelos e funções de ligação que o pacote $stats$ pôde oferecer, conforme mencionados na seção anterior, porém apenas dois modelos obtiveram os resultados desejados na procura de um bom ajuste aos dados. Os coeficientes descritos na tabela abaixo são dos modelos log-linear Normal e log-linear Gama. Todos os coeficientes e ambos os modelos apresentaram bons níveis de significância ($< 0{,}001$) para o modelo. Através da medida da qualidade do ajuste AIC, observamos uma melhora em relação ao modelo normal linear, pois obtive para o modelo log-linear Normal AIC igual a $199{,}58$ e para o modelo log-linear Gama AIC igual a $207{,}29$ reduções significativas no valor do AIC, ou seja, dizer melhor qualidade relativa do ajuste.

Tabela 4.7 – Estimativas dos coeficientes dos Modelos Lineares Generalizados.

Modelos	log-linear Normal			log-linear Gama		
Efeito	Estimativa	E.Padrão	p-valor	Estimativa	E.Padrão	p-valor
Constante	7,59749	0,11249	**< 2E-16**	7,46383	0,15207	**< 2E-16**
Densidade	-0,75831	0,04916	**1,99E-11**	-0,73526	0,05084	**5,52E-11**
Temperatura1	-0,8052	0,0604	**1,98E-10**	-0,67822	0,10213	**4,17E-06**
Temperatura2	-1,89874	0,10811	**2,48E-12**	-1,68979	0,10815	**1,62E-11**

Fonte: Dados da análise do autor.

Gráfico 11 - Gráficos normais de probabilidades para os modelos ajustados log-linear normal (a) e log-linear gama (b) aos dados sobre tempo e densidade máxima sob condições experimentais de temperatura.

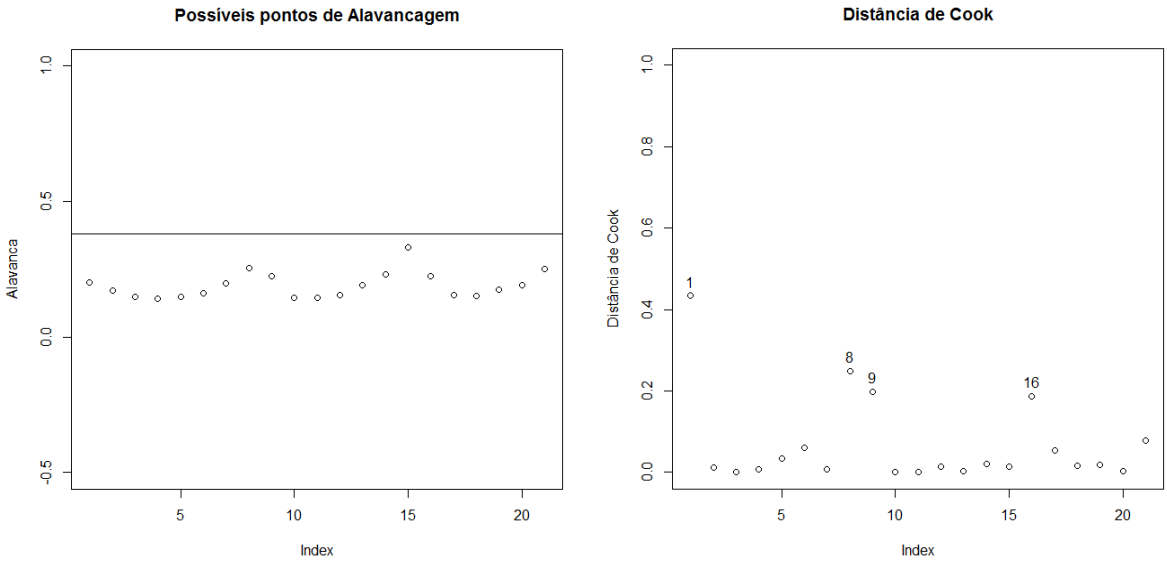


(a) (b)

Fonte: Gráficos (a) e (b) criados utilizando o software R, com o script Paula (2010)

De acordo com os gráficos normais de probabilidades para os modelos com distribuição Normal (Gráfico 11.a) e distribuição Gama (Gráfico 11.b), notou-se uma melhor acomodação dos pontos dentro das bandas do envelope gerado para o segundo modelo. Concluímos que o modelo gama apresentou-se muito bem ao problema de heterocedasticidade dos dados.

