Universidade Federal do Rio de Janeiro
Centro de Ciências Matemáticas e da Natureza
Instituto de Matemática
Departamento de Métodos Estatísticos

Aplicação de Modelos Lineares Generalizados utilizando abordagem clássica para dados de contagem.

Pedro Paulo de Lima Boechat

Rafael Diniz da Rocha

PROJETO FINAL DE CURSO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO TÍTULO DE ATUÁRIO E ESTATÍSTICO.

COMISSÃO EXAMINADORA:

Mariane Branco Alves

Thais Cristina Oliveira da Fonseca

Flavia Maria Pinto Ferreira Landim

Rio de Janeiro

2015

Resumo

Este projeto tem como objetivo analisar o comportamento de utilização dos serviços da cobertura de Assistência 24 horas numa carteira de clientes da seguradora fictícia P&R Seguros. Diante disso, estudou-se a metodologia dos Modelos Lineares Generalizados, no contexto clássico, e suas propriedades a fim de modelar os dados e verificar a existência de possíveis associações entre a contagem de utilização dos serviços e demais características ligadas ao perfil dos clientes e de seus veículos. Ademais, para alcançar tal finalidade foi utilizado o software R-Project para testar a significância das variáveis e estruturar os 7 (sete) modelos candidatos à verificação proposta. A partir dos critérios: AIC, Erros Absoluto e Relativo e Erro Quadrático Médio (EQM), verificaram-se qual dentre todos os modelos melhor se adaptou aos dados dada uma razoável capacidade preditiva. A partir do conhecimento adquirido no estudo sobre o perfil de utilização dos seus segurados, a empresa planeja agravar o preço da cobertura para clientes que se enquadram nessas características e fazer as seguintes modificações no produto: manter a cobertura atual que não possui limite de utilização, porém com o preço mais elevado e criar uma nova cobertura, sendo esta mais econômica e limitada na quantidade de utilizações durante o tempo de vigência do contrato de seguro. Além dessa medida, a empresa irá analisar a viabilidade de demais propostas de alteração do pacote de serviços comercializados, no intuito de otimizar os custos do processo e melhor se direcionar a cada grupo de clientes.

Abstract

This project´s goal is to analyze the behavior of a 24-hour assistance service usage in a client portfolio of a fictitious insurance company named *P&R Seguros*. Therefore, we studied the Generalized Linear Models, in a classical context and their properties, in order to model the data and check for possible associations between use count of services and other characteristics related to the profile of the costumers and their vehicles. Moreover, to achieve such purpose we used R-Project software to test the significance of the variables and structure seven possible models for the proposed verification. Based on AIC criterion, absolute error, relative error and mean square error (MSE), it was verified, among all models, one that best adapted to the data, given a good predictive capacity. After acquiring knowledge on the study about the profile of the insured´s usage of the service, the company plans to aggravate the price of coverage for customers who fall into these characteristics and make the following changes in the product: maintain current coverage that has no usage limit, but with a higher price and create a new coverage, which is more economical and limited in the amount of uses during the period of time in which the insurance contract is valid. In addition to this measure, the company will examine the feasibility of other proposed amendments to the package of services sold, in order to optimize process costs and better target every customer group.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O mercado segurador brasileiro vem experimentando grande avanço, de forma que se observa um consistente aumento das receitas ao longo dos anos. De acordo com o material empresarial divulgado pela empresa de consultoria KPMG (2013), nos últimos anos o mercado de seguros tem apresentado taxas de crescimento em torno de 10% a 15% ao ano. Na década de 80, este segmento de negócio representava cerca de 1% a 2% do PIB brasileiro, em 2013, estava entre 3% a 4% do PIB.

O ramo de seguro para automóveis responde por grande parte desse crescimento. Este pode ser explicado principalmente pelo aumento das vendas de veículos novos e usados aquecido por incentivos fiscais, financeiros e econômicos nos últimos anos e consequentemente maior poder de compra da população. Ao adquirir um automóvel, grande parte dos seus consumidores teme por qualquer risco que possa afetar o seu patrimônio, tais como: vandalismo, roubo, furto, enchentes, entre outros, que podem gerar prejuízo econômico-financeiro. Por esse motivo, a busca pelo seguro é uma boa opção de transferência do risco às seguradoras, mediante ao pagamento do prêmio de seguro, tendo que se dispor unicamente de um capital bem inferior quando comparado ao valor do bem.

Adjunto a esse crescimento, as empresas seguradoras vislumbraram a oportunidade de diversificar seus produtos (antes com coberturas básicas), oferecendo serviços agregados ao seguro, no intuito de gerar aumento de receita e ampliar o atendimento para diversos clientes. Em geral, são comercializadas coberturas de assistência 24h que englobam os serviços de chaveiro, guincho, técnico, mecânico, entre outros que prestam atendimento aos segurados a qualquer hora e em qualquer lugar, utilizando sempre que precisar e sem ter que pagar nada a mais além do prêmio do seguro.

Dada a importância do seguro de automóveis para a sociedade e para esse ramo de mercado, este projeto terá como foco analisar a frequência de utilização do produto Assistência 24 horas, numa carteira de segurados da empresa P&R Seguros. Teremos como base para estudo, dados do comportamento de utilização dos serviços de cada indivíduo e demais informações ligadas ao perfil dos clientes fornecidas na data de contratação do seguro.

Mais especificamente, o objeto de estudo serão os dados de contagens de utilização dos serviços de assistência 24h na referida seguradora, buscando-se analisar possíveis associações entre tais contagens e variáveis que refletem características dos segurados e seus veículos, por meio da aplicação de modelos lineares generalizados.

1.1 - Objetivo Geral

O objetivo geral deste trabalho é estudar os fundamentos do Modelo Linear Generalizado, sua metodologia para estimação dos parâmetros, construção de intervalos de confiança e testes de hipótese, bem como os critérios para seleção dos modelos. A fim de aplicá-los em uma análise experimental na carteira de clientes da companhia, utilizando dados de contagem de utilização dos serviços de assistência 24h e estruturar um modelo linear generalizado com o objetivo de verificar se os dados do comportamento de utilização desses segurados podem ser explicados pelos dados do perfil de cada indivíduo, tais como: sexo, idade e estado civil do condutor e modelo, marca, tipo de combustível e tempo de fabricação do veículo. Ao final, analisar os resultados e utilizar tal modelo para prever a frequência de utilização do serviço, por usuário e verificar a partir dos critérios: AIC e Mínimos Quadrados, qual modelo melhor se adapta aos dados e se é possível, ou não, propor mudanças no preço e/ou características do produto oferecido.

1.2 - Objetivos Específicos

- Testar as variáveis da base de segurados a fim de verificar quais são as mais significativas para explicação do desfecho em questão.
- Estruturar e eleger modelos candidatos à representação dos dados;
- Selecionar aleatoriamente alguns segurados e fazer previsão de seus desfechos a partir dos modelos propostos;
- Utilizar os métodos de Mínimos Quadrados e AIC, para critério de comparação entre os resultados obtidos na previsão dos modelos e os dados reais;
- Apresentar o modelo de previsão para a utilização dos serviços comercializados pela empresa a partir dos resultados obtidos, além de propor alguma melhoria/mudança no produto oferecido.

1.3 - Organização dos Capítulos

O presente trabalho está estruturado de forma que o capítulo 2 abordará a fundamentação teórica sobre Modelos Lineares Generalizados juntamente com alguns conceitos básicos como notações e terminologias utilizadas. Em seguida serão discutidos os métodos de estimação dos parâmetros nessa classe, propriedades assintóticas dos estimadores de máxima verossimilhança, teste de hipóteses, critérios para seleção de modelos e apresentação de alguns exemplos aplicados com destaque para modelos de resposta Poisson devido aos dados de estudo serem provenientes de um processo de contagem.

No capítulo 3 será apresentada uma análise descritiva dos dados estudados, assim como todo o tratamento realizado na base para chegar ao formato adequado para utilização nos modelos e aplicação ao R, além disso, serão apresentados os resultados obtidos dos

modelos propostos. Por fim, no capítulo 4 serão descritas as análises e considerações finais do trabalho.

Capítulo 2 - Modelos Lineares Generalizados

2.1 – Introdução

Neste capítulo apresenta-se uma revisão sobre a classe de modelos lineares generalizados (MLG), abordando o aspecto teórico juntamente com alguns conceitos básicos como notação e terminologia utilizada, bem como alguns exemplos e propriedades relacionadas à família de distribuição exponencial. Em seguida trata-se de estimação dos parâmetros, as propriedades assintóticas dos estimadores de máxima verossimilhança, teste de hipóteses e alguns critérios de seleção de modelos. As seções apresentadas a seguir são fortemente baseadas em Dobson e Barnett (2008).

Segundo Turkman e Silva (2000), os modelos lineares generalizados foram formulados por Nelder e Wedderburn nos anos 70, desempenhando um importante papel no desenvolvimento da Estatística Aplicada. No princípio, o uso era restrito, devido à falta de bibliografia e à complexidade inicial do primeiro software dirigido ao uso da metodologia, o GLIM. O domínio público da teoria e prática dos Modelos Lineares Generalizados foi pleno somente 20 anos depois. A importância dos MLG, do ponto de vista teórico, advém, essencialmente, do fato da metodologia destes modelos constituir uma abordagem unificada de muitos procedimentos estatísticos correntemente usados nas aplicações e promover o papel central da verossimilhança na teoria da inferência. Todavia, a ideia básica proposta por Nelder e Wedderburn (1972) era diversificar as opções para a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições, bem como tornar mais maleável a relação entre a média da variável dependente no experimento e a parte sistemática não aleatória (o chamado preditor linear), através da chamada função de ligação.

Pode-se citar alguns casos particulares dos MLG: modelo de regressão linear clássico, modelo de regressão logística, modelo de regressão Poisson, modelos de análise de variância e covariância, modelos logit e probit para estudos de proporções, modelos log-lineares para dados de contagens, entre outros. Isto lhes permitiu desenvolver um algoritmo geral para a estimativa de máxima verossimilhança em todos estes modelos. Contudo, estes modelos apresentam uma estrutura de regressão linear e possuem em comum o fato da distribuição de probabilidade associada à variável resposta pertencer à família exponencial de distribuições, não se restringindo unicamente à distribuição normal como pressupõem os modelos de regressão linear normal.
Assim, os MLG's vêm sendo largamente utilizados, devido à abrangência no que se refere à quantidade de modelos que contemplam, facilidade nas análises associada ao rápido desenvolvimento computacional que se tem verificado nos últimos anos, propiciado pelos avanços tecnológicos com a criação, aprimoramento e otimização dos softwares

estatísticos e na literatura com o aperfeiçoamento das teorias de modelagem estatística. Ainda que apresente restrições, como manter a estrutura de linearidade, restringir as distribuições à família exponencial e por exigirem a independência das variáveis resposta, a classe de modelos lineares generalizados é extremamente ampla, dando tratamento adequado a dados em sua escala original, em diversas aplicações práticas relevantes.

2.2 - Notação e Terminologia

Os modelos lineares generalizados são uma extensão do modelo linear normal

$$\boldsymbol{Y} = X\beta + \varepsilon \quad (\mathbf{2.1})$$

onde **X** é uma matriz de dimensão n x p de especificação do modelo (em geral é uma matriz de covariáveis com um primeiro vetor unitário) associada a um específico vetor $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^{\boldsymbol{T}}$ de parâmetros e ε é um vetor de erros aleatórios com distribuição que se supõe $N_N(0, \sigma^2$), sendo a matriz identidade de dimensão NxN.

