Universidade Federal de Pernambuco
Centro de Ciências Sociais Aplicadas
Departamento de Ciências Contábeis e Atuariais

Karina Körössy Leite

UMA INTRODUÇÃO À MODELAGEM LINEAR GENERALIZADA APLICADA A DADOS DE PLANO DE SAÚDE

Recife, 2013

Karina Körössy Leite

UMA INTRODUÇÃO À MODELAGEM LINEAR GENERALIZADA APLICADA A DADOS DE PLANO DE SAÚDE

Monografia apresentada ao Curso de ciências atuariais, para obtenção do bacharelado em ciências atuariais sob a orientação do Prof. Cícero Rafael Barros Dias.

Recife, 2013

KARINA KOROSSY LEITE

TITULO

Trabalho de Conclusão de Curso, da Universidade Federal de Pernambuco – UFPE, apresentado como pré-requisito para obtenção do título de Bacharel em Ciências Atuariais.

DATA DA APROVAÇÃO: Recife, ____ de _____ de 2013.

COMISSÃO EXAMINADORA

_____

Prof. Mestre Cícero Rafael Barros Dias

_____

Prof. Doutor Josenildo dos Santos

_____

Prof. Doutor Alessandro

AGRADECIMENTOS

Em primeiro lugar, agradeço à Deus e à minha família (primordialmente meus pais), pois eles são a base de tudo. Agradeço à minha irmã, Nathália Körössy, pelas dicas da ABNT.

Agradeço também ao meu namorado, Jahyr César, pela ajuda com a organização dos dados e pelo apoio dado nesses últimos meses.

Agradeço ao meu orientador Cícero Dias pela paciência com que me auxiliou neste trabalho e, juntamente com os professores Maurício Assuero e Josenildo dos Santos, pelo apoio valioso dado desde o início do curso.

Agradeço também aos meus professores, que desde a época do curso de estatística, tanto me acrescentaram, academicamente e pessoalmente: Cláudia Regina, Manoel Raimundo, Audrey Helen, Sylvio Santos, Isaac Xavier, Raydonal Ospina, Renato Cintra, Leandro Rêgo.

Agradeço aos meus amigos estatísticos Adriano, Bruna Eloize, Daniel Cassimiro e Edijan Cavalcanti, cujos conhecimentos me auxiliaram imensamente na resolução deste trabalho.

Agradeço aos meus companheiros de sala de aula: Jorge Tiago, Olívia Maria e Helena Lourenço, que muito contribuíram para o êxito obtido durante o curso.

RESUMO

Modelos lineares generalizados é uma metodologia usada para se modelar relações entre variáveis. Generaliza o clássico modelo linear normal, por relaxar algumas de suas restrições e fornece métodos para a análise de dados não-normais. A importância de tais modelos não é apenas de índole prática. Do ponto de vista teórico, a sua importância ocorre essencialmente do fato de possuir uma metodologia constituída por uma abordagem unificada de muitos procedimentos estatísticos correntemente usados nas aplicações e promover o papel central da verossimilhança na teoria da inferência. Este texto apresenta a metodologia do modelo linear generalizado (MLG). Ao final, é feita uma aplicação a dados de plano de saúde. No entanto, os resultados não foram tão satisfatórios, talvez por conta da limitação dos dados com a ausência de algumas variáveis importantes, como por exemplo: valor das consultas e mensalidade paga pelo usuário do plano.

ABSTRACT

Generalized linear models is a methodology used to model relations between variables. Generalize the classic normal linear model, by relaxing some of your restrictions and provides methods to analyze non-normal datas. The importance of such facts is not only a practical matter. In a theoretical way of view, your importance occurs essentially by the fact of posses a methodology constituted by a unified approach of many statistics procedures commonly used in the applications and promote the central role of likelihood in the theory of interference. This Text shows a methodology of the generalized linear model (GLM). Finally, an application in the data of the health insurance is made. however, the results is not satisfactory, maybe because of the data limitation with absence of some important variables, for example: the price of the consults and the values paid to the health insurance.

LISTA DE FIGURAS

LISTA DE TABELAS

Sumário

1 INTRODUÇÃO

Na modelagem estatística, por diversas vezes, somos confrontados por problemas em que o objetivo principal é estudar a relação entre variáveis, ou até mesmo analisar a influência que uma ou mais variáveis (chamadas explicativas) têm sobre uma determinada variável de interesse (chamada variável resposta). O modelo estatístico explica a conexão entre a variável resposta e as explicativas.

Ao longo de muitos anos os modelos normais lineares foram utilizados, mesmo quando o fenômeno estudado não apresentava uma resposta seguindo a suposição de normalidade (para isso, usava-se algum tipo de transformação a fim de alcançar a normalidade). Com o desenvolvimento computacional ocorrido na década de 1970, alguns modelos começaram a ser mais aplicados, como por exemplo o modelo normal não linear. Todavia, a proposta mais interessante foi apresentada por Nelder e Wedderburn (1972), que propuseram os modelos lineares generalizados (MLGs), cuja ideia básica é a de que a variável resposta possui uma distribuição pertencente à família exponencial e uma maior flexibilidade para a relação funcional entre a média da variável resposta e as variáveis explicativas do modelo (PAULA, 2010, p.2).

Turkman e Silva (2000, p.2) enfatizam que: "os Modelos Lineares Generalizados introduzidos por Nelder e Wedderburn (1972) vieram a unificar, tanto do ponto de vista teórico como conceitual a teoria da modelagem estatística até então desenvolvida". Acrescentam também que:

> "Devido ao grande número de modelos e à facilidade associada ao desenvolvimento computacional ocorrido nas últimas décadas, os MLGs desempenham um papel cada vez mais importante na análise estatística, apesar das limitações ainda impostas, nomeadamente por manterem a estrutura de linearidade, pelo fato das distribuições se restringirem à família exponencial e por exigirem a independência das respostas" (TURKMAN e SILVA, 2000, p.3).

No capítulo 1 é feita uma contextualização do problema; no capítulo 2 é discutido sobre os procedimentos metodológicos e os dados escolhidos para a aplicação; no capítulo 3 é feita uma abordagem da literatura de MLG, com seus conceitos e notações; no capítulo 4 são apresentados os resultados da modelagem de um banco de dados de plano de saúde e no capítulo 5 é realizada a conclusão.

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA

Modelagem estatística, em geral, nos ajuda a responder determinadas questões através de modelos, tais como: quais variáveis explicativas são influentes na resposta e qual a escala apropriada para incluí-las no modelo? A variabilidade na resposta pode ser bem explicada pela variabilidade das variáveis explicativas? Qual é a predição da resposta para determinados valores das variáveis explicativas e qual a precisão associada a esta predição? (JONG e HELLER, 2008, p.1).

Nas ciências atuariais, podemos citar questionamentos da seguinte natureza: Qual é a relação entre o prazo de liquidação e o número de reclamações de sinistros resolvidos? Qual o impacto do valor de sinistros no nível de dano de uma seguradora? (JONG e HELLER, 2008, p.2). Qual o impacto do número de consultas e internações na despesa de um plano de saúde?

1.2 OBJETIVOS

1.2.1 Objetivo geral

O presente trabalho tem por objetivo utilizar a metodologia dos modelos lineares generalizados aplicada a dados de seguro, mais precisamente dados de plano de saúde, e apresentar uma visão atuarial dos resultados.

1.2.2 Objetivos específicos

a) Descrever a metodologia da modelagem linear generalizada;
b) Fazer uma aplicação da metodologia a dados de plano de saúde.

1.3 JUSTIFICATIVA

Justifica-se esta pesquisa com base na importância da análise de dados para uma operadora de planos de saúde. Visto que o cálculo de reservas e prêmios de seguro é feito pelo atuário com base na despesa estimada, fica então exposta a ponderação de tal trabalho.

2 PROCEDIMENTOS METODOLÓGICOS

De acordo com Bruyne (1991 p. 29), “a metodologia é a lógica dos procedimentos científicos em sua gênese e em seu desenvolvimento, não se reduz, portanto, a uma “metrologia” ou tecnologia da medida dos fatos científicos”. Ou seja, a metodologia representa toda a coerência dos artifícios científicos desde a sua formação até o seu desenvolvimento prático.

2.1 TIPO E MÉTODO DE PESQUISA

Como afirmou Minayo (1993, p.23), “a pesquisa é considerada como atividade básica das ciências na sua indagação e descoberta da realidade. É uma atividade de aproximação sucessiva da realidade que nunca se esgota, fazendo uma combinação particular entre teoria e dados”.

Trata-se de uma pesquisa aplicada, quanto à natureza; quantitativa, quanto à forma de abordagem; descritiva, quanto aos objetivos, pois visa o estabelecimento de relações entre variáveis; e bibliográfica, quanto aos procedimentos técnicos.

Segundo Moresi (2003, p.8), pesquisa aplicada "objetiva gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos". Afirma também na mesma obra que pesquisa quantitativa significa "traduzir em números opiniões e informações para classificá-las e analisá-las. Requer o uso de recurso e de técnicas estatísticas (percentagem, média, moda, mediana, desvio-padrão, coeficiente de correlação, análise de regressão, etc)".

Macedo (1994, p.13) argumenta que “pesquisa bibliográfica é a busca de informações bibliográficas, seleção de documentos que se relacionam com o problema e o respectivo fichamento das referências para que sejam posteriormente utilizadas”. Os pesquisadores Silva & Menezes (2000, p.21) definem pesquisa descritiva como sendo “a pesquisa que visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis”.