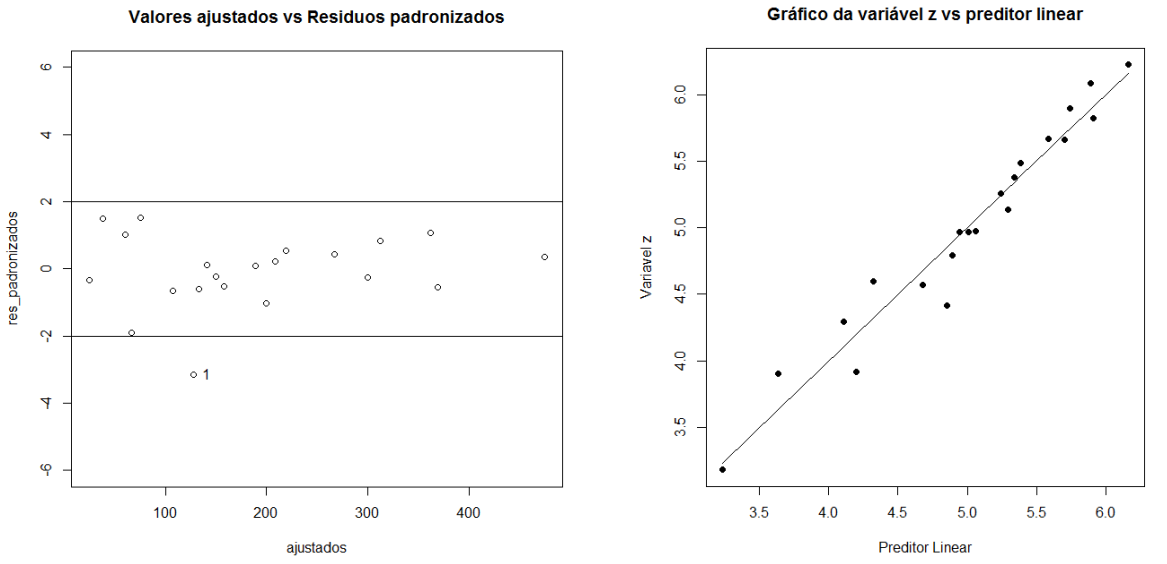
Gráfico 12 – Gráfico para verificação de pontos de alavanca (a) e Gráfico para verificação de pontos de influência (b).



(a) (b)

Fonte: Criado pelo autor através de um script, criado por (Ferreira, 2010) notas de aula de MLG.

Gráficos 13 – Gráfico para verificação de pontos aberrantes (a) e Gráfico para verificação de adequação da função de ligação (b).



(a) (b)

Fonte: Criado pelo autor através de um script, criado por (Ferreira, 2010) notas de aula de MLG.

4.2.4 Diagnóstico do Modelo Selecionado.

São apresentados alguns gráficos de diagnóstico (Gráficos 12 e 13). No Gráfico (12.a) são apresentados os valores de h_{ii}, observando que nenhum dos pontos são destacados como possivelmente de alavanca. Já pelo Gráfico (12.b) notamos pelo menos um ponto com mais destaque como influente em $\widehat{\boldsymbol{\beta}}$, os filmes #1, #8, #9 e #16. Os 4 filmes apresentam densidade máxima maior ou igual a 2,91 e com tempo de duração menor ou igual a 96 horas. Os filmes #8 e #9 foram experimentados sob a mesma condição de temperatura igual a 82ºC e os filmes #1 e #16 foram experimentados respectivamente sob as condições de temperatura de 72ºC e 92ºC. Pelo Gráfico (13.a) notamos apenas um ponto (#1) com mais destaque como aberrante. Esses recrutas tiveram um número alto de ocorrências de infecções, um é mulher (#47) e o outro (#249) é homem. Em geral os pontos de influência e o ponto aberrante desse exemplo referem-se a filmes com um valor elevado de densidade máxima e tempo de duração menor em relação aos demais filmes experimentados. Finalmente, o Gráfico (13.b) indica que a escolha da ligação logarítmica não parece ser inadequada.