Estas hipóteses implicam obviamente que o valor esperado da variável resposta é uma função linear das covariáveis na medida em que, condicionando à X, a esperança é igual a $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

Considere uma única variável aleatória Y cuja distribuição de probabilidade depende de um único parâmetro θ. Diz-se dessa variável aleatória tem distribuição pertencente à família exponencial se a sua função densidade de probabilidade, ou função massa de probabilidade, puder ser escrita na forma:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)}$$

onde a, b, s e t são funções conhecidas. Pode-se perceber uma simetria entre y e θ. Tal simetria é enfatizada se a equação é reescrita:

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \quad (\mathbf{2.2})$$

onde $s(y) = exp\,[d(y)]$ e $t(\theta) = exp\,[c(\theta)]$.

Se $a(y) = y$ a distribuição está na forma canônica (ou seja, forma padrão) e $b(\theta)$ é às vezes chamado de parâmetro natural (também chamado parâmetro canônico) da distribuição.

Se há outros parâmetros, além do parâmetro de interesse principal θ, esses são chamados de "parâmetros de perturbação", formando partes das funções a, b, c, d, e são tratados como se fossem conhecidos, ou seja, constantes.

Várias distribuições conhecidas pertencem à família exponencial, por exemplo, as distribuições Poisson, Normal e Binomial podem ser escritas na forma canônica.

Sejam, então,$Y_1, \ldots, Y_N$ variáveis aleatórias de interesse independentes e com distribuição pertencente à família exponencial. Os Y_i ´s serão as variáveis com as seguintes propriedades:

1) A distribuição de cada Y_i tem a forma canônica e depende de um único parâmetro θ_i (os θ_i´s não têm que ser necessariamente o mesmo), então

$$f(y_i;\ \theta_i) = exp[y_i b_i(\theta_i) + c_i(\theta_i) + \ d_i(\theta_i)].$$

2) As distribuições de todos os $Y_i´s$ são da mesma forma (por exemplo, todas Normais ou todas Binomiais) de modo que os subscritos em b, c e d não são necessários.
   Assim, a função densidade de probabilidade conjunta de $Y_1, \ldots, Y_N$ é

$$f(y_1, y_2, \ldots y_N;\ \theta_1, \theta_2, \ldots \theta_N) = \prod_{i=1}^{N} \exp[y_i\, b(\theta_i) + c(\theta_i) + d(y_i)]$$

$$= \exp\left[\sum_{i=1}^{N} y_i\, b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right]$$

Os parâmetros θ_i´s não são tipicamente de interesse direto (já que pode haver um para cada observação).

Na definição de McCullagh e Nelder (1989), um Modelo Linear Generalizado baseia-se em três componentes fundamentais:

1) Componente aleatório

Dado o vetor de covariáveis $\boldsymbol{x_i}$, as variáveis Y_i são (condicionalmente) independentes com distribuição pertencente à família exponencial, com $E(Y_i|x_i) = \mu_i$,para $i = 1, \ldots, N$.

2) Componente estrutural ou sistemática

Consiste numa combinação linear de variáveis preditoras, tendo p variáveis preditoras e N observações, tal que:

$$\eta = X\beta\ = \ \beta_0 + \ \beta_1 x_1(i) + \ \beta_2 x_2(i) + \ \beta_3 x_3(i) + \ldots + \beta_p x_p(i)\ \text{, para } i = 1, \ldots, N.$$

3) Função de ligação

É a função que relaciona o componente aleatório ao componente sistemático, ou seja, o valor esperado μ_i ao preditor linear $\eta_i = z_i{}^T \beta$, como segue:

$$g(\mu_i) = \eta_i \text{ , para } i = 1, \ldots, N \text{ ,}$$

onde $\boldsymbol{g(\mu_i)}$ é uma função monótona e diferenciável.

Em geral $z_i = (1, x_{i1}, \ldots, x_{ik})^T$ com $k = p \quad 1$. Contudo, quando existem covariáveis qualitativas elas têm que ser, em geral, convenientemente codificadas à custa de variáveis binárias dummies, por exemplo, se uma variável qualitativa (ou fator) tem q categorias, então são necessárias $q \quad 1$ variáveis binárias para representar. Estas variáveis têm que ser incluídas no vetor z.

Para especificação de modelo estamos mais interessados nos parâmetros $\beta_1, \beta_2, \ldots \beta_p$ (onde $p < N$). Supomos que $E(Y_i) = \mu_i$, onde μ_i é alguma função de θ_i. Pode-se escrever então

$$g(\mu_i) = \boldsymbol{x_i}^T \beta,$$

com

$$\boldsymbol{x_i} = \begin{bmatrix} x_{i1} \\ \cdots \\ x_{ip} \end{bmatrix} \text{ de forma que } \boldsymbol{x_i}^T = \begin{bmatrix} x_{i1} \ \ldots \ x_{ip} \end{bmatrix}$$

e β é um vetor p x 1 de parâmetros $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \cdots \\ \beta_p \end{bmatrix}$. O vetor $\boldsymbol{x_i}$ é a i-ésima coluna da matriz $\boldsymbol{X}$. Então o modelo linear generalizado tem três componentes:

I) Variáveis resposta $Y_1, \ldots, Y_N$ que supostamente têm a mesma distribuição da família exponencial.

II) A gama de parâmetros β e variáveis explicativas

$$\boldsymbol{X} = \begin{matrix} x_1^T \\ . \\ . \\ . \\ x_N^T \end{matrix} = \begin{matrix} x_{11} \ldots x_{1p} \\ . \\ . \\ . \\ x_{N1} \ldots x_{Np} \end{matrix}$$

III) Uma função de ligação monótona g de modo que

$$g(\mu_i) = \boldsymbol{x_i}^T \beta$$

onde

$$\mu_i = E(Y_i)$$

O caso especial mais conhecido de modelo linear generalizado é o modelo

$$E(Y_i) = \mu_i = \boldsymbol{x_i}^T \beta \qquad Y_i \sim N(\mu_i; \sigma^2)$$

onde $Y_1, \ldots, Y_N$ são independentes. Aqui a função de ligação é a função Identidade, ou seja, $g(\mu_i) = \mu_i$. Este modelo é usualmente escrito na forma da equação **(2.1).**

Desta forma, o componente linear $\mu = \boldsymbol{X\beta}$ representa o 'sinal' e ε representa o **'**ruído'**,** variação aleatória ou 'erro'. Regressão Múltipla, Análise de Variância e Análise de Covariância são todas dessa forma.

Focalizando a distribuição observacional em um modelo linear generalizado, vê-se, nos exemplos a seguir, que diversas distribuições pertencem à família exponencial, evidenciando a abrangência dessa classe de modelos.

Exemplo 1: Distribuição Poisson

A função de probabilidade para a variável aleatória discreta Y é:

$$f(y;\theta) = \frac{\theta^y e^{-\theta}}{y!}$$

onde y assume os valores $0,1,2...$ Essa equação pode ser reescrita como

$$f(y;\theta) = \exp(ylog\theta \quad \theta \quad logy!)$$

que está na forma canônica porque $a(y) = y$. Também, o parâmetro natural é $log\theta$.

A distribuição Poisson, denotada por $Y \sim Poisson(Y)$, é bastante usada para modelar dados de contagem. Tipicamente estes são números de ocorrências de determinado evento em um período de tempo ou espaço definido, onde a probabilidade de um evento ocorrer em um pequeno espaço de tempo é muito baixa, e os eventos ocorrem independentemente.

São exemplos o número de condições médicas reportadas por uma pessoa; número de erros de soletração numa página de jornal; ou o número de componentes defeituosos em um computador ou num montante de itens manufaturados.

Se uma variável aleatória tem distribuição Poisson, seu valor esperado e variância são iguais. $E(Y) = Var(Y) = \theta$. Dados reais que podem ser razoavelmente modelados por uma distribuição Poisson normalmente têm uma variância maior. Tal fenômeno é conhecido como sobredispersão, e, em muitos casos, o modelo deverá ser adaptado, através de transformações, para representar melhor os dados.

Exemplo 2: Distribuição Normal

A função densidade de probabilidade é:

$$f(y;\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left[\quad \frac{1}{2\sigma^2} (y \quad \mu)^2 \right]$$

onde μ é o parâmetro de interesse e σ^2 é visto como um "parâmetro de perturbação", por isso tratado como constante. Tal equação pode ser reescrita da seguinte forma:

$$f(y;\mu) = \exp\left[\quad \frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} \quad \frac{\mu^2}{2\sigma^2} \quad \frac{1}{2}\log(2\pi\sigma^2) \right]$$

Esta nova equação está na forma canônica. O parâmetro canônico é $b(\mu) = \mu/\sigma^2$ e os outros termos são facilmente encontrados.

$$c(\mu) = \frac{\mu^2}{2\sigma^2} \quad \frac{1}{2}\log(2\pi\sigma^2) \qquad \text{e} \qquad d(y) = \frac{y^2}{2\sigma^2}$$

A distribuição Normal é usada para modelar dados contínuos que têm distribuição simétrica. É amplamente usada para três razões principais. Primeiramente, muitos fenômenos que ocorrem naturalmente são bem descritos pela distribuição Normal. Um exemplo disso é medição de altura ou pressão sanguínea de pessoas. Em segundo lugar, mesmo se os dados não forem normalmente distribuídos (se a distribuição é assimétrica) a média ou total de uma amostra aleatória simples será, aproximadamente, normalmente distribuída, caso o tamanho amostral seja suficientemente grande. Tal resultado é provado pelo Teorema Central do Limite. E em última instância, há uma bagagem considerável de teoria estatística desenvolvida concernindo a distribuição normal, incluindo distribuições de amostras derivadas dela e aproximações para outras distribuições. Por estas razões, se um dado contínuo y não é normalmente distribuído é frequentemente plausível tentar identificar uma transformação, como $y' = logy$ ou $y' = \sqrt{y}$, que produz dado y'que é aproximadamente normal.

Exemplo 3: Distribuição Binomial

Consideremos uma série de eventos binários, chamados tentativas, cada um com somente duas possibilidades de resultado: sucesso ou fracasso. Seja a variável aleatória Y o número de 'sucessos' em N tentativas independentes cuja probabilidade de sucesso, π, é a mesma em todas as tentativas. Então Y tem distribuição Binomial com função densidade de probabilidade

$$f(y;\pi) = \binom{N}{y} \pi^y \ (1 \quad \pi)^{N-y},$$

onde y assume os valores $0,1,2,..N$. Isto é denotado por $Y \sim Binomial(N,\pi)$. Aqui π é o parâmetro de interesse e N é conhecido. A função de probabilidade pode ser reescrita como:

$$f(y;\mu) = \exp(\ ylog\pi \quad ylog(1 \quad \pi) + Nlog(1 \quad \pi) + \log\binom{N}{y}$$

que é da forma da equação **2.2** com $b(\pi) = log\pi - log(1 \quad \pi) = log\ [\pi\ /\ (1 \quad \pi)]$.

A distribuição Binomial é normalmente a primeira escolha de modelo para observações de processos com resultados binários. São exemplos disso: número de candidatos que passaram em um exame (os resultados possíveis sendo passar ou não passar), ou número de pacientes com alguma doença, ainda vivos em um determinado tempo desde a diagnose (os resultados possíveis são estar vivo ou morto). No caso particular em que

N=1, tem-se o modelo Bernoulli (π), amplamente utilizado para classificações binárias, como por exemplo, risco de crédito.

2.3 - Propriedades das distribuições da Família Exponencial

As expressões para o valor esperado e variância de $a(Y)$ podem ser obtidas utilizando-se os seguintes resultados que valem nas operações para qualquer função densidade de probabilidade cuja ordem de integração e diferenciação possa ser trocada, da definição de função densidade de probabilidade, a área abaixo da curva é a unidade, de forma que

$$\int f(y;\theta)dy = 1\ (\mathbf{2.3}).$$

Onde a integração é sobre todos os valores de y. Quando a variável aleatória Y é discreta a integração é substituída por um somatório.