O método fornece os meios para se alcançar o objetivo proposto, ou seja, são as ferramentas das quais fazemos uso na pesquisa, a fim de responder nossa questão.

2.2 DELIMITAÇÃO DA PESQUISA E DADOS UTILIZADOS

A presente pesquisa delimita-se pelo estudo de publicações relacionadas aos modelos lineares generalizados, seus conceitos e notações. Além disso, faz uso de 17.006 dados de plano de saúde, a fim de exemplificar o modelo adotado (MLG).

Tais dados são pertencentes ao ano de 2001. As variáveis presentes são: sexo (M ou F), idade, quantidade de consultas, tempo no plano, quantidade de internações e tipo de acomodação contratado (apartamento ou enfermaria) .

3 REVISÃO DA LITERATURA

3.1 CONCEITO DE MODELAGEM

Como enfatizam Jong e Heller (2008, p.3):

> "Modelagem não é um fim em si, tem o objetivo de proporcionar resoluções de questões de interesse. Modelos diferentes muitas vezes são aplicados aos mesmos dados, dependendo das questões. Isto enfatiza que a modelagem é uma atividade pragmática e não há um modelo verdadeiro".

Para determinada situação não há um único modelo adequado, mas um leque de modelos que podem ser utilizados convenientemente.

Modelos conectam variáveis, e isto requer um entendimento da natureza de tais variáveis. Um modelo estatístico é tão bom quanto os dados utilizados nele. Consequentemente, um bom entendimento dos dados é um ponto inicial essencial para a modelagem. Uma quantidade significante de tempo é gasta na limpeza e exploração dos dados. De acordo com Cordeiro e Paula (1989, p.227), "algumas características nos dados podem não ser descobertas, mesmo por um modelo muito bom e, portanto, um conjunto razoável de modelos adequados aumenta a possibilidade de se detectar essas características".

Exploração de dados utilizando representações gráficas e tabulações apropriadas é o primeiro passo na construção de modelos. É feita para uma compreensão global das relações entre as variáveis, permitindo-nos a visualização de relações entre a resposta e as potenciais variáveis explicativas e relações entre as potenciais variáveis explicativas entre si.

As variáveis podem assumir diferentes formas: discretas ou contínuas, nominais, ordinais, categóricas e assim por diante. É importante distinguir entre os diferentes tipos, pois a maneira como elas podem razoavelmente introduzir um modelo depende disto. Exibições de dados diferem fundamentalmente se as variáveis são contínuas ou categóricas. Por exemplo, se queremos saber a relação entre duas variáveis contínuas, a melhor maneira de visualizar é através de um gráfico de dispersão. Se ambas forem categóricas, a forma mais usual é a tabela de frequências. Se for uma contínua e a outra categórica, boxplots são mais apropriados (JONG e HELLER, 2008, p.6-9).

Outro ponto a ser discutido na modelagem são os problemas que podem ocorrer com principalmente em dados com grandes números de observações. Na prática, dados de plano de

saúde são tipicamente extensos, pois englobam todos os usuários existentes nele. Problemas como valores faltantes e inconsistência ou registros inválidos devem ser resolvidos antes da modelagem estatística. Gráficos exploratórios podem revelar os erros de registro mais grosseiros.

Outro problema que pode ocorrer é o aparecimento de tendências. Uma análise estatística é idealmente imparcial. Ou seja, resulta em meios que não favorecem a outras conclusões. Tendências podem surgir de diversas formas, entre outras: através de resultados censurados (como exemplo, ao se estudar o tempo de vida médio de nascidos em 1950, é provável que nenhuma observação possa ser feita para o tempo de vida dos sobreviventes, ainda); através de casos dependentes, tais como, vários acidentes envolvendo dois carros, mas cada carro deve constituir um sinistro distinto (JONG e HELLER, 2008, p.13-14).

O próximo passo a ser feito na modelagem é a escolha da distribuição resposta, que é discutido no próximo tópico.

3.2 DISTRIBUIÇÕES RESPOSTA

Este tópico introduz as distribuições estatísticas utilizadas em análise de dados de seguro e modelos lineares generalizados (JONG e HELLER, 2008, p.20).

Demétrio e Cordeiro (2010, p.23) afirmam que as possíveis distribuições de variável resposta englobam as distribuições: “normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens”.

Em análise estatística, um resultado tal como o número de reclamações de um sinistro, ou número de vezes que um usuário de plano de saúde precisa de consulta, ou até mesmo quando tal usuário precisa de cirurgia ou internação, é considerado como pelo menos parcialmente determinado por probabilidade. Esta configuração é formalizada pela introdução das variáveis aleatórias.

3.2.1 Variáveis aleatórias

Uma variável aleatória Y é um número determinado por probabilidade. O conjunto de valores que Y pode assumir é chamado espaço amostral, denotado por Ω.

O espaço amostral é um conjunto finito de números reais. Possui uma função de probabilidade $f(y)$ que indica para cada $y \in \Omega$, a probabilidade que a variável aleatória assume a cada valor y. A função $f(y)$ é não-negativa em Ω, ou seja $f(y) > 0$ se $y \in \Omega$, e 0, caso contrário. Além disso,

$$\sum_{y \in \Omega} f(y) = 1,$$

se Y assume valores discretos, ou

$$\int_{-\infty}^{+\infty} f(y)dy = 1,$$

se Y assume valores contínuos. O valor esperado e a variância de uma variável aleatória discreta são definidos como:

$$\mu = E(Y) \equiv \sum_{y \in \Omega} yf(y), \qquad Var(Y) \equiv E\{(Y-\mu)^2\} = \sum_{y \in \Omega} (y-\mu)^2 f(y)$$

e

$$\mu = E(Y) \equiv \int_{-\infty}^{\infty} yf(y)dy, \quad Var(Y) \equiv E\{(Y-\mu)^2\} = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy,$$

se for para uma variável aleatória contínua.

(i) Bernoulli

A distribuição de Bernoulli admite apenas dois possíveis resultados, 0 ou 1, portanto $\Omega = \{0,1\}$. O evento $y = 1$ é chamado "sucesso", enquanto que $y = 0$ quer dizer "fracasso". Além do mais, $f(1) = p$ e $f(0) = 1 - p$, onde $0 < p < 1$.

Como exemplo aplicado a seguro, 1 e 0 podem corresponder a ocorrência ou não ocorrência de sinistro em uma apólice, ou até mesmo morte e vida, respectivamente. Tais eventos (sinistro ou morte) ocorrem com probabilidade p.

$$E(Y) = p; \quad Var(Y) = p(1-p).$$

A função de probabilidade é

$$f(y) = p^y(1-p)^{1-y}, y = 0, 1.$$

(ii) Binomial

A distribuição binomial é uma generalização da distribuição de Bernoulli e é usada para modelar situações tais como o número de apólices que sofrem sinistros. Considere n variáveis aleatórias independentes de Bernoulli com probabilidade de sucesso p. Então o número total de sucessos segue a distribuição binomial denotada por $Y \sim B(n, p)$. A função de probabilidade é dada por

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0, 1, \dots, n,$$

onde p é desconhecido e n é conhecido. Apresenta por média e variância, respectivamente,

$$E(Y) = np, \;\; Var(Y) = np(1-p).$$

(iii) Poisson

Considere $y \sim P(\mu)$. Então, a função de probabilidade de y é

$$f(y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0, 1, 2, \dots$$

e $E(Y) = \mu = Var(Y)$.

(iv) Binomial Negativa

A derivação clássica da binomial negativa é o número de falhas em ensaios de Bernoulli até que haja r sucessos. Havendo y falhas, implica que o ensaio $r + y$ é um sucesso e em $r + y - 1$ ensaios houve $r - 1$ sucessos e y falhas. Se p é a probabilidade de sucesso em cada ensaio de Bernoulli, então o número de falhas Y tem função de probabilidade

$$f(y) = p \times \binom{r+y-1}{r-1} p^{r-1}(1-p)^y = \binom{r+y-1}{r-1} p^r (1-p)^y, \;\; y = 0, 1, 2, \dots$$

que depende de π e r inteiro positivo.

A distribuição binomial negativa pode ser definida por quaisquer valores positivos de r, usando-se a função gama ao invés de fatoriais:

$$f(y) = \frac{\Gamma(y+r)}{y!\,\Gamma(r)} p^r (1-p)^y, \qquad y = 0, 1, 2, \ldots$$

Em modelos lineares generalizados, usa-se a parametrização a seguir:

$$\mu = \frac{r(1-p)}{p}, \qquad k = \frac{1}{r}.$$

Consequentemente, a função de probabilidade de y apresenta-se:

$$f(y) = \frac{\Gamma\left(y + \frac{1}{k}\right)}{y!\,\Gamma\left(\frac{1}{k}\right)} \left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}} \left(\frac{k\mu}{1+k\mu}\right)^{y}, \qquad y = 0, 1, 2, \ldots$$

com

$$E(Y) = \mu, \qquad Var(Y) = \mu(1 + k\mu),$$

onde k é chamado de parâmetro de dispersão. A variável aleatória y tendo a distribuição acima é denotada por $y \sim NB(\mu, k)$.