4.1.5 Interpretação do modelo final

Portanto, o modelo final ajustado log-linear Gama fica

$$\widehat{\boldsymbol{y}} = e^{7{,}46383 - 0{,}73526 * \boldsymbol{Densidade} - 0{,}67822 * \boldsymbol{Temperatura1} - 1{,}68979 * \boldsymbol{Temperatura2}}$$

Desse modelo, podemos extrair as seguintes interpretações: $e^{\widehat{\beta}_1} = e^{-0{,}73526} = 0{,}4793$ isso significa que o aumento de uma unidade da densidade máxima corresponde a $(1 - 0{,}4793) * 100 = 42{,}07\%$ de redução em média no tempo de duração do filme. Observando agora a variável temperatura podemos concluir para a temperatura de 82ºC que $e^{\widehat{\gamma}_1} = e^{-0{,}67822} = 0,5075$ corresponde uma redução em média no tempo de duração do filme de $(1 - 0,5075) * 100 = 49{,}25\%$ em relação a 72ºC e o filme sob a condição experimental de 92ºC $e^{\widehat{\gamma}_2} = e^{-1.68979} = 0{,}1845$ corresponde a uma redução em média no tempo de duração do filme de $(1 - 0,1845) * 100 = 81{,}55\%$ em relação a temperatura de 72ºC.

Capítulo 5

CONCLUSÃO

Embora seja bastante comum encontrarmos conjuntos de dados que apresentem distribuições diferentes da normal, e que uma quantidade razoável de material já tenha sido publicada nessa área, a análise de diagnóstico é pouca explorada. O enfoque computacional nesse caso é de uma enorme importância, dada a complexidade dos algoritmos a serem utilizados. Um segundo objetivo foi o de ajustar estes modelos com um algoritmo implementado no software livre R. A linguagem R é bastante intuitiva e versátil, mostrou-se muito eficaz com as funcionalidades já desenvolvidas para os modelos aqui estudados (Normal, Poisson, Binomial Negativa, Normal Inversa e Gama) facilitando de certa maneira o trabalho desenvolvido.

No exemplo abordado no Capítulo 4 observamos através da Tabela 4.7, que as estimativas dos parâmetros correspondentes ao modelo log-linear Normal e log-linear Gama são próximas, no entanto, quando verificamos a análise dos gráficos de envelope simulados do modelo log-linear Normal mostra um ajuste pobre dos dados em estudo, evidenciando a falta de robustez do modelo para observações mais afastadas dos dados. Já o modelo log-linear gama não observa o ponto discrepante que mereça atenção como no outro modelo, indicando um ajuste mais adequado onde contempla de forma satisfatória a heterocedasticidade presente nos dados.

Sugestões de novas linhas de estudo são os Modelos Aditivos Generalizados para locação, escala e forma (GAMLSS), uma vasta classe de modelos apresentados por (Rigby e Stasinopoulos (2005). Diversas outras famílias de modelos de regressão são casos particulares de um modelo GAMLSS como, por exemplo, os Modelos Lineares Generalizados. Dessa forma, o GAMLSS, segundo Rigby e Stasinopoulos (2005), permite a inclusão em um mesmo modelo de termos fixos paramétricos e não paramétricos e a inclusão de fatores aleatórios. E a principal diferença entre os MLGs e estes novos modelos é que podem ser utilizados para modelar distribuições que não pertencem à família exponencial e ainda possibilitam o ajuste de todos os parâmetros da distribuição da variável resposta em função das variáveis preditoras. Da mesma forma quando foram apresentados os MLGs, os modelos GAMLSS abrem assim um leque de opções ainda maior para a distribuição da variável resposta, bem como dar maior flexibilidade para ligação entre a média da variável resposta e a parte sistemática do modelo, o preditor linear.

REFERÊNCIAS

Atkinson, A.C. (1981). Robustness, transformations and two graphical displays for outlying and influential observations in regression. Biometrika, 68, 13-20.

Box, G.E.P. e Cox, D.R. (1964). An analysis of transformations. Journal of the Royal Statistical Society(B), 26 (2):211–252.

Cordeiro, G.M. (1986). Modelos lineares generalizados. VII SINAPE, UNICAMP.

Cordeiro, G.M. e Demétrio, C.G.B. (2008). Modelos Lineares Generalizados e Extensões, Piracicaba: ESALQ, Departamento de Ciências Exatas.

Cordeiro, G.M. e Lima Neto, E.A. (2006). Modelos Paramétricos. Recife: Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática.

Demétrio, C.G.B. e Zocchi, S.S. (2008). Modelos de Regressão, Piracicaba: ESALQ, Departamento de Ciências Exatas.

Ferreira, C.S. (2010). Diagnostico em Modelos Lineares Generalizados, Juiz de Fora: Universidade Federal de Juiz de Fora, Notas de Aula.txt.

Fisher, R.A. (1925). Statistical methods for research workres. Oliver and Boyd, Edinburgh.