Ao diferenciar os dois lados da equação **(2.3)** com respeito a θ obtém-se

$$\frac{d}{d\theta}\int f(y;\theta)\,dy = \frac{d}{d\theta}.1 = 0.$$

Se a ordem de integração e diferenciação no primeiro termo for invertida, então a equação acima se torna

$$\int \frac{df(y;\theta)}{d\theta}\,dy = 0\ (\mathbf{2.4}).$$

Similarmente se **(2.3)** é diferenciada duas vezes com respeito a θ e a ordem de integração e diferenciação é trocada obtemos

$$\int \frac{d^2f(y;\theta)}{d\theta^2}\,dy = 0\ (\mathbf{2.5}).$$

Estes resultados podem agora ser usados para distribuições na família exponencial. Da equação **(2.2)**

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)].$$

Então

$$\frac{df(y;\theta)}{d\theta} = [a(y)b'(\theta) + c'(\theta)]\,f(y;\theta)$$

Pela equação **(2.4)**

$$\int [a(y)b'(\theta) + c'(\theta)]\,f(y;\theta)dy = 0.$$

Esta equação pode ser simplificada para

$$b'(\theta)E[a(y)] + c'(\theta) = 0 \ (\mathbf{2.6})$$

porque $\int a(y)f(y;\theta)dy = E[a(y)]$ pela definição do valor esperado e $\int c'(\theta)f(y;\theta)dy = c'(\theta)$ por **(2.3).** Rearranjando a equação **(2.6)** temos:

$$E[a(Y)] = \frac{c'(\theta)}{b'(\theta)} \ (\mathbf{2.7}).$$

Um argumento similar pode ser usado para obter $var[a(Y)]$.

$$\frac{d^2 f(y;\theta)}{d\theta^2} = [a(y)b''(\theta) + c''(\theta)]\, f(y;\theta) + [a(y)b'(\theta) + c'(\theta)]^2\, f(y;\theta)$$

O segundo termo do lado direito da equação acima pode ser reescrito assim

$$[b'(\theta)]^2 \{a(y) \quad E[a(Y)]\}^2\, f(y;\theta)$$

usando a equação **(2.7).** Então pela equação **(2.5)**

$$\int \frac{d^2 f(y;\theta)}{d\theta^2} = b''(\theta)E[a(Y)] + c''(\theta) + [b'(\theta)]^2\, var[a(Y)] = 0 \ (\mathbf{2.8})$$

porque $\int \{a(y) \quad E[a(Y)]\}^2 \ f(y;\theta)dy = var[a(Y)]$ por definição.

Rearranjando a equação **(2.8)** e substituindo a equação **(2.7)** temos

$$var[a(Y)] = \frac{b''(\theta)c'(\theta) \quad c''(\theta)b'(\theta)}{[b'(\theta)]^3} \ (\mathbf{2.9}).$$

As equações **(2.7)** e **(2.9)** podem ser prontamente verificadas para distribuições Poisson, Normal e Binomial e usadas para obter valor esperado e variância de outras distribuições na família exponencial.

Expressões para o valor esperado e variância das derivadas da função log-verossimilhança serão necessárias. De **(2.2)**, a função log-verossimilhança para a distribuição na família exponencial é

$$l(\theta; y) = a(y)b(\theta) + c(\theta) + d(y).$$

A derivada de $l\ (\theta;\ y)$ com respeito a θ é

$$U(\theta; y) = \frac{dl(\theta; y)}{d\theta} = a(y)b'(\theta) + c'(\theta).$$

A função U é chamada de **estatística escore** e, como depende de y, pode ser vista como uma variável aleatória da seguinte forma

$$U = a(Y)b'(\theta) + c'(\theta) \ (\mathbf{2.10}).$$

O seu valor esperado é

$$E(U) = b'(\theta)E[a(Y)] + c'(\theta).$$

Usando equação **(2.7)** temos que

$$E(U) = b'(\theta)\left[\ \frac{c'(\theta)}{b'(\theta)}\right] + c'(\theta) = 0\ (\mathbf{2.11}).$$

A variância de U é chamada **Informação** e será denotada por . Usando a fórmula da variância de transformação linear de variáveis aleatórias temos que

$$= var(U) = [b'(\theta)]^2 var[a(Y)].$$

Substituindo **(2.9)** chega-se, finalmente, na equação da variância:

$$var(U) = \frac{b''(\theta)c'(\theta)}{b'(\theta)} \quad c''(\theta)\ (\mathbf{2.12}).$$

A estatística escore U é usada para inferência sobre os valores dos parâmetros em modelos lineares generalizados.

Uma outra propriedade de U, que usaremos mais adiante, é que

$$var(U) = E(U^2) = \quad E(U').$$

A primeira igualdade segue do resultado geral que

$$var(X) = E(X^2) \quad [E(X)]^2$$

para cada variável aleatória, e o fato de que $E(U) = 0$, visto em **(2.11).** Para se chegar na segunda igualdade, deriva-se U com respeito a θ, a partir de **(2.10)**

$$U' = \frac{dU}{d\theta} = a(Y)b''(\theta) + c''(\theta).$$

Então o valor esperado para U'é calculado da seguinte forma:

$$E(U') = b''(\theta)E[a(Y)] + c''(\theta)$$

$$= b''(\theta)\left[\ \frac{c'(\theta)}{b'(\theta)}\right] + c''(\theta)$$

$$= \quad var(U) = \quad .$$

2.4 – Inferência sobre Modelos Lineares Generalizados

2.4.1 – Estimação via Máxima Verossimilhança

Neste tópico considera-se um problema geral de estimação dos parâmetros β de um Modelo Linear Generalizado utilizando o método de máxima verossimilhança.

Para cada Y_i, a função log-verossimilhança é

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i) \ (\mathbf{2.13})$$

sendo que

$$E(Y_i) = \mu_i = \frac{c'(\theta_i)}{b'(\theta_i)} \ (\mathbf{2.14})$$

$$Var(Y_i) = \frac{b''(\theta_i)c'(\theta_i) \quad c''(\theta_i)b'(\theta_i)}{[b'(\theta_i)]^3} \ (\mathbf{2.15}) \ \text{ e}$$

$$g(\mu_i) = x_i{}^T\beta = \eta_i \ (\mathbf{2.16})$$

onde X_i é o vetor com elementos $x_{ij}, j = 1, \ldots, p$.

A função log-verossimilhança para todos os Y_i's é

$$l = \sum_{i=1}^{N} l_i = \sum y_i \, b(\theta_i) + \sum c(\theta_i) + \sum d(y_i)$$

Calculando as derivadas parciais de 1ª ordem da função log-verossimilhança tem-se que,

$$\frac{\delta l}{\delta \beta_j} = U_j = \sum_{i=1}^{N} \left[\frac{\delta l_i}{\delta \beta_j}\right] = \sum_{i=1}^{N} \left[\frac{\delta l_i}{\delta \theta_i} . \frac{\delta \theta_i}{\delta \mu_i} . \frac{\delta \mu_i}{\delta \beta_j}\right] \ (\mathbf{2.17}).$$

Considerando cada termo do lado direito de $(\mathbf{2.17})$ separadamente,

$$\frac{\delta l_i}{\delta \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i \quad \mu_i),$$

diferenciando $(\mathbf{2.13})$ e substituindo $(\mathbf{2.14})$. Teremos ainda que,

$$\frac{\delta \theta_i}{\delta \mu_i} = \frac{1}{\left(\frac{\delta \mu_i}{\delta \theta_i}\right)}.$$

Derivando $(\mathbf{2.14})$, tem-se:

$$\frac{\delta \mu_i}{\delta \theta_i} = \frac{c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i) b''(\theta_i)}{[b'(\theta_i)]^2} = b'(\theta_i) var(Y_i)$$

de $(\mathbf{2.15})$. Assim,

$$\frac{\delta \mu_i}{\delta \beta_j} = \frac{\delta \mu_i}{\delta \eta_i} . \frac{\delta \eta_i}{\beta_j} = \frac{\delta \mu_i}{\delta \eta_i} x_{ij}$$

Por fim, substituindo os termos desenvolvidos na expressão de $(\mathbf{2.17})$, tem-se que a função escore é expressa por

$$U_j = \sum_{i=1}^{N} \left[\frac{(y_i \quad \mu_i)}{var(Y_i)} x_{ij} \frac{\delta \mu_i}{\delta \eta_i} \right] \quad (\mathbf{2.18}).$$

Logo, o estimador de máxima verossimilhança $\hat{\beta}$ é a solução da equação $U_j = 0$, para j Є {1,...,p}. Em geral, os resultados obtidos dessa igualdade não são lineares e, dessa maneira, deve-se obter uma solução numérica utilizando a aproximação pelo método de Newton-Rapson ou o escore de Fisher.

Seja f uma função real e assuma que se tenha interesse em obter sua(s) raiz(es). O princípio do método iterativo de Newton-Rapson é encontrar o valor de x no qual f $(x) = 0$ baseado na aproximação de Taylor através da função f (x) nas vizinhanças do ponto t, ou seja,

$$f(x) = f(t) + (x \quad t) \left[\frac{df}{dx} \right]_{x=t},$$

obtendo-se

$$x = t \quad \frac{f(t)}{f'(t)}$$

Procede-se iterativamente e na iteração *m* tem-se:

$$x^{(m)} = x^{(m-1)} \quad \frac{f\left(x^{(m-1)}\right)}{f'(x^{(m-1)})}$$

onde $x^{(m)}$ representa o valor de x na iteração *m*, $x^{(m-1)}$ o valor de x na iteração (*m* - 1), $f\left(x^{(m-1)}\right)$ a função $f(x)$ avaliada em $x^{(m-1)}$ e $f'\left(x^{(m-1)}\right)$ a derivada da função $f(x)$ avaliada em $x^{(m-1)}$.

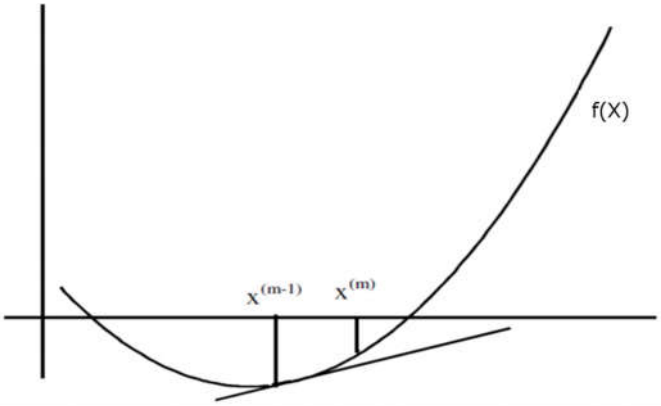

Figura 2.1 - Método de Newton-Raphson para encontrar a solução da equação f (x) =0.

A *Figura 2.1* ilustra o princípio do algoritmo de Newton-Raphson. O objetivo é encontrar o valor de x para o qual a função f cruza o eixo-x, ou seja, onde $f\ (x)\ =\ 0$, a inclinação da t a um valor $x^{(m-1)}$ é dada pela

$$\left[\frac{df}{dx}\right]_{x=x^{(m-1)}} = f'\left(x^{(m-1)}\right) = \frac{f(x^{(m)} \quad f(x^{(m-1)}))}{x^{(m)} \quad x^{(m-1)}} \quad (\mathbf{2.19})$$

onde a distância $x^{(m)} \quad x^{(m-1)}$ é pequena. Se $x^{(m)}$ é a solução necessária de modo que $f(x^m) = 0$, então $(\mathbf{2.19})$ pode ser reescrito como

$$x^{(m)} = x^{(m-1)} \quad \frac{f\left(x^{(m-1)}\right)}{f'(x^{(m-1)})} \quad (\mathbf{2.20})$$

A partir das suposições iniciais $x^{(1)}$, aproximações sucessivas são obtidas usando $(\mathbf{2.20})$ até que a processo iterativo convirja.