(v) Normal

Em seguros e finanças, quantidades de interesse como número de sinistros, renda pessoal ou tempo para a ocorrência de um sinistro, são não negativos e, geralmente, têm distribuições que são predominantemente desviadas para a direita. A distribuição normal é, no entanto, importante para a análise de dados de seguro, uma vez que, muitas vezes pode ser possível a aplicação de uma transformação, tal como a transformação log, a fim de alcançar a normalidade (JONG e HELLER, 2008, p.26).

A função de probabilidade é

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}, -\infty < y < \infty,$$

onde μ é a média e σ o desvio padrão. A função de probabilidade normal é uma curva simétrica centrada em μ.

A notação $y \sim N(\mu, \sigma^2)$ é usada para indicar que y segue a distribuição normal e $N(0,1)$ é chamada de distribuição normal padrão.

(vi) Qui-quadrado e Gama

A distribuição qui-quadrado é a distribuição da soma dos quadrados das v variáveis aleatórias independentes $N(0,1)$, denotada por $Y \sim \chi_v^2$. O parâmetro v é o grau de liberdade e é maior do que zero. Tais variáveis aleatórias são não-negativas, e sua distribuição é inclinada para a direita. Possui média v e variância $2v$. Para grandes valores de v, Y se aproxima da distribuição normal.

Multiplicando uma variável aleatória χ_{2v}^2 por $\mu/2v$ teremos uma variável aleatória gama com parâmetros μ e v, denotada por $G(\mu, v)$. A gama é um ajuste bem adequado para acomodação de variáveis, tais quais, número de sinistros e rendimento anual. É contínua, não-negativa e inclinada para a direita, com a possibilidade de grandes valores de cauda superior.

A função de probabilidade $G(\mu, v)$ é

$$f(y) = \frac{y^{-1}}{\Gamma(v)}\left(\frac{yv}{\mu}\right)^v e^{-yv/\mu}, \qquad y > 0,$$

com $E(Y) = \mu,\ Var(Y) = \frac{\mu^2}{2}$.

Valores pequenos de v resultam em uma distribuição com uma longa cauda à direita, i.e., uma distribuição mais inclinada à direita.

(vii) Gaussiana inversa

A distribuição Gaussiana inversa é uma distribuição contínua com densidade similar à gama, porém com maior assimetria e maior pico. Possui dois parâmetros, μ e σ^2. Existem diversas parametrizações, mas usaremos

$$f(y) = \frac{1}{\sqrt{2\pi y^3}\sigma} exp\left\{-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right\}, \qquad y > 0,$$

denotada por Y $\sim IG(\mu, \sigma^2)$.

Possui média μ e variância $\sigma^2\mu^3$, σ^2 é o parâmetro de dispersão. Esta distribuição é usada em situações de extrema assimetria. A denominação Gaussiana inversa deriva da função cumulante, que tem uma relação inversa com a função cumulante da distribuição normal (JONG e HELLER, 2008, p.30).

3.3 FAMÍLIA EXPONENCIAL – RESPOSTA E ESTIMAÇÃO

Como já dito anteriormente, os MLGs pressupõem que a variável resposta tenha uma distribuição pertencente à família exponencial. Demétrio e Cordeiro (2010, p.1) salientam que:

> "O conceito de família exponencial foi introduzido na Estatística por Fisher, mas os modelos da família exponencial surgiram na Mecânica Estatística no final do século XIX e foram desenvolvidos por Maxwell, Boltzmann e Gibbs. A importância da família exponencial de distribuições teve maior destaque, na área dos modelos de regressão, a partir do trabalho pioneiro de Nelder e Wedderburn (1972) que definiram os modelos lineares generalizados (MLG)".

Todas as funções de probabilidade discutidas no último subtópico são da forma:

$$f(y) = c(y, \phi) exp\left\{\frac{y\theta - a(\theta)}{\phi}\right\} \qquad (3.3.1)$$

onde θ e ϕ são parâmetros. O parâmetro θ é chamado de parâmetro canônico e ϕ o parâmetro de dispersão.

Funções de probabilidade que podem ser escritas na forma (3.3.1) são chamadas de membros da família exponencial. Em termos de $a(\theta)$,

$$E(y) = \dot{a}(\theta),\ Var(y) = \phi\ddot{a}(\theta) \qquad (3.3.2)$$

onde $\dot{a}(\theta)$ e $\ddot{a}(\theta)$ são as primeiras e segundas derivadas de $a(\theta)$ com respeito a θ, respectivamente.

A Tabela 1 (abaixo) mostra diferentes escolhas de θ e $a(\theta)$. Provas das relações podem ser encontradas em Jong e Heller (2008, p.37).

Tabela 1. Distribuições de família exponencial e seus parâmetros

Distribuição	$\boldsymbol{\theta}$	$\boldsymbol{a(\theta)}$	$\boldsymbol{\phi}$	$\boldsymbol{E(y)}$	$\boldsymbol{var(\mu) = \frac{var(y)}{\phi}}$
$B(n,p)$	$\ln\frac{p}{1-p}$	$n\ln(1+e^{\theta})$	1	np	$np(1-p)$
$P(\mu)$	$\ln\mu$	e^{θ}	1	μ	μ
$N(\mu,\sigma^2)$	μ	$\frac{1}{2}\theta^2$	σ^2	μ	1
$G(\mu,v)$	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\frac{1}{v}$	μ	μ^2
$IG(\mu,\sigma^2)$	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2	μ	μ^3
$NB(\mu,\kappa)$	$\ln\frac{\kappa\mu}{1+\kappa\mu}$	$-\frac{1}{\kappa}\ln(1-\kappa e^{\theta})$	1	μ	$\mu(1+\kappa\mu)$

Fonte: Jong e Heller (2008, p.36)

3.3.1 A função variância

Para distribuições resposta de família exponencial,

$$\ddot{a}(\theta) = \frac{\partial\dot{a}(\theta)}{\partial\theta} = \frac{\partial\mu}{\partial\theta} \equiv V(\mu), \qquad (3.3.3)$$

e portanto pode-se sempre escrever $Var(Y) = \phi V(\mu)$ onde $V(\mu)$ é chamado função variância, indicando relação entre média e variância. Paula (2010, p.4) sustenta que:

> "A função de variância desempenha um papel importante na família exponencial, uma vez que a mesma caracteriza a distribuição. Isto é, dada a função de variância, tem-se uma classe de distribuições correspondentes, e vice-versa. Essa propriedade permite a comparação de distribuições através de testes simples para a função de variância".

A função variância $V(\mu)$ é uma construção fundamental. Em MLG, a média μ varia com as variáveis explicativas e o mesmo acontece com a variância. Um modelo que conecta a média com as variáveis explicativas é, ao mesmo tempo, um modelo de relacionamento entre a variância e as variáveis explicativas (JONG e HELLER, 2008, p.36).

Deve ser enfatizado que há muitas funções $V(\mu)$ que não podem surgir a partir de uma distribuição de família exponencial. Esta questão é discutida mais detalhadamente em Jong e Heller (2008, p.94-96).

3.4 ESTIMAÇÃO DOS PARÂMETROS

As funções de probabilidade $f(y)$ discutidas anteriormente possuem um ou dois parâmetros, valores dos quais geralmente são desconhecidos. As funções são tipicamente ajustadas a uma amostra de dados $y_1, \dots, y_n$, isto é, os parâmetros são estimados com base na amostra. Na discussão abaixo é importante entender que neste subtópico é assumido que cada observação y_i parte exatamente da mesma distribuição, isto é, um dado membro da família exponencial com parâmetros fixos, porém desconhecidos.

3.4.1 Estimação de máxima verossimilhança (EMV)

A estimação de máxima verossimilhança é baseada em escolher parâmetros estimados que maximizam a verossimilhança da amostra observada $y_1, \dots, y_n$.

Se os y_i são independentes e identicamente distribuídos, i.i.d., então possuem função de probabilidade conjunta:

$$f(y_i; \theta, \phi) = \prod_{i=1}^{n} f(y_i; \theta, \phi).$$

A verossimilhança da amostra $(y_1, \dots, y_n)$ é a expressão acima considerada como uma função de θ e ϕ. A log-verossimilhança $\ell(\theta, \phi)$ é o logaritmo da verossimilhança:

$$\ell(\theta, \phi) = \sum_{i=1}^{n} \ln\, f(y_i; \theta, \phi).$$

O método de máxima verossimilhança escolhe os valores de θ e ϕ que maximizam a verossimilhança, ou equivalentemente, a log-verossimilhança. Os estimadores de máxima verossimilhança de θ e ϕ são denotados por $\hat{\theta}$ e $\hat{\phi}$, respectivamente.

Quando $f(y_i; \theta, \phi)$é uma função de probabilidade com família exponencial então $\ell(\theta, \phi)$ é

$$\sum_{i=1}^{n}\left\{\ln c(y_i,\phi)+\frac{y_i\theta-a(\theta)}{\phi}\right\}=\frac{n\{\bar{y}\theta-a(\theta)\}}{\phi}+\sum_{i=1}^{n}\ln c(y_i,\phi).$$

Diferenciando $\ell(\theta,\phi)$ com relação a θ e igualando a zero leva à primeira condição para a maximização de verossimilhança

$$\frac{n\{\bar{y}-\dot{a}(\theta)\}}{\phi}=0\ \Rightarrow\ \dot{a}(\theta)=\bar{y}.$$

Consequentemente, o estimador de máxima verossimilhança de θ é obtido encontrando θ tal que $\dot{a}(\theta)\equiv\mu$. Portanto, para qualquer distribuição de família exponencial, $\hat{\mu}=\bar{y}$.