Hinkley, D.V. (1985). Transformation diagnostic for linear models. Biometrika, 72, 487-496.

Hoaglin, D.C. e Welsch, R. (1978). The hat matrix in regression and ANOVA. The American Statistician, 32, 17-22.

Lauretto, M.S. (2011). Análise Exploratória de Dados, São Paulo: Universidade de São Paulo, Disciplina de Estatística Computacional. Notas de Aula. PDF.

Lee, A.H. (1987). Diagnostic displays for assessing leverage and influence in generalized linear models. Austral. J. Statist., 29, 233-243.

McCullagh, P. e Nelder, J.A. (1989). Generalized linear models. Chapman and Hall, London.

Nelder, J.A. e Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A*, 135, 370–384.

Paula, G.A. (2010). Modelos de Regressão com Apoio Computacional, São Paulo: IME – Universidade de São Paulo.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape, (with discussion).Appl. Statist., 54: 507–554.

Williams, D. A. (1984). Residuals in generalized linear models. In: Proceedings of the 12th. International Biometrics Conference, Tokyo, pp. 59-68.

APÊNDICE

Distribuição de Y	***Funções de ligação*** (disponíveis no *software* R)		
Gaussiana	Identidade $\mu = \beta_0 + \beta_1 X_1 + \ldots$		
Gama	Identidade $\mu = \beta_0 + \beta_1 X_1 + \ldots$	Inversa $\frac{1}{\mu} = \beta_0 + \beta_1 X_1 + \ldots$	Logarítmica $\log(\mu) = \beta_0 + \beta_1 X_1 + \ldots$
Poisson	Identidade $\mu = \beta_0 + \beta_1 X_1 + \ldots$	Raiz quadrada $\sqrt{\mu} = \beta_0 + \beta_1 X_1 + \ldots$	Logarítmica $\log(\mu) = \beta_0 + \beta_1 X_1 + \ldots$
Binomial	Logit $\log\left(\frac{\mu}{1+\mu}\right) = \beta_0 + \beta_1 X_1 + \ldots$	Probit $\Phi^{-1}(\mu) = \beta_0 + \beta_1 X_1 + \ldots$	Complementar log-log $\log\left[-\log(1-\mu)\right] = \beta_0 + \beta_1 X_1 + \ldots$
Binomial Negativa*	Identidade $\mu = \beta_0 + \beta_1 X_1 + \ldots$	Raiz quadrada $\sqrt{\mu} = \beta_0 + \beta_1 X_1 + \ldots$	Logarítmica $\log(\mu) = \beta_0 + \beta_1 X_1 + \ldots$
Gaussiana Inversa	Quadrática Inversa $\frac{1}{\mu^2} = \beta_0 + \beta_1 X_1 + \ldots$		

Para a realização desta monografia foi utilizado o software livre R, os comandos usados na criação dos gráficos de envelope e outros podem ser encontrados na página virtual do Professor Gilberto Alvarenga Paula: http://www.ime.usp.br/~giapaula/ .