Voltando à questão de maximização da função de log-verossimilhança, como visto, deseja-se obter zeros da função escore.

A partir da expressão de $(\mathbf{2.18})$ tem-se que a matriz de variância-covariância da U_j's tem os termos

$$_{jk} = E\left[U_j U_k\right]$$

que formam a matriz de informação , ou seja,

$$_{jk} = E\left\{\sum_{i=1}^{N}\left[\frac{(Y_i \quad \mu_i)}{var(Y_i)}\ x_{ij}\left(\frac{\delta\mu_i}{\delta\eta_i}\right)\right]\sum_{l=1}^{N}\left[\frac{(Y_l \quad \mu_l)}{var(Y_l)}\ x_{lk}\left(\frac{\delta\mu_l}{\delta\eta_l}\right)\right]\right\}$$
$$= \sum_{i=1}^{N}\frac{E[(Y_i \quad \mu_i)^2]x_{ij}x_{ik}}{[var(Y_i)]^2}\left(\frac{\delta\mu_i}{\delta\eta_i}\right)^2 \quad (\mathbf{2.21}).$$

Sabe-se que $E[(Y_i \quad \mu_i)(Y_l \quad \mu_l)] = 0$, para $i \neq l$, já que os Y_i's são independentes. Usando $E[(Y_i \quad \mu_i)^2] = var(Y_i)$, $(\mathbf{2.21})$ pode ser simplificado para

$$_{jk} = \sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{var(Y_i)} \left(\frac{\delta\mu_i}{\delta\eta_i}\right)^2 (\mathbf{2.22}).$$

A equação $(\mathbf{2.10})$ de estimativa para o método de scoring é então generalizada para

$$b^{(m)} = b^{(m-1)} + \left[\quad ^{(m-1)}\right]^{-1} U^{(m-1)} (\mathbf{2.23})$$

onde $b^{(m)}$ é o vetor das estimativas dos parâmetros $\beta_1, \ldots, \beta_p$ na m-ésima iteração. Na equação $(\mathbf{2.23})$, $\left[\quad ^{(m-1)}\right]^{-1}$ é o inverso da matriz de informação com os elementos $_{jk}$ dada por $(\mathbf{2.22})$ e $U^{(m-1)}$ é o vetor de elementos dada por **(2.18)**, todos avaliados em $b^{(m-1)}$. Se ambos os lados da equação $(\mathbf{2.23})$ forem multiplicados por $^{(m-1)}$ obtemos

$$^{(m-1)}b^{(m)} = \quad ^{(m-1)}b^{(m-1)} + U^{(m-1)} (\mathbf{2.24}).$$

A partir de $(\mathbf{2.22})$, pode ser escrito como $= \mathbf{X^T W X}$

onde $\boldsymbol{W}$ é a matriz diagonal $N \; x \; N$ com elementos

$$w_{ii} = \frac{1}{var(Y_i)} \left(\frac{\delta\mu_i}{\delta\eta_i}\right)^2 (\mathbf{2.25}).$$

A expressão ao lado direito de $(\mathbf{2.24})$ é o vetor com elementos

$$\sum_{k=1}^{p}\sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{var(Y_i)} \left(\frac{\delta\mu_i}{\delta\eta_i}\right)^2 b_k{}^{(m-1)} + \sum_{i=1}^{N} \frac{(y_i \quad \mu_i)x_{ij}}{var(Y_i)} \left(\frac{\delta\mu_i}{\delta\eta_i}\right)$$

avaliada em $b^{(m-1)}$, isso decorre a partir das equações $(\mathbf{2.22})$ e $(\mathbf{2.18})$. Assim, a expressão do lado direito da equação $(\mathbf{2.24})$ pode ser escrita como

$$\boldsymbol{X^T W z}$$

onde $\boldsymbol{z}$ tem elementos

$$z_i = \sum_{k=1}^{p} x_{ik} b_k{}^{(m-1)} + (y_i \quad \mu_i)\left(\frac{\delta\eta_i}{\delta\mu_i}\right) (\mathbf{2.26}).$$

Com μ_i e $^{\delta\eta_i}/_{\delta\mu_i}$ avaliados em $\boldsymbol{b}^{(m-1)}$. Portanto, a equação iterativa $(\mathbf{2.23})$ pode ser escrita como

$$\boldsymbol{X^T W X b^{(m)} = X^T W z} \ \boldsymbol{(2.27).}$$

A maioria dos pacotes estatísticos que incluem procedimentos para ajuste de modelos lineares generalizados tem um algoritmo eficiente com base em $\boldsymbol{(2.27)}$. Eles começam usando algumas aproximações iniciais $\boldsymbol{b^{(0)}}$ para avaliar a $\boldsymbol{z}$ e $\boldsymbol{W}$, em seguida, $\boldsymbol{(2.27)}$ é resolvido para fornecer $\boldsymbol{b^{(1)}}$, que por sua vez é utilizada para obter melhores aproximações para $\boldsymbol{z}$ e $\boldsymbol{W}$, e assim por diante até a convergência adequada é obtida. Quando a diferença entre as sucessivas aproximações $\boldsymbol{b^{(m-1)}}$ e $\boldsymbol{b^{(m)}}$ é suficientemente pequena, $\boldsymbol{b^{(m)}}$ é tomada como a estimativa de máxima verosimilhança.

Para ilustrar esse princípio, começaremos com um exemplo numérico.

Exemplo 1: Tempos de falha para vasos de pressão

Os dados da *Tabela 2.1* são os tempos de vida (tempo até a falha em horas) de vasos de pressão fio Kevlar epóxi em 70% o nível de estresse. Eles são apresentados na Tabela 29.1 do livro de conjuntos de dados por Andrews e Herzberg (1985). A Figura *2.2* mostra a forma da sua distribuição.

Um modelo comumente usado para tempos até a falha (ou o tempo de sobrevivência) é a distribuição de Weibull, que tem a função de densidade de probabilidade

$$f(y;\lambda,\theta) = \frac{\lambda y^{\lambda-1}}{\theta^{\lambda}} \exp\left[\ \left(\frac{y}{\theta}\right)^{\lambda} \right] \ \boldsymbol{(2.28)},$$

onde $y > 0$ é o tempo até à falha, λ é um parâmetro que determina a forma da distribuição e θ é o parâmetro que determina a escala. A *Figura 2.2* mostra o gráfico de probabilidade dos dados na *Tabela 2.1* em comparação com a distribuição Weibull com $\lambda = 2$. Embora existam discrepâncias entre a distribuição e os dados para alguns dos tempos mais curtos, para a maior parte das observações a distribuição proporciona um bom modelo para os dados.

Tabela 2.1 – Tempos de vida dos vasos de pressão

1051	4921	7886	10861	13520
1337	5445	8108	11026	13670
1389	5620	8546	11214	14110
1921	5817	8666	11362	14496
1942	5905	8831	11604	15395
2322	5956	9106	11608	16179
3629	6068	9711	11745	17092
4006	6121	9806	11762	17568
4012	6473	10205	11895	17568
4063	7501	10396	12044	

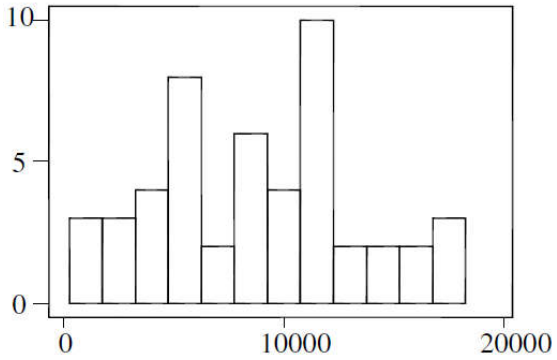

Figura 2.2 – Distribuição dos tempos de vida dos vasos de pressão.

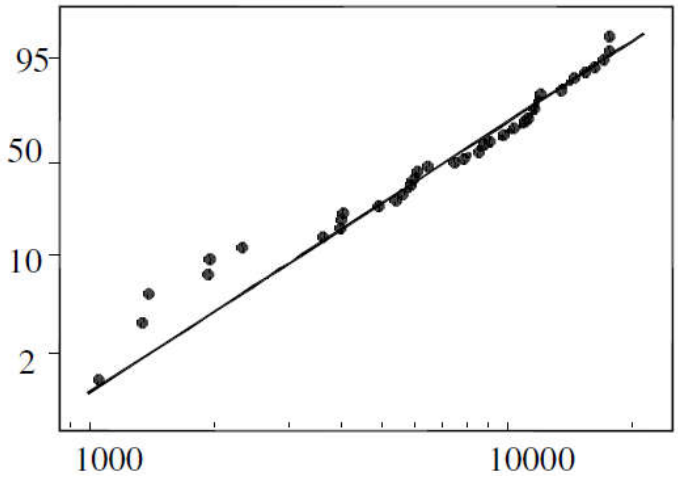

Figura 2.3 – Grafico de probabilidades dos dados de tempos de vida dos vasos de pressão em relação à distribuição Weibull com parâmetro de forma 2.

Por isso, pode-se utilizar uma distribuição Weibull com $\lambda = 2$ e uma estimativa de θ. A distribuição de $(\mathbf{2.27})$ pode ser escrita como

$$f(y;\theta) = \exp\left[log\lambda + (\lambda \quad 1)logy \quad \lambda\, log\theta \quad \left(\frac{y}{\theta}\right)^{\lambda}\right]$$

Esta pertence à família exponencial com

$$a(y) = y^{\lambda}, b(\theta) = \quad \theta^{-\lambda}, c(\theta) = logy \quad \lambda log\theta \; e \; d(y) = (\lambda \quad 1)logy,$$

onde λ é um parâmetro de deformidade. Esta não é uma forma canônica (a menos que $\lambda = 1$, o que corresponde à distribuição exponencial) e por isso não pode ser usada

diretamente na especificação de um modelo linear generalizado. No entanto, é apropriado para ilustrar a estimativa de parâmetros para as distribuições na família exponencial.

Denote por $Y_1, \ldots, Y_N$ os dados, com $N = 49$. Se os dados são da forma de uma amostra aleatória de vasos de pressão, assume-se que os $Y_i's$ são variáveis aleatórias independentes. Se todos eles têm a distribuição Weibull com os mesmos parâmetros, a distribuição de probabilidade conjunta é

$$f(y_1, \ldots, y_N; \theta; \lambda) = \prod_{i=1}^{N} \frac{\lambda {y_i}^{\lambda-1}}{\theta^{\lambda}} \exp\left[\left(\frac{y_i}{\theta}\right)^{\lambda} \right]$$

A função log-verossimilhança é:

$$f(\theta; y_1, \ldots, y_N, \lambda) = \sum_{i=1}^{N} \left[[(\lambda \quad 1) \log y_i + \log y \quad \lambda \, log\theta] \quad \left(\frac{y_i}{\theta}\right) \lambda \right]. \ (\mathbf{2.29}).$$

Para maximizar esta função, é exigida a derivada em relação a θ. Esta é a função escore

$$\frac{dl}{d\theta} = U = \sum_{i=1}^{N} \left[\quad \frac{\lambda}{\theta} + \frac{\lambda {y_i}^{\lambda}}{\theta^{\lambda+1}} \right] (\mathbf{2.30}).$$

O estimador de máxima verossimilhança $\hat{\theta}$ é a solução da equação $U(\theta) = 0$. Neste caso, é fácil encontrar uma expressão explícita para $\hat{\theta}$ se λ é uma constante conhecida, mas para efeitos ilustrativos, será obtida uma solução numérica utilizando a aproximação de Newton-Raphson.