(i) Propriedades dos estimadores de máxima verossimilhança

Qualquer estimador $\hat{\theta}$ depende dos valores amostrais, e irá variar de amostra para amostra, extraídas da mesma população. Um estimador é assim uma variável aleatória, e duas importantes propriedades de um estimador são o viés e a variância. Um estimador $\hat{\theta}$ é não-viesado (ou não viciado) se $E(\hat{\theta})=\theta$. A variância de um estimador indica sua precisão, em que um estimador não-viesado com variância pequena é suscetível de produzir estimativas confiáveis, enquanto que um estimador com variância alta produz pouca precisão.

Suponha que $\hat{\theta}$ é o estimador de máxima verossimilhança, EMV, do parâmetro θ. Propriedades desejáveis possuídas por um EMV incluem:

a) Invariância: Se h é uma função monotônica e $\hat{\theta}$ é o EMV de θ, então $h(\hat{\theta})$ é o EMV de $h(\theta)$.

b) Assintoticamente não-viesado: O valor esperado $\mathrm{E}(\hat{\theta})$ aproxima de θ quando o tamanho da amostra aumenta. Ou seja, $E(\theta)\xrightarrow[n\to\infty]{}\theta$.

c) Consistência: à medida que o tamanho da amostra aumenta a distribuição de probabilidade $\hat{\theta}$ aproxima de θ.

d) Variância mínima: Na classe de todos os estimadores, para amostras grandes, $\hat{\theta}$ tem variância mínima e, portanto, é o estimador mais preciso possível.

Os EMV possuem algumas desvantagens, tais quais: maior o viés em amostras pequenas e dificuldade computacional, pois nem sempre são fáceis de calcular manualmente (geralmente precisa-se de iterações computacionais) (JONG e HELLER, 2008, p.41).

3.5 O MODELO LINEAR GENERALIZADO

Demétrio e Cordeiro (2010, p.23) enfatizam que:

> "A seleção de modelos é uma parte importante de toda pesquisa em modelagem estatística e envolve a procura de um modelo que seja o mais simples possível e que descreva bem o processo gerador dos valores observados que surgem em diversas áreas do conhecimento".

Eles também defendem que os modelos lineares generalizados envolvem uma variável resposta univariada, variáveis explicativas e uma amostra aleatória de n observações independentes, onde

a) A variável resposta é o componente aleatório do modelo e possui distribuição: normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens;
b) Variáveis explicativas como uma estrutura linear, constituindo o componente sistemático do modelo;
c) A ligação entre os componentes aleatório e sistemático é feita de uma função denominada função de ligação.

Dada uma resposta Y, O MLG é dado por

$$f(y) = c(y,\phi)exp\left\{\frac{y\theta - a(\theta)}{\phi}\right\}, \qquad g(\mu) = x'\beta. \qquad (3.5.1)$$

A equação para $f(y)$ especifica que a distribuição da resposta é da família exponencial. A segunda equação apresenta que uma transformação da média, $g(\mu)$, é linearmente relacionada com as variáveis explicativas contidas em x e β é o vetor de parâmetros desconhecidos (DEMÉTRIO e CORDEIRO, 2010, p.41).

Para ser feita a modelagem linear generalizada, é preciso considerar os seguintes pontos:

(i) A escolha de $a(\theta)$ determina a distribuição resposta.
(ii) A escolha de $g(\mu)$, chamada de função de ligação, determina como a média está relacionada com as variáveis explicativas x. No modelo linear normal, a relação entre a média e as variáveis explicativas é $\mu = x'\beta$. Em MLG, isto é generalizado para $g(\mu) = x'\beta$, onde g é uma função monotônica diferenciável (tal como log ou raiz quadrada).

(iii) A configuração em (3.5.1) afirma que, dado x, μ é determinado através de $g(\mu)$. Dado μ, θ é determinado através de $\dot{a}(\theta) = \mu$. Dado θ, y é determinado a partir da densidade exponencial especificada em $a(\theta)$.

(iv) Observações em y são i.i.d, ou seja, como dito anteriormente, independentes e identicamente distribuídas.

3.5.1 Etapas da modelagem linear generalizada

Dada uma variável resposta Y, contruir um MLG consiste dos seguintes passos:

(i) Escolher uma distribuição resposta $f(y)$ em (3.5.1). A distribuição resposta é adaptada à situação dada.

(ii) Escolher uma função de ligação $g(\mu)$.

(iii) Escolher variáveis explicativas x em termos de $g(\mu)$. Considerações semelhantes se aplicam como na modelagem de regressão comum.

(iv) Coletar observações $y_1, \dots, y_n$ na resposta Y e os valores correspondentes $x_1, \dots, x_n$ das variáveis explicativas X. Observações sucessivas são i.i.d., i.e., a amostra será considerada como uma amostra aleatória da população estudada.

(v) Ajustar o modelo por estimação de β e de ϕ (se este for desconhecido). O ajuste é geralmente feito utilizando softwares estatísticos, que implementam estimações de máxima verossimilhança e suas variantes.

(vi) Dada a estimativa de β, gerar predições (ou valores ajustados) de Y para diferentes configurações de X e examinar como os modelos se ajustam examinando a saída dos valores ajustados com os valores reais, bem como outros diagnósticos do modelo. Além disso, o valor estimado de β será utilizado para descobrirmos se as variáveis explicativas são importantes ou não na determinação de μ.

A escolha de $a(\theta)$ é guiada pela natureza da variável resposta. A escolha da ligação é sugerida pela forma funcional das relações entre a resposta e as variáveis explicativas.

3.5.2 Funções de ligação

Funções de ligação comumente utilizadas são apresentadas na Tabela 2. Com exceção da função *logit*, as funções de ligação são da forma $g(\mu) = \mu^p$, com o caso logarítmico sendo o limite de $(y^p - 1)/p$ com $p \to 0$.

Se $g(\mu) = \theta$ então g é chamada de ligação canônica correspondente a $a(\theta)$. Neste caso $\theta = x'\beta$. A escolha da ligação canônica g correspondente à distribuição resposta f simplifica a estimação, embora com a computação moderna isso não seja mais uma consideração primordial (JONG e HELLER, 2008, p.66).

Tabela 2. Ligações comumente utilizadas

Função de ligação	$\boldsymbol{g(\mu)}$	**Ligação canônica para**
Identidade	μ	Normal
Log	$\ln \mu$	Poisson
Potência	μ^p	Gama $(p = -1)$ Gaussiana inversa $(p = 2)$
Raiz quadrada	$\sqrt{\mu}$	
Logit	$\ln \frac{\mu}{1-\mu}$	Binomial

Fonte: Jong e Heller (2008, p.67)

3.5.3 Estimação de máxima verossimilhança

O estimador de máxima verossimilhança, EMV, de β e ϕ são derivados pela maximização da log-verossimilhança, definida similarmente à seção 3.4.2, como

$$\ell(\beta, \phi) = \sum_{i=1}^{n} \ln f(y_i; \beta, \phi) = \sum_{i=1}^{n} \left\{ \ln c(y_i, \phi) + \frac{y_i \theta_i - a(\theta_i)}{\phi} \right\}, \qquad (3.5.2)$$

que assume família exponencial independente de resposta y_i.

Considere o EMV de β_j. Para encontrar o máximo, $\ell(\beta, \phi)$ é diferenciado em relação a β_j:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j},$$

onde

$$\frac{\partial \ell}{\partial \theta_i} = \frac{y_i - \dot{a}(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi}, \qquad \frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial n_i}\frac{\partial n_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial n_i} x_{ij}.$$

Aqui, $n_i = x'_i\beta$ e x_{ij} é a componente i de x_j. Fixando-se $\partial \ell / \partial \beta_j = 0$, segue implícita nas condições de primeira ordem para a maximização de verossimilhança:

$$\sum_{i=1}^{n} \frac{\partial \theta_i}{\partial n_i} x_{ij}(y_i - \mu_i) = 0 \iff X'D(y - \mu) = 0, \qquad (3.5.3)$$

onde D é a matriz diagonal com entradas $\partial \theta_i / \partial n_i$,

$$\left(\frac{\partial \theta_i}{\partial n_i}\right)^{-1} = \frac{\partial n_i}{\partial \theta_i} = \frac{\partial n_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \theta_i} = \dot{g}(\mu_i)\ddot{a}(\theta_i) = \dot{g}(\mu_i)V(\mu_i).$$

Assim, D é a diagonal com entradas $\{\dot{g}(\mu_i)V(\mu_i)\}^{-1}$. As equações em (3.5.3) são chamadas de equações de estimação para β.

Definindo as matrizes diagonais G e W com entradas diagonais $\dot{g}(\mu_i)$ e $\{\dot{g}(\mu_i)^2 V(\mu_i)\}^{-1}$, respectivamente, então $D = WG$ é equivalente a

$$X'WG(y - \mu) = 0. \qquad (3.5.4)$$

(i) Iteração de Newton-Raphson

As condições de primeira ordem (3.5.3) para a maximização de verossimilhança são normalmente difíceis de resolver diretamente, exceto para casos como o da normal, com a ligação identidade. Uma sugestão proveitosa assume que a primeira e segunda derivada da função a ser maximizada pode ser facilmente avaliada em cada ponto (JONG e HELLER, 2008, p.69).