1-Comandos utilizados para a análise do banco de dados **recrutas.txt.**

```
##############################################
### Monografia Samuel de Oliveira 200755020##
##############################################

setwd("C:/Users/Samuel/Desktop/Segundo conjunto de dados - Monografia") ###
Mudar Diretório ###

#####################################
##Carregar biblioteca e ler os dados##
#####################################

library(MASS)

dados= read.table(file="recrutas.txt", header=T)
attach(dados)

####################
#Analise descritiva#
####################

summary(dados)

faixa.tb=table(fetaria)
barplot(faixa.tb,main="Gráfico de barras para faixa
etária",ylab="Frequências",xlab="Faixa Etária")

casos.tb = table(casos)
barplot(casos.tb,,main="Gráfico de barras para o número de
casos",ylab="Frequências",xlab="Número de ocorrência de infecções no
ouvido")

par(mfrow=c(2,2))
boxplot(casos~habito, xlab="Habito",ylab="Nº de casos de
infecções",main="Boxplot do nº de casos segundo o habito")
title(sub="(a)")
boxplot(casos~local, xlab="Local",ylab="Nº de casos de infecções",
main="Boxplot do nº de casos segundo o local")
title(sub="(b)")
boxplot(casos~sexo, xlab="Sexo",ylab="Nº de casos de infecções",
main="Boxplot do nº de casos segundo o sexo")
title(sub="(c)")
boxplot(casos~fetaria, xlab="Faixa Etária",ylab="Nº de casos de
infecções",main="Boxplot do nº de casos segundo a f.etária")
title(sub="(d)")

#################################
#Ajuste por Modelo Linear Clássico#
#################################

rcasos=log(casos+1)

fit.model=glm(rcasos ~ habito + local + fetaria + sexo + habito*local +
habito*fetaria + habito*sexo + local*fetaria + local*sexo + fetaria*sexo,
family= gaussian(link="identity"))
ajuste=stepAIC(fit.model)

fit.model=glm(rcasos ~ habito + local + sexo + local:sexo, family=
gaussian(link="identity"))
summary(fit.model)
source("envel_norm.txt")
```

```
#####################################
#Modelo Linear Generalizado - Poisson#
#####################################

fit.model=glm(casos~ habito + local + fetaria + sexo + habito*local +
habito*fetaria + habito*sexo + local*fetaria + local*sexo + fetaria*sexo,
family= poisson(link="identity"))
ajuste=stepAIC(fit.model)
summary(fit.model)
source("envel_pois.txt")

fit.model=glm(casos~ habito + local + fetaria + sexo + habito*local +
habito*fetaria + habito*sexo + local*fetaria + local*sexo + fetaria*sexo,
family= poisson(link="log"))
ajuste=stepAIC(fit.model)

fit.model=glm(casos ~ habito + local + fetaria + sexo + habito:local +
habito:fetaria + local:fetaria + local:sexo, family= poisson(link="log"))
summary(fit.model)
source("envel_pois.txt")

fit.model=glm(casos~ habito + local + fetaria + sexo + habito*local +
habito*fetaria + habito*sexo + local*fetaria + local*sexo + fetaria*sexo,
family= poisson(link="sqrt"))
ajuste=stepAIC(fit.model)

fit.model=glm(casos ~ habito + local + fetaria + sexo + habito:local +
habito:fetaria + local:fetaria + local:sexo, family= poisson(link="sqrt"))
summary(fit.model)
source("envel_pois.txt")


####################
##Binomial Negativa##
####################

fit.model=glm.nb(casos~ habito + local + fetaria + sexo + habito*local +
habito*fetaria + habito*sexo + local*fetaria + local*sexo +
fetaria*sexo,link=log)
ajuste=stepAIC(fit.model)

fit.model=glm.nb(casos ~ habito + local + sexo + local:sexo,link=log)
summary(fit.model)
source("envel_nbin.txt")


fit.model=glm.nb(casos~ habito + local + fetaria + sexo + habito*local +
habito*fetaria + habito*sexo + local*fetaria + local*sexo +
fetaria*sexo,link=sqrt)
ajuste=stepAIC(fit.model,direction = both)

fit.model=glm.nb(casos ~ habito + local + sexo + local:sexo,link=sqrt)
summary(fit.model)
source("envel_nbin.txt")

#Não podemos utilizar essa função de ligação, pois...da erro#
fit.model=glm.nb(casos~ habito + local + fetaria + sexo + habito*local +
habito*fetaria + habito*sexo + local*fetaria + local*sexo +
fetaria*sexo,link=identity)
```

```
#############################################################
##Pressupostos de normalidade e homocedasticidade dos erros##
#############################################################

# Pressupostos de normalidade e homocedasticidade dos erros #

res_padronizados=rstandard(fit.model)
ajustados=fit.model$fitted.values
plot(ajustados,res_padronizados,main="Valores ajustados vs Residuos
padronizados")
abline(h=0)

# normalidade

plot(fit.model) # os 4 gráficos de diagnostico#


# Ponto aberrante #

plot(ajustados,res_padronizados,ylim=c(-6,6),main="Valores ajustados vs
Residuos padronizados")
abline(h=c(-2,2))
identify(ajustados,res_padronizados,pos=T)

# Alavanca #

fit=influence.measures(fit.model)


hii=fit$infmat[,9]
plot(hii,ylab="Alavanca Medida de hii",ylim=c(-.5,1),main="Possiveis pontos
de Alavancagem")
abline(h=2*length(fit.model$coefficients)/length(casos))
identify(hii,pos=T)

# ponto de influencia #

cook=fit$infmat[,8]
plot(cook,ylab="Distância de Cook",ylim=c(0,1),main="Distância de Cook")
identify(cook,pos=T)


# Variável zi # Verificar a função de ligação#

fi <- fit.model$theta
w <- fi*fitted(fit.model)/(fi + fitted(fit.model))

eta = predict(fit.model)
z = eta + resid(fit.model, type="pearson")/sqrt(w)
plot(predict(fit.model),z,xlab="Preditor Linear",
ylab="Variavel z", pch=16 ,main= "Gráfico da variável adicionada" )
lines(smooth.spline(predict(fit.model), z, df=2))
```