Para a estimativa de máxima verossimilhança usando a função escore, tem-se:

$$\theta^{(m)} = \theta^{(m-1)} \quad \frac{U^{(m-1)}}{U'^{(m-1)}} \ (\mathbf{2.31}).$$

A partir de $(\mathbf{2.30})$, para a distribuição de Weibull com $\lambda = 2$,

$$U = \quad \frac{2\,N}{\theta} + \frac{2 \sum {y_i}^2}{\theta^3} \ (\mathbf{2.32})$$

que é avaliada em estimativas sucessivas $\theta^{(m)}$. A derivada de U, obtida diferenciando $(\mathbf{2.30})$, é

$$\frac{dU}{d\theta} = U' = \sum_{i=1}^{N} \left[\frac{\lambda}{\theta^2} \quad \frac{\lambda(\lambda+1){y_i}^{\lambda}}{\theta^{\lambda+2}} \right]$$

$$= \frac{2\,N}{\theta^2} \quad \frac{2\sum {y_i}^2}{\theta^4} \quad (\mathbf{2.33})$$

Para a estimativa de máxima verossimilhança, é comum para aproximar U' pelo valor esperado $E\,(U')$. Para distribuições na família exponencial, este é facilmente obtido utilizando $E(U') = b''(\theta)E[a(Y)] + c''(\theta)$.

A informação que é

$$= E(\ \ U') = E\left[\sum_{i-1}^{N} U_i{}'\right] = \sum_{i=1}^{N}[E(\ \ U_i')]$$

$$= \sum_{i=1}^{N}\left[\frac{b''(\theta)c'(\theta)}{b'(\theta)} \quad c''(\theta)\right] = \frac{\lambda^2 N}{\theta^2} \quad (\mathbf{2.34})$$

onde U_i é a função escore para Y_i e as expressões para b e c são dadas em $(\mathbf{2.28})$. Assim, uma equação de estimação alternativa é:

$$\theta^{(m)} = \theta^{(m-1)} + \frac{U^{(m-1)}}{(m-1)} \ (\mathbf{2.35}).$$

Tabela 2.2 - Detalhes de iterações de Newton-Raphson para obter uma estimativa de máxima verosimilhança para o parâmetro de escala para a distribuição de Weibull para modelar os dados da Tabela 2.1.

Iteration	1	2	3	4
θ	8805.9	9633.9	9876.4	9892.1
$U \times 10^6$	2915.10	552.80	31.78	0.21
$U' \times 10^6$	-3.52	-2.28	-2.02	-2.00
$\mathrm{E}(U') \times 10^6$	-2.53	-2.11	-2.01	-2.00
U/U'	-827.98	-242.46	-15.73	-0.105
$U/\mathrm{E}(U')$	-1152.21	-261.99	-15.81	-0.105

O procedimento acima é denominado método de "scoring".

A *Tabela 2.2* mostra os resultados da utilização da equação $(\mathbf{2.31})$, tomando iterativamente a média dos dados da Tabela 1, $\bar{y}$ = 8805,9, como o valor inicial $\theta^{(1)}$, e estas aproximações posteriores são mostradas na linha superior da *Tabela 2.2*. Os números na segunda fila foram obtidos avaliando $(\mathbf{2.32})$ em $\theta^{(m)}$ e os valores dos dados, eles se aproximam de zero rapidamente. A terceira e quarta linhas, U' e $E\,(U') =$ têm valores semelhantes ilustrando que qualquer um pode ser usado, isto é ainda mostrado pela similaridade dos números na quinta e sexta linhas. A estimativa final é $\theta^{(5)}$= 9.892,1 - (- 0,105) = 9.892,2. Esta é a estimativa da máxima verosimilhança $\hat{\theta}$ para estes dados. Neste valor a função de log-verossimilhança, calculada a partir de $(\mathbf{2.29})$, é l = -480,850.

Exemplo 2: Para o caso específico de interesse neste trabalho, em que teremos um modelo Log-Linear, com N observações independentes provenientes de uma variável aleatória com distribuição de Poisson.

A função de verossimilhança para as N observações é dada por:

$$L(y;\lambda) = e^{-\lambda_i}\frac{{\lambda_i}^{y_i}}{y_i!}.$$

Além disso, a log-verossimilhança é dada por:

$$(y;\lambda) = \sum_{i=1}^{N}(\quad \lambda_i + y_i ln\lambda_i \quad lny_i!).$$

A função de ligação é dada por $g(\lambda) = \ln(\lambda) = \boldsymbol{X}^T \beta$. A expressão para a log-verossimilhança em função dos β's:

$$(\boldsymbol{\beta}) = \sum_{i=1}^{N}[e^{-{\boldsymbol{X_i}}^T\boldsymbol{\beta}} + Y_i{\boldsymbol{X}_i}^T\boldsymbol{\beta} \quad ln(y_i!)]$$

desconsiderando a última parcela da expressão acima pois é constante nos parâmetros β_j, logo dispensável na identificação dos máximos:

$$(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left[e^{\sum_{k=0}^{p} x_{k(i)}\boldsymbol{\beta}_k} + y_i\sum_{i=1}^{N} X_{k(i)}\beta_k\right].$$

A condição necessária para a existência de extremo da log-verossimilhança no ponto $\beta = \hat{\beta}$ é:

$$\frac{\delta\ (\widehat{\boldsymbol{\beta}})}{\delta(\boldsymbol{\beta_j})} = \sum_{i=1}^{N}\left[x_j(i)\left[y_i \quad e^{\sum_{k=0}^{p} x_{k(i)}\boldsymbol{\beta}_k}\right]\right] = 0, \qquad para\ qualquer\ j = 0, \ldots p.$$

É notório que é um sistema não linear cujas soluções são não analíticas.
Métodos númericos computacionais entrariam como alternativa para solucionar o sistema e dar os possíveis candidatos a ponto de máximo, os estimadores de máxima verossimilhança.

2.4.2 - Distribuição amostral para estatística escore

Suponha que $Y_1, \ldots Y_N$ são variáveis aleatórias independentes em um modelo linear generalizado com parâmetros β onde $E(Y_i) = \mu$ e $g(\mu_i) = \boldsymbol{x_i}^T\beta = \eta_i$

As estatísticas escore são

$$U_j = \frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{N} \left[\frac{(Y_i \quad \mu_i}{var(Y_i)} \, x_{ij} \, \left(\frac{\delta \mu_i}{\delta \eta_i} \right) \right] \text{ para } j = 1, \ldots p$$

Como $E(Y_i) = \mu$ para todo i

$$E(U_J) = 0 \text{ para } j = 1, \ldots p$$

A matriz de covariância da estatística escore é matriz de informação com elementos

$$_{jk} = E[U_j U_k]$$

Se há somente um parâmetro β, a estatística escore tem distribuição amostral assintótica

$$\frac{U}{\sqrt{}} \sim N(0,1), \text{ou equivalentemente } \frac{U^2}{} \sim \chi^2(1)$$

porque $E(U) = 0$ e $Var(U) =$

Se há um vetor de parâmetros $\begin{bmatrix} \beta_1 \\ \ldots \\ \beta_p \end{bmatrix}$ então o vetor escore $\boldsymbol{U} = \begin{bmatrix} U_1 \\ \ldots \\ U_P \end{bmatrix}$ tem distribuição Normal Multivariada $U \sim N(0, \ )$, pelo menos assintoticamente, de forma que

$$\boldsymbol{U}^T \ ^{-1} \boldsymbol{U} \sim \chi^2(p)$$

para amostras grandes.

Exemplos: 1) Estatística escore para a distribuição Normal

Sejam $Y_1, \ldots Y_N$ variáveis aleatórias independentes e identicamente distribuídas com $Y_i \sim N(\mu, \sigma^2)$ onde σ^2 é uma constante conhecida. A função log-verossimilhança é

$$l = \quad \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i \quad \mu)^2 \quad N log\left(\sigma\sqrt{2\pi}\right).$$

A estatística escore é

$$U = \frac{dl}{d\mu} = \frac{1}{\sigma^2} \sum (Y_i \quad \mu) = \frac{N}{\sigma^2} \, (\bar{Y} \quad \mu),$$

de modo que o estimador de máxima verossimilhança, obtido ao solucionar a equação $U = 0$, é $\mu = \bar{Y}$. O valor esperado da estatística U é

$$E(U) = \frac{1}{\sigma^2} \sum [E(Y_i \quad \mu)].$$

Como $E(Y_i) = \mu$, segue que $E(U) = 0$ como era esperado. A variância de U é

$$= var(U) = \frac{1}{\sigma^4} \sum var(Y_i) = \frac{N}{\sigma^2}$$

e $Var(Y_i) = \sigma^2$. Portanto

$$\frac{U}{\sqrt{}} = \frac{(\bar{Y} \quad \mu)}{\sigma/\sqrt{N}}.$$

Este resultado tem distribuição $N(0,1)$ assintótica e, de fato, é exato porque $\bar{Y} \sim N(\mu, \sigma^2/N)$. Similarmente,

$$U^T \ ^{-1}U = \frac{U^2}{} = \frac{(Y \quad \mu)^2}{\sigma^2/N} \sim \chi^2(1)$$

é um resultado exato. A distribuição amostral de U pode ser usada para fazer inferências sobre μ. Por exemplo, um intervalo de 95% de confiança para μ é

$$\bar{y} \pm 1{,}96\, \sigma/N$$

onde σ é conhecido.

2) Estatística "escore" para a distribuição Binomial

Se $Y \sim Binomial\ (N, \pi)$, então a função de verossimilhança é:

$$l(\pi, y) = y \log \pi + (N \quad y) \log(1 \quad \pi) + log \binom{N}{y}$$

E a estatística "escore", que a primeira derivada da função de verossimilhança avaliada no parâmetro π é

$$U = \frac{dl}{d\pi} = \frac{Y}{\pi} \quad \frac{N \quad Y}{1 \quad \pi} = \frac{Y \quad N\pi}{\pi(1 \quad \pi)}$$

Mas $E(Y) = N\pi$, e então $E(U) = 0$, como esperado. Também, $Var(Y) = N\pi(1 \quad \pi)$ então

$$= Var(U) = \frac{1}{\pi^2(1 \quad \pi)^2} Var(Y) = \frac{N}{\pi(1 \quad \pi)}$$

E dessa forma

$$\frac{U}{\sqrt{}} = \frac{Y \quad N\pi}{\sqrt{N\pi(1 \quad \pi)}} \sim N(0,1)$$

aproximadamente. Esta é a aproximação Normal para a distribuição Binomial (sem nenhuma correção de continuidade). É usada para achar intervalo de confiança para a Binomial, além de testar hipóteses para π.