Suponha ϕ conhecido e $\ell(\beta)$ a log-verossimilhança como uma função de parâmetros desconhecidos β. Se β contém um único parâmetro, então a aproximação por séries de Taylor no ponto β é

$$\ell(\beta + \delta) \approx \ell(\beta) + \ell(\beta)\delta + \frac{\delta^2}{2}\ddot{\ell}(\beta).$$

Diferenciando o lado direito em relação a δ e igualando a 0, tem-se que

$$\dot{\ell}(\beta) + \delta\ddot{\ell}(\beta) = 0 \implies \delta = -\left\{\ddot{\ell}(\beta)\right\}^{-1}\dot{\ell}(\beta).$$

Denotando-se $\beta^{(m)}$ como o valor para β na iteração m, a equação de atualização é

$$\beta^{(m+1)} = \beta^{(m)} - \left\{\ddot{\ell}\left(\beta^{(m)}\right)\right\}^{-1}\dot{\ell}\left(\beta^{(m)}\right). \qquad (3.5.7)$$

Para o máximo, $\ddot{\ell}(\beta) < 0$.

A abordagem pode ser adaptada quando β é um vetor. Neste caso uma aproximação quadrática é feita para a superfície $\ell(\beta)$, a qual será maximizada. A equação de atualização é como em (3.5.7) tendo inversa como a matriz de inversão e $\dot{\ell}(\beta)$ é o vetor de derivadas parciais de ℓ em relação a β_j, chamado vetor score. $\ddot{\ell}(\beta)$ é a matriz Hessiana. A condição para um máximo é que a Hessiana seja não-positiva definida, i.e., $-\ddot{\ell}(\beta)$ é não-negativa definida. O procedimento de repetidamente avaliar o score e a Hessiana para melhorar a estimação é chamado de iteração de Newton-Raphson.

3.5.4 Intervalo de confiança e predição

Dados os valores das variáveis aleatórias, x, o valor estimado da média de Y é $\hat{\mu}$ onde $g(\hat{\mu}) = x'\hat{\beta}$. Um intervalo de confiança em torno da estimação é usado para indicar precisão. A computação disto requer a distribuição amostral de $\hat{\mu}$. A variância de $x'\hat{\beta}$ é

$$Var\left(x'\hat{\beta}\right) = \phi x'(X'WX)^{-1}x.$$

Deste modo, um intervalo de confiança para a média é (μ_ℓ, μ_u), onde

$$g(\mu_\ell) = x'\hat{\beta} - Z\sqrt{\phi x'(X'WX)^{-1}x},$$

e μ_u similarmente definido com o sinal positivo. Z é o ponto aproximado da distribuição N(0,1). O intervalo de confiança é exato no caso de uma ligação identidade e resposta normal. Em outros casos, é uma aproximação. A dispersão ϕ é substituída por uma estimativa, provocando erros de uma maior aproximação (JONG e HELLER, 2008, p.71).

Um intervalo construído para um caso atual é chamado de predição intervalar. A largura de tal intervalo é baseada na incerteza associada ao $\hat{\mu}$ e no resultado da distribuição resposta.

3.5.5 Avaliando ajuste e desvio

A qualidade do ajuste de um modelo de dados é uma questão natural resultante de todas as modelagens estatísticas. Os princípios de teste de significância, seleção de modelos e testes de diagnóstico, são os mesmos para MLG como para regressão normal, no entanto, os detalhes técnicos dos métodos diferem um pouco (JONG e HELLER, 2008, p.71).

Uma forma de avaliar o ajuste de um determinado modelo é compará-lo ao modelo com o melhor ajuste possível. O melhor ajuste é obtido quando há muitos parâmetros e observações (o chamado modelo saturado). Um modelo saturado irá garantir que há total flexibilidade no ajuste θ_i. A partir de

$$\frac{\partial \ell}{\partial \theta_i} = \frac{y_i - \mu_i}{\phi} = \frac{y_i - \dot{a}(\theta_i)}{\phi},$$

o EMV de θ_i no modelo saturado é $\breve{\theta}_i$, onde $\dot{a}(\breve{\theta}_i) = y_i$. Assim, cada valor ajustado é igual à observação e o modelo saturado possui ajuste perfeito. O valor da log-verossimilhança saturada é

$$\breve{\ell} \equiv \sum_{i=1}^{n} \left\{ \ln c(y_i, \phi) + \frac{y_i \breve{\theta}_i - a(\breve{\theta}_i)}{\phi} \right\},$$

que é a máxima log-verossimilhança possível para y dada a distribuição resposta especificada por $a(\theta)$. Este valor é comparado com o $\hat{\ell}$, o valor de máximo da log-verossimilhança baseada em y e dadas as variáveis explicativas. O desvio, denotado por Δ, é definido como uma medida de distância entre o modelo saturado e o modelo ajustado:

$$\Delta \equiv 2(\breve{\ell} - \hat{\ell}).$$

Quando o modelo fornece um bom ajuste, $\hat{\ell}$ é próximo de $\breve{\ell}$. O tamanho de Δ é avaliado em relação à distribuição χ^2_{n-p}, que é a distribuição amostral do desvio, assumindo que o modelo ajustado está correto e n é grande. O valor esperado do desvio é $n - p$, e normalmente o desvio dividido pelo seu grau de liberdade é examinado: um valor muito maior indica um modelo deficientemente apropriado.

Um cálculo direto mostra que, para a família exponencial,

$$\Delta = 2 \sum_{i=1}^{n} \left\{ \frac{y_i(\breve{\theta}_i - \hat{\theta}_i) - a(\breve{\theta}_i) + a(\hat{\theta}_i)}{\phi} \right\}. \qquad (3.510)$$

Quando ϕ é desconhecido e estimado, a distribuição χ^2_{n-p} para o desvio é comprometida. No caso da distribuição de Poisson $\phi = 1$, a aproximação χ^2 é eficiente. No caso da distribuição normal, quando σ^2 é conhecido, a distribuição χ^2 para o desvio é exata; no entanto, quando σ^2 é estimado, não se pode contar com tal afirmação (JONG e HELLER, 2008, p.72).

A Tabela 3 disponibiliza as expressões para o desvio de distribuições de família exponencial:

Tabela 3. Desvio para distribuições resposta de família exponencial

Distribuição	**Desvio Δ**
Normal	$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}$
Binomial	$2\sum_{i=1}^{n} n_i\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{1 - \hat{\mu}_i}\right)\right\}$
Gamma	$2v\sum_{i=1}^{n}\left\{-\ln\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right\}$
Gaussiana inversa	$\frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i{}^2 y_i}$
Binomial negativa	$2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i + \frac{1}{k}\right)\ln\left(\frac{y_i + 1/k}{\hat{\mu}_i + 1/k}\right)\right\}$

Fonte: Jong e Heller (2008, p.73).

3.5.6 Testando a significância das variáveis explicativas

Hipóteses são escritas como $C\beta = r$ onde C é a matriz conhecida de hipóteses e r o conjunto de valores dados. Existem três abordagens para tal teste da forma acima, e cada abordagem considera a verossimilhança ou log-verossimilhança como em (3.5.3). Escreva $\hat{\beta}$ como o EMV de β, e denote $\tilde{\beta}$ o EMV de β quando ℓ é maximizada sujeita às restrições $C\beta = r$.

(i) Teste razão de verossimilhança

É feita uma comparação entre $\hat{\ell}$ e $\tilde{\ell}$. Um valor de $\hat{\ell}$ muito maior do que $\tilde{\ell}$ é uma prova contra as restrições. A razão de verossimilhança é definida como $\lambda = \hat{L}/\tilde{L}$ onde $\hat{L}$ e $\tilde{L}$ são a verossimilhança dos modelos irrestrito e restrito, respectivamente. O teste de razão de verossimilhança é

$$2\ln\lambda = 2\left(\hat{\ell} - \tilde{\ell}\right). \qquad (3.5.11)$$

A expressão acima é sempre não-negativa e possui a distribuição χ_q^2 se $C\beta = r$, onde q é o número de linhas de C, i.e., o número de restrições em β. Se $2\ln\lambda$ é próximo de zero, então o modelo restrito é quase tão bom quanto o modelo irrestrito. A região de rejeição para o teste é a parte superior da distribuição χ_q^2.