2-Comandos utilizados para a análise do banco de dados **dfilme.txt.**

```
################################################
### Monografia Samuel de Oliveira 200755020 ###
################################################

setwd("C:/Users/Samuel/Desktop/Segundo conjunto de dados - Monografia") ###
Mudar Diretório ###

#######################################
##Carregar biblioteca e ler os dados##
#######################################

library(MASS)

dados=read.table("dfilme.txt",header=T)
attach(dados)

temperatura <- factor(temperatura)
temperatura <- C(temperatura,treatment)


###################
#Analise descritiva#
###################

summary(dados)

par(mfrow=c(2,2))
plot(densidade,tempo, pch=19,
col=c("lightblue3","orange","red3")[unclass(temperatura)], main="Gráfico de
Dispersão")
legend("topright",
legend=c("Temp1=72ºC","Temp2=82ºC","Temp3=92ºC"),pch=c(19,19,19),col=c("lig
htblue3","orange","red3"), bty="n")

hist(tempo,main="Histograma da Variável Tempo")
hist(densidade, main="Histograma da Variável Densidade")
plot(tempo~temperatura, main="Boxplot da Variável Tempo vs Temperatura")

#################################
#Ajuste por Modelo Linear Clássico#
#################################

fit.model=glm(tempo~densidade+temperatura,
family=gaussian(link="identity"))

summary(fit.model)
Source("envel_norm.txt")

#######################################
#Ajuste por Modelo Linear Generalizado#
#######################################

fit.model=glm(tempo~densidade+temperatura, family=gaussian(link="inverse"))
summary(fit.model)
Source("envel_norm.txt")

fit.model=glm(tempo~densidade+temperatura, family=gaussian(link="log"))
summary(fit.model)
Source("envel_norm.txt")

fit.model=glm(tempo~densidade+temperatura,
family=inverse.gaussian(link="identity"))
summary(fit.model)
source("envel_ninv.txt")

fit.model=glm(tempo~densidade+temperatura,
family=inverse.gaussian(link="log"))
summary(fit.model)
```

```
source("envel_ninv.txt")

fit.model=glm(tempo~densidade+temperatura,
family=inverse.gaussian(link="inverse"))
summary(fit.model)
source("envel_ninv.txt")

fit.model=glm(tempo~densidade+temperatura, family=Gamma(link="identity"))
summary(fit.model)
source("envel_gama.txt")

fit.model=glm(tempo~densidade+temperatura, family=Gamma(link="inverse"))
summary(fit.model)
source("envel_gama.txt")

fit.model=glm(tempo~densidade+temperatura, family=Gamma(link="log"))
summary(fit.model)
source("envel_gama.txt")

############################################################
#Pressupostos de normalidade e homocedasticidade dos erros#
############################################################

res_padronizados=rstandard(fit.model)
ajustados=fit.model$fitted.values
plot(ajustados,res_padronizados,main="Valores ajustados vs Residuos
padronizados")
abline(h=0)
identify(ajustados,res_padronizados,pos=T)

#normalidade#

plot(fit.model) # os 4 gráficos de diagnostico#

#Ponto aberrante #

plot(ajustados,res_padronizados,ylim=c(-6,6),main="Valores ajustados vs
Residuos padronizados")
abline(h=c(-2,2))
identify(ajustados,res_padronizados,pos=T)

#Alavanca#

fit=influence.measures(fit.model)
hii=fit$infmat[,8]
plot(hii,ylab="Alavanca",ylim=c(-.5,1),main="Possíveis pontos de
Alavancagem")
abline(h=2*length(fit.model$coefficients)/length(tempo))
identify(hii,pos=T)

#ponto de influencia#

cook=fit$infmat[,7]
plot(cook,ylab="Distância de Cook",ylim=c(0,1),main="Distância de Cook")
identify(cook,pos=T)

# Variável zi # Verificar a função de ligação#

fi <- fit.model$theta
w <- fi*fitted(fit.model)/(fi + fitted(fit.model))

eta = predict(fit.model)
z = eta + resid(fit.model, type="pearson")/sqrt(w)
plot(predict(fit.model),z,xlab="Preditor Linear",
ylab="Variavel z", pch=16 ,main= "Gráfico da variável adicionada" )
lines(smooth.spline(predict(fit.model), z, df=2))
```