2.4.3 - Distribuição amostral para estimadores de máxima verossimilhança

Para uma função log-verossimilhança de um único parâmetro β, os primeiros três termos da Série de aproximação de Taylor perto de uma estimativa b são:

$$l(\beta) = l(b) + (\beta \quad b)U(b) + \frac{1}{2}(\beta \quad b)^2 U'(b)$$

onde $U(b) = {dl}/{d\beta}$ é a função escore avaliada em $\beta = b$. Se $U' = {d^2 l}/{d\beta^2}$ é aproximada pelo seu valor esperado $E(U') =$, a aproximação se torna

$$l(\beta) = l(b) + (\beta \quad b)U(b) \quad \frac{1}{2}(\beta \quad b)^2 \quad (b)$$

onde (b) é a informação avaliada no ponto $\beta = b$. A aproximação correspondente para a função log-verossimilhança para o vetor de parâmetros **β** é:

$$l(\boldsymbol{\beta}) = l(\boldsymbol{b}) + (\boldsymbol{\beta} \quad \boldsymbol{b})^T \boldsymbol{U}(\boldsymbol{b}) \quad \frac{1}{2}(\boldsymbol{\beta} \quad \boldsymbol{b})^T \boldsymbol{\mathcal{I}}(\boldsymbol{b})(\boldsymbol{\beta} \quad \boldsymbol{b}) \ (\mathbf{2.36})$$

onde $\boldsymbol{U}$ é o vetor de estatísticas escore e $\boldsymbol{\mathcal{I}}$ é matriz informação.

Para a função escore de parâmetro único β os dois primeiros termos da aproximação por Série de Taylor perto de uma estimativa b são:

$$U(\beta) = U(b) + (\beta \quad b)U'(b)$$

Se U' é aproximada por $E(U') =$, obtemos

$$U(\beta) = U(b) \quad (\beta \quad b) \ (b)$$

A expressão correspondente para um vetor de parâmetros **β** é:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{b}) \quad \boldsymbol{\mathcal{I}}(\boldsymbol{b})(\boldsymbol{\beta} \quad \boldsymbol{b}) \ (\mathbf{2.37})$$

Supondo que o logaritmo da função de verossimilhança tem um único máximo em $\hat{\beta}$ e que está próximo do verdadeiro valor de β. Então a equação $(\mathbf{2.37})$ pode ser usada para obter a distribuição amostral do estimador de máxima verossimilhança $\boldsymbol{b} = \widehat{\boldsymbol{\beta}}$**.** Por definição, $\boldsymbol{b}$ é o estimador que maximiza $l(\boldsymbol{b})$ e então $\boldsymbol{U}(\boldsymbol{b}) = \mathbf{0}$**.** Portanto,

$$\boldsymbol{U}(\boldsymbol{b}) = \quad \boldsymbol{\mathcal{I}}(\boldsymbol{b})(\boldsymbol{\beta} \quad \boldsymbol{b})$$

ou, equivalentemente

$$(\boldsymbol{b} \quad \boldsymbol{\beta}) = \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{U}$$

sendo que $\boldsymbol{\mathcal{I}}$ é não singular. Se $\boldsymbol{\mathcal{I}}$ é vista como constante então $E(\boldsymbol{b} \quad \boldsymbol{\beta}) = 0$ porque $E(\boldsymbol{U}) = \boldsymbol{0}$. Portanto $E(\boldsymbol{b}) = \boldsymbol{\beta}$, ao menos assintoticamente, então $\boldsymbol{b}$ é um estimador consistente para $\boldsymbol{\beta}$. A matriz de variância-covariância para b é:

$$E[(\boldsymbol{b} \quad \boldsymbol{\beta})(\boldsymbol{b} \quad \boldsymbol{\beta})^T] = \boldsymbol{\mathcal{I}}^{-1} E(\boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{\mathcal{I}} = \boldsymbol{\mathcal{I}}^{-1} \quad (\mathbf{2.38}).$$

Porque $\boldsymbol{\mathcal{I}} = E(\boldsymbol{U}\boldsymbol{U}^T)$ e $(\boldsymbol{\mathcal{I}}^{-1})^T = \boldsymbol{\mathcal{I}}^{-1}$, como é simétrica.

A distribuição amostral assintótica para $\boldsymbol{b}$, por $(\mathbf{2.38})$ é:

$$(\boldsymbol{b} \quad \boldsymbol{\beta})^T \ (\boldsymbol{b})(\boldsymbol{b} \quad \boldsymbol{\beta}) \sim \chi^2(p) \ (\mathbf{2.39}).$$

Esta é a Estatística de Wald. Para o caso de um parâmetro, a forma assintótica mais comumente usada é:

$$b \sim N(\beta, \boldsymbol{\mathcal{I}}^{-1})(\mathbf{2.40}).$$

Se as variáveis respostas no modelo linear generalizado são normalmente distribuídas então $(\mathbf{2.39})$ e $(\mathbf{2.40})$ são resultados exatos. O exemplo abaixo comprovará tal fato:

Exemplo: Estimadores de Máxima Verossimilhança para o Modelo linear Normal

Considere o modelo

$$Y_i = \mu_i = \boldsymbol{x_i}^{\boldsymbol{T}}\boldsymbol{\beta}, \quad Y_i \sim N(\mu, \sigma^2) \ (\mathbf{2.41})$$

onde os $Y_i´s$ são N variáveis aleatórias independentes e $\boldsymbol{\beta}$ é um vetor de p parâmetros $(p < N)$. Este é um modelo linear generalizado com função identidade como função de ligação. Devido a tal fato, na equação $(\mathbf{2.16})$, $\mu_i = \eta_i$ e então ${}^{\delta\mu_i}/_{\delta\eta_i} = 0$.

Os elementos da matriz informação, dada pela equação $(\mathbf{2.22})$, têm uma forma mais simples:

$$_{ik} = \sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{\sigma^2}$$

Porque $Var(Y_i) = \sigma^2$. Portanto a matriz informação pode ser escrita assim:

$$\boldsymbol{\mathcal{I}} = \frac{1}{2}\boldsymbol{X^T X} \ (\mathbf{2.42}).$$

Similarmente a expressão em $(\mathbf{2.25})$ tem uma forma mais simples:

$$z_i = \sum_{i=1}^{p} x_{ik} \, b_k{}^{(m-1)} + (y_i \quad \mu_i).$$

Mas μ_i avaliado em $\boldsymbol{b}^{(m-1)}$ é $\boldsymbol{x_i}^T\boldsymbol{b}^{(\boldsymbol{m-1})} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{p}} \boldsymbol{x_{ik}} \, \boldsymbol{b_k}^{(\boldsymbol{m-1})}$. Portanto $z_i = y_i$ neste caso.

A equação de estimação $(\mathbf{2.25})$ é:

$$\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}$$

e então o estimador de máxima verossimilhança é

$$\boldsymbol{b} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\ (\mathbf{2.43}).$$

O modelo $(\mathbf{2.41})$ pode ser escrito em notação vetorial como $\boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{\mathcal{I}})$ onde $\boldsymbol{\mathcal{I}}$ é uma matrix unitária NxN com 1´s na diagonal principal e zero nos outros campos.

$$E(\boldsymbol{b}) = (\boldsymbol{X^T X})^{-1}(\boldsymbol{X^T X\beta}) = \boldsymbol{\beta}$$

então $\boldsymbol{b}$ é um estimador não viciado para $\boldsymbol{\beta}$**.**

Para obter a matriz de variância-covariância para $\boldsymbol{b}$ usamos

$$\begin{aligned} \boldsymbol{b} \quad \boldsymbol{\beta} &= (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y} \quad \boldsymbol{\beta} \\ &= (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T}(\boldsymbol{y} \quad \boldsymbol{X\beta}) \end{aligned}$$

Então

$$\begin{aligned} E[(\boldsymbol{b} \quad \boldsymbol{\beta})(\boldsymbol{b} \quad \boldsymbol{\beta})^T] &= (\boldsymbol{X^T X})^{-1}\boldsymbol{X}^T E[(\boldsymbol{y} \quad \boldsymbol{X\beta})(\boldsymbol{y} \quad \boldsymbol{X\beta})^T]\boldsymbol{X}(\boldsymbol{X^T X})^{-1} \\ &= (\boldsymbol{X^T X})^{-1}\boldsymbol{X}^T[var(\boldsymbol{y})]\boldsymbol{X}(\boldsymbol{X^T X})^{-1} \\ &= \sigma^2(\boldsymbol{X^T X})^{-1} \end{aligned}$$

Mas $\sigma^2(\boldsymbol{X^T X})^{-1} = \boldsymbol{\mathcal{I}^{-1}}$**,** de acordo com $(\mathbf{2.42})$, então a matriz de variância-covariância para $\boldsymbol{b}$ é $\boldsymbol{\mathcal{I}^{-1}}$ como visto em $(\mathbf{2.38})$**.**

O estimador de máxima verossimilhança **b** é uma combinação linear dos elementos Y_i´s de $\boldsymbol{y}$ de acordo com $(\mathbf{2.43})$**.** Como os Y_i´s são normalmente distribuídos, de resultados já vistos, os elementos de $\boldsymbol{b}$ também são. Dessa forma a distribuição amostral de $\boldsymbol{b}$, neste caso, é:

$$\boldsymbol{b} \sim N(\boldsymbol{\beta}, \boldsymbol{\mathcal{I}^{-1}})$$

ou

$$(\boldsymbol{b} \quad \boldsymbol{\beta})^T\boldsymbol{\mathcal{I}}(\boldsymbol{b} \quad \boldsymbol{\beta}) \sim \chi^2(p)$$

Sendo assim, foi encontrada uma quantidade pivotal razoável para construção de intervalos de confiança para o parâmetro β utilizando a estatística, descrita acima, $(\boldsymbol{b} \quad \boldsymbol{\beta})^T\boldsymbol{\mathcal{I}}(\boldsymbol{b} \quad \boldsymbol{\beta})$**.** Esta expressão depende claramente do valor do parâmetro de interesse, todavia sua distribuição não depende, pois é uma qui-quadrado de parâmetro p.

2.4.4 – Estatística razão da Log-verossimilhança

Um modo de verificar a adequação de um modelo é compará-lo com um modelo mais geral com o maior número de parâmetros possível. Este é chamado de modelo saturado. É um modelo linear generalizado com a mesma distribuição e a mesma função de ligação do modelo de interesse.

Se temos N observações $Y_i, i = 1, \ldots, N$, todas com valores totalmente diferentes para o componente linear $\boldsymbol{x_i}^T\boldsymbol{\beta}$ então um modelo saturado pode conter N parâmetros. Este pode ser chamado um modelo "máximo" ou "cheio".

Se algumas das observações têm o mesmo componente linear ou covariável padrão, ou seja, se elas correspondem à mesma combinação de níveis de fatores e têm os mesmos valores de quaisquer variáveis explicativas contínuas, elas são chamadas réplicas. Neste caso, o número máximo de parâmetros que pode ser estimado para o modelo saturado é igual ao número de componentes lineares potencialmente diferentes, que pode ser menos que N.

Em geral, denota-se por m o número máximo de parâmetros que pode ser estimado. Seja $\boldsymbol{\beta}_{max}$ o vetor de parâmetros para o modelo saturado e $\boldsymbol{b}_{max}$ o estimador de máxima verossimilhança para $\boldsymbol{\beta}_{max}$. A função de verossimilhança para o modelo saturado avaliado em $\boldsymbol{b}_{max}$, $L(\boldsymbol{b}_{max}; \boldsymbol{y})$, será maior do que qualquer outra função de verossimilhança para essas observações, com a mesma distribuição e função de ligação assumidas, porque isso garante a descrição mais completa dos dados. Seja $L(\boldsymbol{b}; \boldsymbol{y})$ o valor máximo da função de verossimilhança para o modelo de interesse. Então a razão de verossimilhança:

$$\lambda = \frac{L(\boldsymbol{b}_{max}; \boldsymbol{y})}{L(\boldsymbol{b}; \boldsymbol{y})}$$

provê uma maneira de verificar a bondade do ajuste para o modelo. Na prática, o logaritmo da razão de verossimilhança, que é a diferença entre as funções de verossimilhança,

$$\log \lambda = l(\boldsymbol{b}_{max}; \boldsymbol{y}) \quad l(\boldsymbol{b}; \boldsymbol{y})$$

é usada**.** Valores altos para o $\log \lambda$ sugerem que o modelo de interesse descreve pobremente a realidade dos dados relativamente ao modelo saturado. Para determinar a região crítica para $\log \lambda$ precisa-se da sua distribuição amostral.