(ii) Teste Wald

Este teste mede o quão longe $C\hat{\beta}$ está de r, com uma grande distância fornecendo evidências contra as restrições. O estimador $\hat{\beta}$ é necessário, mas não $\tilde{\beta}$. Se $C\beta = r$, então a partir de $\hat{\beta} \sim \mathrm{N}\{\beta, \phi(\mathrm{X'WX})^{-1}\}$ segue que

$$C\hat{\beta} - r \sim \mathrm{N}\{0, \phi\mathrm{C}(\mathrm{X'WX})^{-1}\mathrm{C'}\}.$$

Isto conduz à estatística de Wald para testar $C\beta = r$:

$$\left(C\hat{\beta} - r\right)'\{\phi\mathrm{C}(\mathrm{X'WX})^{-1}\mathrm{C'}\}^{-1}\left(\mathrm{C}\hat{\beta} - \mathrm{r}\right) \sim \chi_{\mathrm{q}}^2. \qquad (3.5.12)$$

Na prática, W é substituído por uma estimativa e, portanto, a distribuição χ_{q}^2 é aproximada.

a) Testando coeficientes individuais: Quando testamos $\beta_{\mathrm{j}} = \mathrm{r}$, a matriz C é um vetor linha de zeros exceto na posição j onde é igual a 1. Ele assume que todas as outras variáveis explicativas em X estão no modelo. O termo em chaves em (3.5.12) se reduz a uma única entrada diagonal de $\phi(\mathrm{X'WX})^{-1}$, a variância de $\hat{\beta}_{\mathrm{j}}$, que é denotada por $\phi\psi_{\mathrm{j}}$. A estatística de Wald torna-se então

$$\frac{\left(\hat{\beta}_{\mathrm{j}} - \mathrm{r}\right)^2}{\phi\psi_{\mathrm{j}}} \sim \chi_1^2.$$

Onde ϕ é desconhecido, ele é substituído por sua estimativa $\widehat{\phi}$. A raiz quadrada da estatística de Wald é relatada por alguns softwares estatísticos, com valor-p calculado na distribuição normal padrão ou t.

c) Testando todos os coeficientes: Para o teste global em que todos os coeficientes com exceção do intercepto são zero, C é a matriz $(p-1) \times p$:

$$C = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

3.5.7 Resíduos

Resíduos são usados para verificar a adequação da distribuição de resposta escolhida, e para a identificação de *outliers*. Os resíduos $\hat{\epsilon}_i = y_i - \hat{y}_i$, são centrais para verificar o modelo com o modelo linear normal. Para o MLG, em qualquer distribuição de resposta diferente da normal, estes resíduos não são nem normalmente distribuídos, nem têm variância constante. A definição de resíduo é ampliada a partir da noção de diferença entre os valores observados e ajustados, para uma quantificação mais geral da conformidade de um caso para a especificação do modelo.

(i) Desvios residuais

Suponha δ_i^2 denotando o termo Δ dado em (3.5.6)

$$\delta_i^2 \equiv \frac{2\{y_i(\breve{\theta}_i - \hat{\theta}_i) - a(\breve{\theta}_i) + a(\hat{\theta}_i)\}}{\phi}.$$

Então, δ_i é chamado desvio residual e mensura a raiz quadrada da contribuição do caso i para o desvio (3.5.6).

Se uma contribuição particular δ_i é grande, então o caso i está contribuindo muito para o desvio, indicando um afastamento dos pressupostos do modelo para esse caso. Ou seja, indica que o modelo não está representando fielmente os dados, para este caso. Se o modelo

está correto e n é grande, o desvio é aproximadamente χ^2_{n-p}. O valor esperado do desvio é portanto $n-p$, e espera-se que cada caso contribua com $(n-p)/n \approx 1$ para o desvio. Por este motivo, $|\delta_i|$ muito maior é uma indicação de que o caso i contribui para uma falta de ajuste. De acordo com Jong e Heller (2008, p.78), "Isso sugere uma má especificação do modelo ou um erro de dados". Normalmente, desvios residuais são examinados plotando-os contra valores ajustados ou variáveis explicativas.

3.5.8 Outras ferramentas de diagnóstico

(i) Verificando a função de ligação

A expansão de primeira ordem por séries de Taylor de $g(y_i)$, como em (3.5.5) é

$$g(y_i) \approx g(\mu_i) + \dot{g}(\mu_i)(y_i - \mu_i) \approx x_i'\beta. \qquad (3.5.13)$$

Plotando-se $g(\hat{\mu}_i) + \dot{g}(\hat{\mu}_i)(y_i - \hat{\mu}_i)$ contra $x_i'\hat{\beta}$ deve produzir pontos que aproximadamente ficam em uma linha reta, e curvatura forte sugere que a função de ligação g está incorreta.

(ii) Identificando outliers

Influência para modelos lineares generalizados é mensurada usando a matriz:

$$H = W^{\frac{1}{2}}X(X'WX)^{-1}X'W^{\frac{1}{2}},$$

que é a mesma matriz usada para a regressão linear normal, definida em Jong e Heller (2008, p.59), com X substituído por $W^{\frac{1}{2}}X$.

3.5.9 Seleção de modelos

O princípio do viés equilibrado e da variância aplica-se para o MLG da mesma maneira como para o modelo linear normal.

Cada variável explicativa adicionada a um modelo melhora o ajuste. No entanto a adição de variáveis injustificadas diminui a precisão das estimativas de parâmetros. Embora uma variável explicativa possa ser estatisticamente significativa, acrescentar ao modelo pode

não valer a pena, porque a melhoria no ajuste pode ser compensada pela perda de precisão estimativa. O conceito central é a troca viés-variância. Viés refere-se à falta de ajuste. Variância neste contexto significa a variância das estimativas de regressão.

Aumentar o número de variáveis explicativas aumenta o número de parâmetros p e diminui os erros de ajuste. Um grande p implica em viés baixo. No entanto, quanto mais parâmetros no modelo, maiores as variações de β_j. O objetivo é encontrar um modelo de compromisso entre:

(i) Um grande número de parâmetros de ajuste. Neste caso, estimar parâmetros β_j tem baixa precisão (desvio padrão elevado);

(ii) Um número pequeno de parâmetros com um ajuste pior. Neste caso, estimar parâmetros β_j tem precisão relativamente alta (desvio padrão baixo).

Há inúmeros critérios que se equilibram melhor no ajuste de um modelo com um termo penalidade para o número de parâmetros. Os mais conhecidos são o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC):

$$AIC \equiv -2\ell + 2p, \qquad BIC \equiv -2\ell + p \ln n.$$

Onde ℓ é a log-verossimilhança do modelo dado em (3.5.2). Um bom ajuste forma um alto valor para a verossimilhança, e consequentemente um baixo valor para -2ℓ.

Os primeiros dois termos na expressão para ℓ são geralmente omitidos, rendendo

$$AIC = \frac{S}{\sigma^2} + 2p, \qquad BIC = \frac{S}{\sigma^2} + p \ln n,$$

onde S é a soma dos quadrados dos resíduos. O termo S/σ^2 é o termo viés que quantifica o ajuste do modelo para as observações. Um bom ajuste produz um valor baixo para S. O segundo termo de cada um dos critérios é o termo variância, penalizando para o número de parâmetros do modelo. Um modelo com um grande número de parâmetros e um bom ajuste tem um valor baixo para o viés, mas o critério cresce pelo termo variância. Por outro lado um modelo com p pequeno tem maior viés, mas menor variância. O modelo com o menor valor de AIC ou BIC é selecionado.

A decisão tem de ser feita independentemente de usar AIC ou BIC. Este último aplica uma maior penalidade para o número de parâmetros, de modo que tende-se a escolher modelos com menos variáveis explicativas em comparação com o AIC. Quando *n* é grande, como é o caso da maioria dos conjuntos de dados de seguro, o BIC tende a selecionar modelo que a maioria dos analistas considera muito simples. Neste caso, o AIC é preferível (JONG e HELLER, 2008, p.63).

Comparação de modelos usando AIC ou BIC deve basear-se no mesmo conjunto de observações. Isso impacta no momento de decidir entre os modelos que incluem pelo menos uma variável explicativa com valores faltantes.

3.6 O MODELO DE POISSON

Quando a variável resposta é contável, o modelo utilizando a distribuição de Poisson é normalmente utilizado. O modelo de regressão Poisson é:

$$Y \sim P(\mu), \quad g(\mu) = x'\beta. \qquad (3.6.1)$$

Escolhas populares para $g(\mu)$ são a função identidade $\mu = x'\beta$ e a função log $\ln \mu = x'\beta$. A seguir é apresentado um exemplo descrito por Jong e Heller (2008, p.82).

3.6.1 Exemplo

(i) Número de crianças

Considere o número de filhos de uma amostra de 141 mulheres. Neste exemplo, y, μ e x são o número de filhos, o número esperado de filhos e a idade das mães, respectivamente. Os modelos com as ligações log e identidade apresentam os ajustes

$$\ln \hat{\mu} = -4{,}090 + 0{,}113x \quad \Rightarrow \quad \hat{\mu} = e^{-4{,}090+0{,}113x}$$

e

$$\hat{\mu} = -0{,}965 + 0{,}057x,$$

respectivamente. Os resultados dos modelos são resumidos na Tabela 4.

Tabela 4. Modelos de Poisson para números de filhos

VARIÁVEL RESPOSTA		Número de filhos	
DISTRIBUIÇÃO RESPOSTA		Poisson	
LIGAÇÃO		Log	
DESVIO		165,0	
GRAUS DE LIBERDADE		139	
PARÂMETROS	$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{\chi^2}$	**P valor**
Intercepto	-4,090	32,84	<0,0001
Idade	0,113	28,36	<0,0001

VARIÁVEL RESPOSTA		Número de filhos	
DISTRIBUIÇÃO RESPOSTA		Poisson	
LIGAÇÃO		Identidade	
DESVIO		171,4	
GRAUS DE LIBERDADE		139	
PARÂMETROS	$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{\chi^2}$	**P valor**
Intercepto	-0,965	4,40	0,0359
Idade	0,057	13,39	0,0003

Fonte: Jong e Heller (2008, p.83)

De acordo com a Tabela 4, os desvios para ambos os modelos apresentam ajustes adequados: 165 para o modelo de ligação log e 171 para o modelo de ligação identidade, os dois com 139 graus de liberdade.

4 RESULTADOS

Foram utilizados, inicialmente, 17.006 dados correspondentes às quantidades de consultas para o ano de 2001 por usuários que possuíam tempo de plano superior a 1 ano. No entanto ao realizar a análise, foram identificadas algumas inconsistências impossíveis de se corrigir, ocasionando na exclusão das ocorrências, o que acabou resultando em 16.865 observações.

Para a modelagem, foi escolhido o software R (ver: http://cran.r-project.org/), programa de código aberto amplamente utilizado no meio estatístico.

Tabela 5: Resumo das variáveis

Variável	**Média**	**Desvio padrão**	**0%**	**25%**	**50%**	**75%**	**100%**
Tempo no plano	5,58	3,28	1	3	6	8	30
Idade	33,57	19,72	0	17	34	48	96
Quantidade de consultas	4,08	3,55	0	1	3	6	33
sexo	0,55	0,50	0	0	1	1	1
Tipo de acomodação	0,43	0,50	0	0	0	1	1

A Tabela 5 representa o resumo numérico das variáveis. Vale salientar que “sexo” e “tipo de acomodação” são variáveis dummy, assumindo valores 0 o 1 para sexo masculino ou sexo feminino, respectivamente, e 0 ou 1 para acomodação em apartamento ou enfermaria, concomitantemente.

A variável a ser estudada é “quantidade de consultas”, que apresenta valores numéricos discretos e não negativos. Tais características implicam em escolhermos inicialmente uma distribuição de resposta Poisson com função de ligação log.

Primeiramente, analisamos na Figura 1 o histograma da variável resposta escolhida:

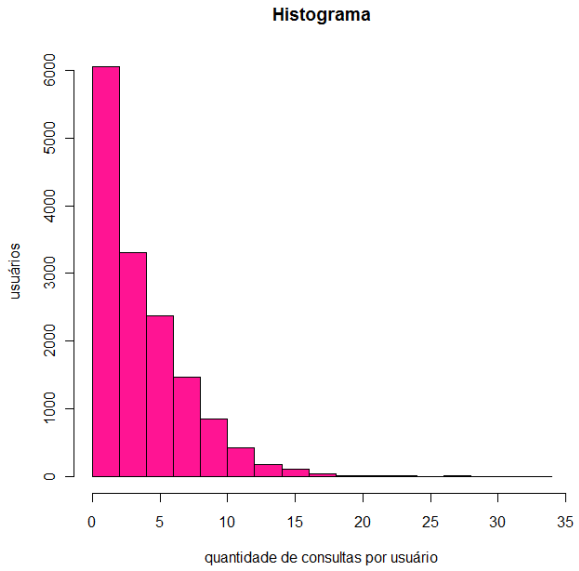

Figura 1. Histograma de quantidades de consultas por usuário

Observe que os dados apresentam uma grande quantidade de zeros e possuem um comportamento decrescente.

Primeiramente, aplicou-se:

$$Quantidade\ de\ consultas \sim\ idade + factor(sexo) +\ tempo\ no\ plano + factor(acomodação)$$

A tabela 6 apresenta os coeficientes estimados no modelo.

Tabela 6: Coeficientes estimados no modelo

VARIÁVEIS	**ESTIMATIVA**	**ERRO PADRÃO**	**VALOR z**	**P > (|Z|)**
INTERCEPTO	1,0195521	0,0132257	77,089	<2e-16 ***
IDADE	0,0067489	0,0002051	32,908	<2e-16 ***
SEXO	0,4343789	0,0085916	50,558	<2e-16 ***
TEMPO NO PLANO	-0,0122691	0,0012910	-9,503	<2e-16 ***
TIPO DE ACOMODAÇÃO	-0,1105618	0,0086387	-12,798	<2e-16 ***

Grau de significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

De acordo com a Tabela 6, o intercepto está incluso no modelo, assim como as outras variáveis. Como pode ser observado, o tempo no plano é inversamente proporcional à quantidade de consultas, ou seja, quanto maior o tempo, menor a quantidade. Também verificou-se que a variável dummy tipo de acomodação (igual a 1, correspondente à enfermaria) apresentou comportamento inverso à resposta. Geralmente, planos de enfermaria são mais baratos do que planos de apartamento. Isso sugere que usuários que desembolsam menores valores para o pagamento do plano são menos propensos a consultas.

O valor dos desvios residuais foi 40.584, em 16.894 graus de liberdade (gl), o que sugere que não é provável que tal modelo tenha gerado os dados. Como apresentado anteriormente no texto e especificado por Jong e Heller (2008, p.78), valores elevados sugerem uma "má especificação do modelo ou um erro de dados".

A Figura 2 (apresentada abaixo) ilustra a relação entre os valores ajustados e os resíduos:

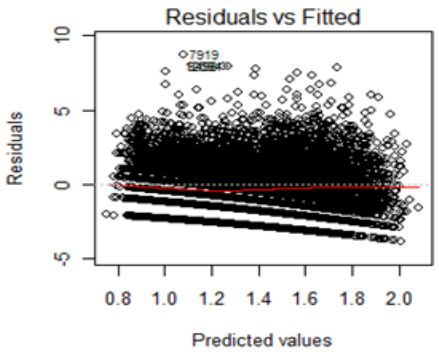

Figura 2. Gráficos dos Resíduos X Valores ajustados

A Figura 2 mostra os dados bastante agrupados, além de uma tendência decrescente visível na parte inferior do gráfico.

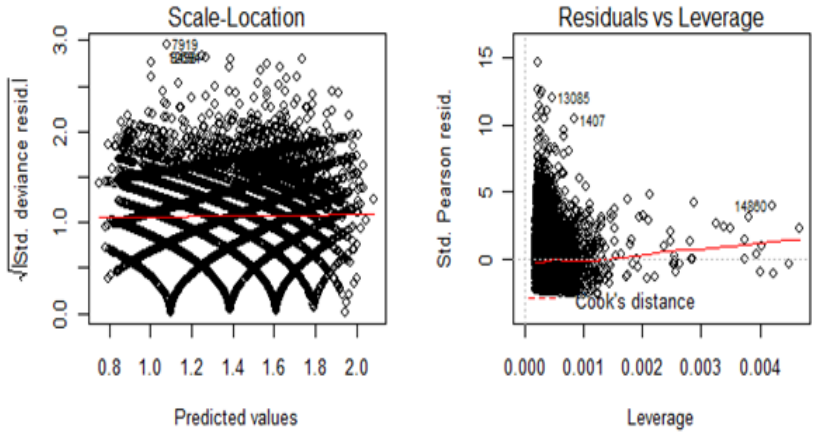

Figura 3. Gráficos de Locação-Escala e Alavanca

Na Figura 3, o primeiro gráfico (Locação-escala) sugere uma tendência na parte inferior. O segundo gráfico (alavanca) mostra que existem muitos pontos distantes (como o 14860) que podem influenciar nas estimativas dos coeficientes e dos valores ajustados.

A fim de encontrar um melhor ajuste, os dados foram separados em dois grupos: sexo feminino e sexo masculino. A Tabela 7 apresenta o resumo das variáveis para cada grupo.

Tabela 7: Resumo das variáveis para os dois grupos

Variável	**Média**	**Desvio padrão**	**0%**	**25%**	**50%**	**75%**	**100%**
Tempo no plano	M: 5,64 F: 5,54	M: 3,25 F: 3,29	M: 1 F: 1	M: 3 F: 3	M: 6 F: 6	M: 8 F: 8	M: 30 F: 30
Idade	M: 32,2 F: 34,68	M: 19,39 F: 19,92	M: 0 F: 0	M: 16 F: 18	M: 32 F: 35	M: 47 F: 48	M: 96 F: 96
Quantidade de consultas	M: 3,10 F: 4,89	M: 2,95 F: 3,78	M: 0 F: 0	M: 1 F: 2	M: 2 F: 4	M: 4 F: 7	M: 28 F: 33
Tipo de acomodação	M: 0,44 F: 0,42	M: 0,50 F: 0,49	M: 0 F: 0	M: 0 F: 0	M: 0 F: 0	M: 1 F: 1	M: 1 F: 1

Note que para ambos os grupos, os valores foram extremamente próximos.