É possível mostrar que $D = 2 \log \lambda$, chamada de estatística deviance, tem distribuição aproximadamente χ^2 não centralizada, com N-p graus de liberdade, onde N é o número de parâmetros no modelo saturado e p, o número de parâmetros no modelo de interesse.

2.4.5) Procedimento geral para teste de hipóteses

Considere a definição do processo usual de testes de hipóteses a fim de verificar quão bem dois modelos relacionados se ajustam aos dados, de maneira que o modelo mais simples, que corresponde à hipótese nula H_0, deve ser um caso especial do outro modelo mais geral. Se o modelo mais simples se adaptar aos dados, bem como o modelo mais geral, então é preferível por motivos de parcimônia H_0 ser mantida. Se o modelo mais geral encaixa significativamente melhor, em seguida, H_0 é rejeitado em favor da hipótese alternativa H_1 que corresponde ao modelo mais geral. Para fazer estas comparações, usamos estatísticas de resumo para descrever quão bem os modelos se ajustam aos dados. Estas estatísticas de bondade do ajuste podem ser baseadas no valor máximo da função de verossimilhança, no valor máximo da função de log-verossimilhança, no valor mínimo do critério da soma dos quadrados ou em uma estatística de síntese com base nos resíduos. O processo e a lógica podem ser resumidos como se segue:

1. Especifique um modelo M_0 correspondente a H_0. Especifique um modelo mais geral M_1 (com M_0 como um caso especial de M_1).

2. Ajuste M_0 e calcule a estatística G_0 de bondade do ajuste. Ajuste M_1 e calcule a estatística G_1 de bondade do ajuste.

3. Calcule a melhoria no ajuste, geralmente $G_1 \quad G_0$, mas G_1/G_0 é outra possibilidade.

4. Use a distribuição amostral de $G_1 \quad G_0$ (ou alguma estatística relacionada) para testar a hipótese nula de que $G_1 = G_0$ contra a hipótese alternativa $G_1 \neq G_0$.

5. Se a hipótese $G_1 = G_0$ não é rejeitada, então H_0 não é rejeitada e M_0 é o modelo preferido. Se a hipótese $G_1 = G_0$ é rejeitada então H_0 é rejeitada e M_1 é considerado como o melhor modelo.

2.4.6) Utilizando o teste de hipóteses

Hipóteses sobre um vetor de parâmetros β de tamanho p podem ser testadas usando a distribuição amostral da estatística de Wald apresentada em $(\mathbf{2.42})$**.** Ocasionalmente a estatística de escore é usada: $\boldsymbol{U}^T \ ^{-1}\boldsymbol{U} \sim \chi^2(p)$.

A seguir, apresenta-se uma abordagem para comparar a qualidade do ajuste de dois modelos. Os modelos têm de ser hierarquizados, isto é, eles devem ter a mesma distribuição de probabilidade e a mesma função de ligação, mas a componente linear do modelo mais simples M_0 é um caso especial do componente linear do modelo mais geral M_1.

Considere a hipótese nula:

$$H_0: \beta = \beta_0 = \begin{bmatrix} \beta_1 \\ \dots \\ \beta_q \end{bmatrix}$$

Correspondente ao Modelo M_0 e uma hipótese mais geral

$$H_1: \beta = \beta_1 = \begin{bmatrix} \beta_1 \\ \dots \\ \beta_p \end{bmatrix}$$

Correspondente ao M_1, com q < p < N.

Podemos testar H_0 contra H_1usando a diferença das estatísticas deviance

$$\Delta D = D_0 \quad D_1 = 2[\, l(b_{max}; y) \quad l(b_0; y)] \quad 2[l(b_{max}; y) \quad l(b_1; y)]$$
$$= 2[l(b_1; y) \quad l(b_0; y)$$

Se ambos os modelos descrevem bem os dados, então $D_0 \sim \chi^2(N \quad q)$ e $D_1 \sim \chi^2(N \quad p)$ de modo que $\Delta D \sim \chi^2(p \quad q)$, desde que certa condição de independência aconteça. Se o valor de ΔD é consistente com distribuição $\chi^2(p \quad q)$ então decidiremos pelo modelo M_0 correspondente a H_0 por ser mais simples.

Se o valor de ΔD está na região crítica (isto é, maior que 100xα% da calda superior da distribuição $\chi^2(p \quad q)$) então rejeitamos H_0 em favor de H_1 devido ao modelo M_1 ser significativamente melhor na descrição dos dados (embora isso também não possa particularmente se ajustar bem aos dados).

2.5) Critérios para seleção de modelos

2.5.1 - Critério do AIC

O método proposto por Akaike (1974) é definido como

$$AIC = \quad 2\log L(\hat{\theta}) + p$$

em que $L(\hat{\theta})$ é a função de máxima verossimilhança do modelo e p é o número de variáveis explicativas consideradas no modelo. O intuito do critério é orientar na seleção dos modelos e prover uma medida de informação que equilibre um bom ajuste com uma quantidade pequena de parâmetros, ou seja, um modelo parcimonioso. Como o logaritmo da função de verossimilhança $L(\hat{\theta})$ cresce com o aumento do número de parâmetros do modelo, naturalmente procuramos um modelo com menor número de variáveis possíveis para a função acima, na medida em que essa escolha não comprometa uma boa capacidade de previsão. Modelos com mais variáveis tendem a produzir menor SQE, porém usam mais parâmetros. A melhor escolha é balancear o ajuste com a quantidade de variáveis.

O critério AIC provê um método de garantir a qualidade do modelo através de uma comparação de modelos relacionados. É baseado na função *desvio*, mas penaliza por deixar o modelo mais complicado. O intuito é impedir que se adicione preditores irrelevantes.

O número de AIC, sozinho, não tem sentido. Se há mais de um modelo candidato similar, onde todas as variáveis do modelo mais simples ocorrem nos modelos mais complexos, então se deve selecionar o modelo com o menor AIC.

A função de verossimilhança, $L(\hat{\theta})$, utilizada foi a Poisson, uma vez que é a distribuição de probabilidade no modelo log-linear embutido no cálculo. Para verificar se o modelo ajustado é adequado para descrever o comportamento dos dados, foram utilizados os

critérios de erro absoluto, erro relativo e erro quadrático médio, conforme são apresentados a seguir.

2.5.2 - Erros Absoluto e Quadrático

Uma vez que sejam obtidos estimadores de máxima verossimilhança$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2,\ \hat{\beta}_3, \ldots \hat{\beta}_p$, pela propriedade de invariância dos estimadores de máxima verossimilhança, tem-se estimador de máxima verossimilhança para a resposta média, dado por $\hat{\mu}_i = g^{-1}(\eta_i)$, sendo g a função de ligação do modelo e η_i o preditor linear para a i-ésima unidade amostral, dado por:

$$\eta_i = X\beta\ = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + \beta_3 x_3(i) + \ldots + \beta_p x_p(i)\text{ , para } i = 1, \ldots, N.$$

É possível, então, obter-se o erro absoluto médio resultante do ajuste:

$$EAM = \frac{\sum_{i=1}^{N} |y_i - \mu_i|}{N}.$$

Pode-se, ainda, relativizar os erros absolutos, com respeito à resposta observada.

O erro quadrático médio do ajuste é dado por:

$$EQM = \frac{\sum_{i=1}^{N} (y_i - \mu_i)^2}{N}.$$

Tendo em mente toda a teoria apresentada neste capítulo, a parte prática do projeto se desenvolverá através de uma abordagem clássica, onde se objetivará encontrar um modelo visando a parcimônia e capacidade preditiva. Uma análise descritiva inicial dos dados será realizada, bem como a devida interpretação de saídas e resultados.

Capítulo 3 - Aplicação a dados reais

3.1) Análise Descritiva

3.1.1 - Apresentação dos dados

A base de dados analisada neste projeto baseia-se nas informações dos 376.810 segurados da empresa P&R Seguros que contrataram a cobertura de Assistência 24h. Esta é uma cobertura, adicional ao seguro, comercializada pela seguradora onde o cliente, optando pela contratação, poderá usufruir de diversos serviços, tais como: chaveiro, guincho, motorista amigo, técnico, entre outros. Para chegar na estrutura final da base de estudo, foi selecionado por um período de dois anos todos os clientes que acionaram ao menos uma vez algum serviço neste espaço de tempo. As quantidades foram sumarizadas para cada segurado, resultando na frequência absoluta (ou contagem) de utilização da carteira segurada, a qual será a covariável resposta de interesse no modelo estudado. Além desta informação, na base há também dados qualitativos ligados tanto ao perfil dos clientes quanto dos seus veículos, distribuídas no total de 51 variáveis, como seguem abaixo:

- **Frequência:** Variável contendo a contagem de utilização de cada indivíduo.
- **Estado Civil**: Casado(a), Separado(a), Solteiro(a), Viúvo(a) e Não preencheu.
- **Sexo**: Feminino e Masculino.
- **Idade do veículo**: As idades dos veículos que estão divididas nas faixas, 0-5; 6-11; 12-17; 18-25 anos.
- **Idade do condutor**: Segregada nas faixas de idade: 18-23; 24-29; 30-35; 36-41; 42-47; 48-53; 54-59; 60-65; 66-71; 72-77; acima de 78 anos.
- **Combustível utilizado**: Gasolina, bicombustível, diesel, tricombustível e álcool.
- **Marca do veículo**: Fiat, Chevrolet, Volkswagen, Ford, Peugeot, Honda, Renault, Citröen, Toyota, Hyundai, Nissan, Kia Motors, Jac Motors, Mitsubishi e Seat Ibiza.
- **Modelo do veículo**: Mini Van, Station Wagon, Pick Up, Blazer, Jipe, Furgão/Van, Sedan e Hatch.

3.1.2 - Análise Exploratória dos dados

Inicialmente foram criadas variáveis *"dummies"* para representar numericamente os dados qualitativos contidos na base estudada. Todas as regressoras disponíveis são variáveis qualitativas ou quantitativas categorizadas. Estas serão as variáveis independentes adotadas em modelos a serem ajustados no pacote R.

Para cada variável, identificou-se a categoria modal. Tais categorias são denominadas, daqui por diante, categorias de base, sendo estas: sexo masculino, estado civil casado, idade na faixa de 30-35 anos, marca Chevrolet, Bicombustível, modelo hatch e 0-5 anos de fabricado, conforme o disposto nas tabelas 3.1 e 3.2, que exibem estatísticas descritivas de covariáveis associadas ao segurado e ao veículo, respectivamente. As linhas em negrito referem-se categorias de base.