A Figura 4 apresenta os histogramas da variável resposta “quantidade de consultas” para homens e mulheres:

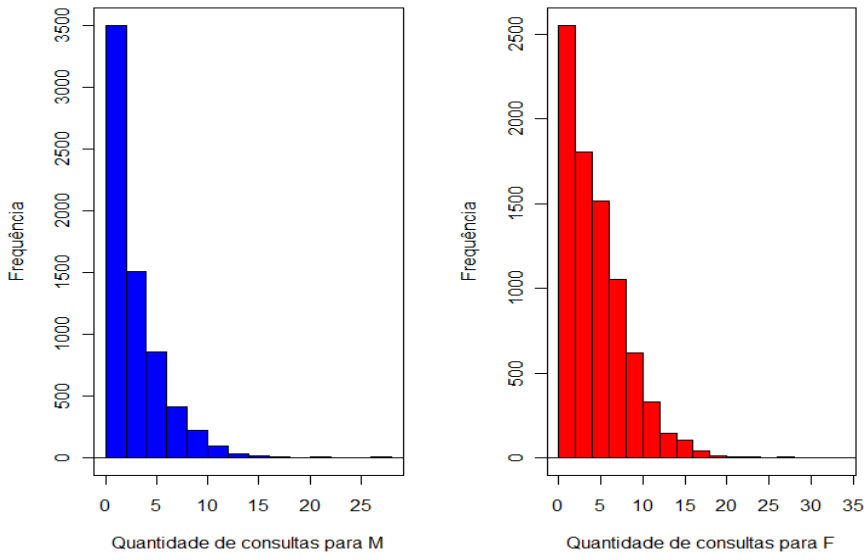

Figura 4. Histogramas de frequência de consultas para os dois grupos

Observe que o grupo “F” apresenta as quantidades distribuídas um pouco mais homogêneas do que o grupo “M”. A seguir, os coeficientes estimados do modelo de Poisson com ligação log e dados agrupados por sexo:

Tabela 8: Coeficientes estimados no modelo agrupado

VARIÁVEIS	ESTIMATIVA	ERRO PADRÃO	VALOR Z	P > (|Z|)
INTERCEPTO	M: 1,1709 F: 1,3735	M: 0,02025 F: 0,01498	M: 57,817 F: 91,681	M: <2e-16 *** F: <2e-16 ***
IDADE	M: 0,0041 F: 0,0080	M: 0,00035 F: 0,00024	M: 11,556 F: 32,14	M: <2e-16 *** F: <2e-16 ***
TEMPO NO PLANO	M: -0,02102 F: -0,0077	M: 0,0022 F: 0,0015	M: -9,296 F: -4,931	M: <2e-16 *** F: 8.2e-07 ***
TIPO DE ACOMODAÇÃO	M: -0,1415 F: -0,0918	M: 0,0146 F: 0,0106	M: -9,819 F: -8,607	M: <2e-16 *** F: <2e-16 ***

Grau de significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

A partir da Tabela 8, mais uma vez, todas as variáveis são significantes. A Tabela 9 apresenta os valores dos desvios residuais e AIC para os dois modelos:

Tabela 9: Valores de desvios residuais e AIC para os modelos agrupados por sexo e desagrupados

VALORES	MODELO DESAGRUPADO	MODELO AGRUPADO POR SEXO (“M”)	MODELO AGRUPADO POR SEXO (“F”)
DESVIOS RESIDUAIS	40.584 em 16.894 gl DR/gl = 2,40	17.414 em 6.658 gl DR/gl = 2,61	23.060 em 8.199 gl DR/gl = 2,81
AIC	82.061	33.917	48.039

Como pode ser visto acima na Tabela 9, a divisão entre os desvios residuais e os graus de liberdade foi crescente, no entanto os valores AIC dos modelos agrupados por sexo foram demasiadamente menores.

Abaixo a figura 5 com os gráficos dos modelos agrupados:

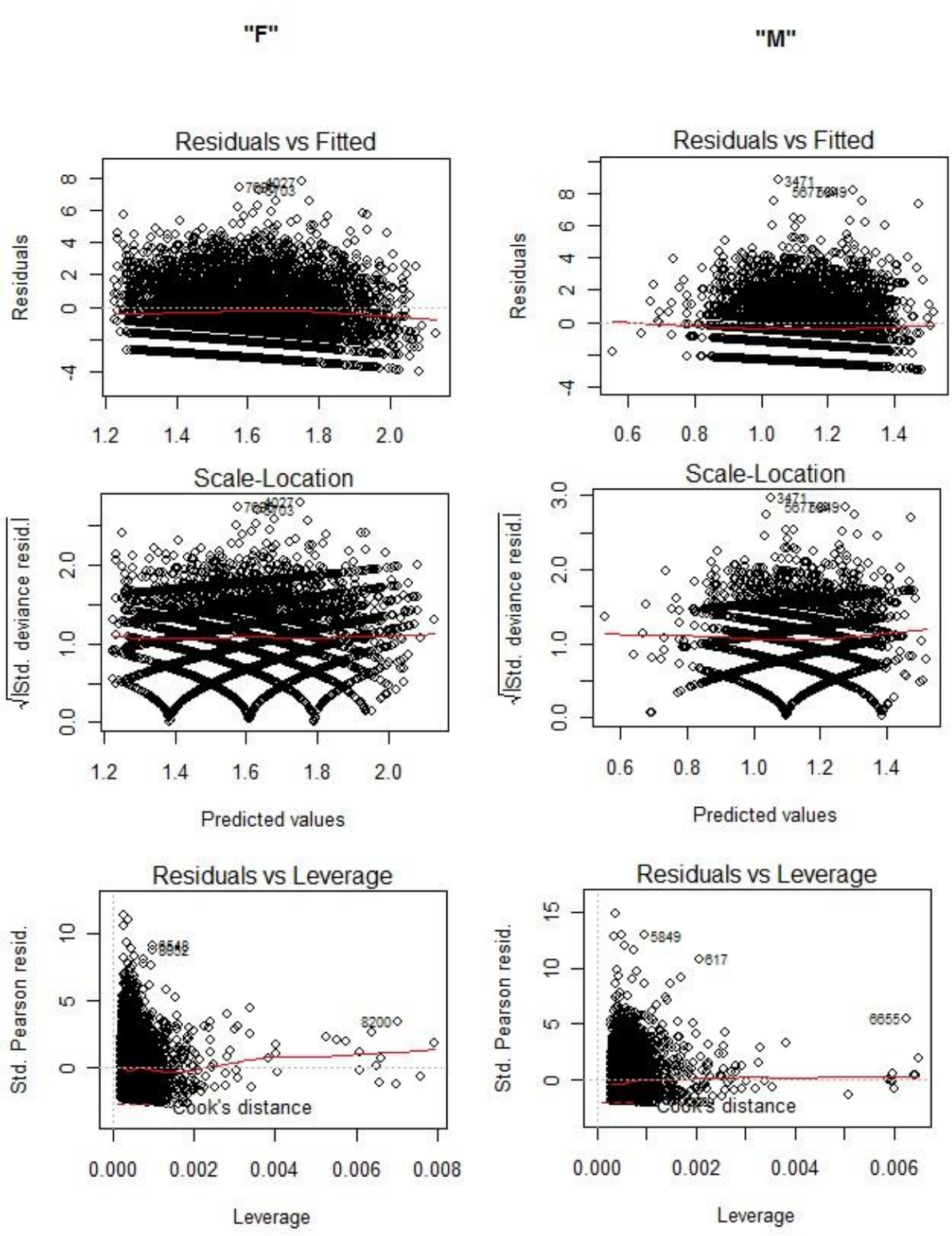

Figura 5: Gráficos do novo modelo

Como pode ser visto acima, os dados continuam agrupados e com diversos pontos extremos (outliers).

.

CONCLUSÃO

Este texto apresentou a metodologia do modelo linear generalizado (MLG). Duas aplicações foram feitas a um banco de dados de plano de saúde: uma utilizando um modelo adaptado à família de distribuição de Poisson com ligação log para os dados desagrupados e outra para a mesma família e ligação, porém com as observações agrupadas por sexo.

O primeiro modelo apresentou pior ajuste do que os seguintes, que também não foram tão adequados. Talvez este resultado seja devido à limitação dos dados com a ausência de algumas variáveis importantes (como por exemplo: valor das consultas e mensalidade paga pelo usuário do plano).

O que se sugere futuramente é que sejam usados dados mais completos, a fim de se obter modelagens mais seguras. Talvez também outros modelos possam ser mais condizentes com as observações que se pretende explicar.

REFERÊNCIAS BIBLIOGRÁFICAS

BRUYNE, P. **Dinâmica da pesquisa em ciências sociais**. Rio de Janeiro: Francisco Alves, 1991.

CORDEIRO, G. M.; PAULA, G. A. **Modelos de regressão para análise de dados univariados**. Rio de Janeiro: Instituto de Matemática Pura e Aplicada do CNPQ, 1989. 353p.

DEMÉTRIO, C. G. B; CORDEIRO, G. M. **Modelos Lineares Generalizados e Extensões**. São Paulo: ESALQ/USP, 2010. 249p. Disponível em: < http://www4.esalq.usp.br/departamentos/lce/arquivos/aulas/2010/LCE5868/livro.pdf> Acesso em: abril de 2013.

JONG, P.; HELLER, G. Z. **Generalized Linear Models for Insurance Data**. Cambridge: Cambridge University Press, 2008.

MACEDO, N. D. **Iniciação à pesquisa bibliográfica: guia do estudante para a fundamentação do trabalho de pesquisa**. 2. Ed. Revista. São Paulo: Edições Loyola, 1994.

MINAYO, M. C. D. S. **O desafio do conhecimento: pesquisa qualitativa em saúde**. São Paulo e Rio de Janeiro: Hucitec-Abrasco, 1993.

MORESI, E. **Metodologia da pesquisa**. Brasília: UCB, 2003.

PAULA, G. A. **Modelos de regressão com apoio computacional**. São Paulo: Instituto de Matemática e Estatística, Universidade de São Paulo, 2010. Disponível em: <http://people.ufpr.br/~lucambio/CE225/2S2011/texto_2010.pdf>. Acesso em: abril de 2013.

SILVA, E.L.; MENEZES, E.M. **Metodologia da pesquisa e elaboração de dissertação**. Florianópolis: UFSC/PPGEP/LED, 2000, 118p.

TURKMAN, M.A.A. e Silva, G.L. **Modelos Lineares Generalizados – da Teoria à Prática**. Edições SPE, Lisboa, 2010. Disponível em: < http://docentes.deio.fc.ul.pt/maturkman/mlg.pdf >. Acesso em: abril de 2013.