Tabela 3.1 - Características do Segurado

Características do Segurado

Estado Civíl	Qtde Registros	Freq de Utilização	Freq Média($\overline{x}$)	Variância (s^2)	$\lvert\overline{x} - s^2\rvert$
Separado(a)	19.710	36.578	1,856	1,773	0,083
Viúvo(a)	6.932	12.561	1,812	1,597	0,215
Solteiro(a)	92.433	170.487	1,844	1,741	0,104
Casado(a)	**252.820**	**451.297**	**1,785**	**1,589**	**0,196**
Não preencheu	4.915	8.718	1,774	1,659	0,115

Faixa Etária	Qtde Registros	Freq de Utilização	Freq Média($\overline{x}$)	Variância (s^2)	$\lvert\overline{x} - s^2\rvert$
Fx_18_23	6.407	11.972	1,869	1,644	0,225
Fx_24_29	41.328	75.470	1,826	1,675	0,151
Fx_30_35	**85.954**	**156.730**	**1,823**	**1,668**	**0,156**
Fx_36_41	64.481	117.571	1,823	1,700	0,123
Fx_42_47	47.404	85.910	1,812	1,736	0,077
Fx_48_53	43.634	78.403	1,797	1,623	0,174
Fx_54_59	37.550	66.854	1,780	1,543	0,237
Fx_60_65	25.732	44.809	1,741	1,493	0,249
Fx_66_71	14.024	24.301	1,733	1,433	0,300
Fx_72_77	6.795	11.565	1,702	1,332	0,370
Fx_>=78	3.501	6.056	1,730	1,532	0,198

Sexo	Qtde Registros	Freq de Utilização	Freq Média($\overline{x}$)	Variância (s^2)	$\lvert\overline{x} - s^2\rvert$
Feminino	153.068	275.570	1,800	1,646	0,154
Masculino	**223.742**	**404.071**	**1,806**	**1,632**	**0,174**

Tabela 3.2 - Características do Veículo

Características do Veículo

Tipo de Combustível	Qtde Registros	Freq de Utilização	Freq Média($\bar{x}$)	Variância (s^2)	$\lvert\bar{x} - s^2\rvert$
Gasolina	97.987	189.149	1,930	2,023	0,093
Diesel	1.568	3.116	1,987	2,172	0,185
Tricombustível	340	602	1,771	1,341	0,429
Álcool	1.085	2.087	1,924	2,080	0,156
Bicombustível	**275.830**	**484.687**	**1,757**	**1,488**	**0,269**

Modelo Veículo	Qtde Registros	Freq de Utilização	Freq Média($\bar{x}$)	Variância (s^2)	$\lvert\bar{x} - s^2\rvert$
Sedan	90.143	161.834	1,795	1,612	0,184
Hatch	**205.723**	**369.064**	**1,794**	**1,606**	**0,188**
Mini Van	39.232	73.157	1,865	1,853	0,012
Station Wagon	14.218	27.002	1,899	1,884	0,015
Pick Up	20.068	35.425	1,765	1,488	0,277
Blazer	4.455	7.726	1,734	1,422	0,312
Jipe	1.940	3.347	1,725	1,249	0,477
Furgão/Van	1.031	2.086	2,023	2,892	0,869

Faixa Idade do VHL	Qtde Registros	Freq de Utilização	Freq Média($\bar{x}$)	Variância (s^2)	$\lvert\bar{x} - s^2\rvert$
Idade_VHL_0_5	**180.317**	**306.486**	**1,700**	**1,303**	**0,397**
Idade_VHL_6_11	154.074	287.399	1,865	1,806	0,059
Idade_VHL_12_17	38.541	78.078	2,026	2,304	0,279
Idade_VHL_18_25	3.878	7.678	1,980	2,691	0,711

Marca do Veículo	Qtde Registros	Freq de Utilização	Freq Média($\bar{x}$)	Variância (s^2)	$\lvert\bar{x} - s^2\rvert$
Fiat	72.309	128.378	1,775	1,568	0,207
Chevrolet	**94.336**	**171.183**	**1,815**	**1,645**	**0,170**
Volkswagen	69.151	125.194	1,810	1,688	0,122
Ford	44.540	81.272	1,825	1,697	0,127
Peugeot	20.168	37.939	1,881	1,774	0,107
Honda	21.424	37.771	1,763	1,517	0,246
Renault	19.917	35.982	1,807	1,708	0,098
Citroen	12.072	22.609	1,873	1,849	0,023
Toyota	7.285	12.200	1,675	1,282	0,393
Hyundai	5.920	10.357	1,749	1,441	0,309
Nissan	4.127	6.855	1,661	1,262	0,399
Kia Motors	2.949	5.282	1,791	1,463	0,328
Jac Motors	1.507	2.714	1,801	1,495	0,306
Mitsubishi	1.104	1.901	1,722	1,237	0,485
Seat Ibiza	1	4	4,000	0,000	4,000

Ao analisar os quadros acima (na 6ª coluna), observa-se em geral que os valores de $\bar{x}$ e s^2 são bem próximos. Além disso, verifica-se que o mesmo ocorre para variável frequência (freq), onde $\bar{x} = 1{,}804$ e $s^2 = 1{,}638$, o que permite desconsiderar a hipótese de sobredispersão, ou seja, $\mu << \sigma^2$.

3.2) Modelagem

Assim, como tem-se dados de contagem, sem indícios de presença de sobredispersão, adota-se um modelo Poisson modelo log-linear para a frequência de utilização do serviço, dado por:

$$Y_i \sim Pois(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_1(i) + \beta_2 x_2(i) + \beta_3 x_3(i) + \ldots + \beta_p x_p(i)\text{, para } i = 1, \ldots, N.$$

Foram testadas diversas combinações de modelos, como descrito adiante. Os p-valores indicados nas saídas do pacote referem-se ao teste de significância para o coeficiente de cada variável regressora. Como todas as variáveis explicativas inseridas nos modelos propostos são dummies e os ajustes foram feitos omitindo-se categorias de base, de forma a evitar colinearidade, tais testes indicam se há diferença significativa na frequência de utilização do serviço por cada categoria de uma dada característica, em relação à categoria de base, preservadas as demais condições. A título de ilustração, tome-se a saída de um modelo contendo apenas estado civil como variável explicativa e casados como categoria de base, como exibido na Tabela 3.3.

Tabela 3.3 - Saída do R referente à análise individual da covariável estado civil

Coeficientes	Estimativa	Std. Error	z value	Pr(>|z|)	
(Intercepto)	0,579448	0,001489	389.265	< 2e-16	***
I_SEPARA	0,038873	0,005436	7.150	8,65E-13	***
I_VIUVO	0,015001	0,009046	1.658	0,0973	.
I_SOLT	0,032727	0,002843	11.512	< 2e-16	***
I_NPREN	-0,006350	0,010813	-0,587	0,557	

AIC: 1152568

Null deviance*: 245309 *on* 361808 *degrees of freedom

Residual deviance*: 245155 *on* 361804 *degrees of freedom

***Number of Fisher Scoring iterations*: 5**

De acordo com esta saída, indivíduos separados e solteiros têm frequência de utilização significativamente diferente de casados, ao contrário do que ocorre com viúvos e indivíduos que não informaram estado civil. Ainda utilizando o resultado acima de forma ilustrativa, teríamos uma estimativa pontual igual a $\exp(0{,}032727) \approx 1{,}03$, indicando uma utilização, por solteiros, em média, 3% superior à de indivíduos casados.

Se utilizássemos somente essa característica no modelo, ou seja, se somente estado civil impactasse a variável resposta, teríamos um modelo para previsão:

$$log\,(\hat{\mu}_i) = 0{,}579448 + 0{,}038873\ \ x_{SEPARA} + 0{,}015001\ \ x_{VIUVO} + 0{,}032727\ \ x_{SOLT} - 0{,}006350\ \ x_{NPREN},$$

de tal forma que a frequência esperada de utilização entre indivíduos casados é $\exp(0{,}579448) \approx 1{,}78$ e ao se exponenciar as estimativas pontuais de cada um dos demais coeficientes, tem-se o incremento ou decréscimo de utilização esperada associada a cada categoria, comparativamente aos casados.

O R reporta duas formas de desvio: "$Null\ deviance$" conhecido como desvio nulo e o "$Residual\ deviance$", desvio residual. O primeiro mostra o quão bem a variável resposta é prevista por um modelo que inclui somente o intercepto. O segundo é relacionado ao modelo que inclui todas as variáveis propostas.
$Desvio\ Nulo = 2(LL(Modelo\ Saturado) \quad LL(Modelo\ nulo))$, sendo LL a função log-verossimilhança, com graus de liberdade igual aos graus de liberdade do modelo saturado menos do modelo nulo.
$Desvio\ Residual = 2\big(LL(Modelo\ Saturado) \quad LL(Modelo\ Proposto)\big)$, sendo LL a função log-verossimilhança, com graus de liberdade igual aos graus de liberdade do modelo saturado menos do modelo residual.

O modelo saturado é um modelo que assume que cada ponto de dados tem seus próprios parâmetros, ou seja, temos N parâmetros para estimar.

O modelo nulo já nos diz exatamente o contrário, de forma que assume um parâmetro para todos os pontos de dados, o que significa que só temos que estimar um parâmetro.

O modelo proposto assume que podemos explicar os pontos de dados com p parâmetros mais um termo para o intercepto, totalizando $p + 1$ parâmetros para estimação. Caso o desvio nulo seja bem pequeno, isso significa que o modelo nulo explica os dados muito bem. Da mesma forma acontece com o desvio residual.

Além disso, pode-se verificar na saída do R a informação de número de iterações de Fisher. O Algoritmo de *scoring* de Fisher é uma derivada do método de Newton Raphson para solucionar problemas de máxima verossimilhança, numericamente. Na saída de R do exemplo está destacado que o Algoritmo precisou de 5 iterações para conseguir realizar o ajuste. Tal fato não diz muito, somente que o modelo de fato convergiu e que não teve dificuldades em fazer isso.

Tal procedimento foi feito para testar individualmente cada uma das variáveis explicativas: estado civil, sexo, idade do veículo, idade do condutor, marca do veículo e modelo do veículo. A partir daí pode-se pensar em modelos com essas variáveis qualitativas significativas, ponderando qual deve ser colocada, e se deve colocar interações entre essas variáveis.

A seguir na tabela 3.4 são listadas as variáveis que compõem os modelos ajustados à base de dados, após retiradas, aleatoriamente, 15.000 observações para fins de previsão.

Tabela 3.4 – Variáveis dos modelos

	Variáveis	Modelo Completo	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6
Estado Civil	Separado(a)	X	X	X	-	-	-	-
	Viúvo(a)	X	X	X	-	-	-	-
	Solteiro(a)	X	X	X	-	-	-	-
	Não preencheu	X	X	X	-	-	-	-
Faixa Etária	Fx_18_23	X	X	X	-	-	-	-
	Fx_24_29	X	X	X	-	-	-	-
	Fx_36_41	X	X	X	-	-	-	-
	Fx_42_47	X	X	X	-	-	-	-
	Fx_48_53	X	X	X	-	-	-	-
	Fx_54_59	X	X	X	-	-	-	-
	Fx_60_65	X	X	X	-	-	-	-
	Fx_66_71	X	X	X	-	-	-	-
	Fx_72_77	X	X	X	-	-	-	-
	Fx_>=78	X	X	X	-	-	-	-
Sexo	Feminino	X	X	X	-	-	-	-
Tipo de Combustível	Gasolina	X	X	X	X	X	X	X
	Diesel	X	X	X	X	X	X	X
	Tricombustível	X	-	-	X	X	X	X
	Álcool	X	-	-	X	X	X	X
Modelo de Veículo	Sedan	X	-	-	X	X	X	X
	Mini Van	X	X	X	X	X	X	X
	Station Wagon	X	X	X	X	X	X	X
	Pick Up	X	-	-	X	X	X	X
	Blazer	X	-	-	X	X	X	X
	Jipe	X	-	-	X	X	X	X
	Furgão/Van	X	X	X	X	X	X	X
Idade do Veículo	Idade_VHL_6_11	X	X	X	X	X	X	X
	Idade_VHL_12_17	X	X	X	X	X	X	X
	Idade_VHL_18_25	X	X	X	X	X	X	X
Montadora	Fiat	X	X	X	X	X	X	X
	Volkswagen	X	-	-	X	X	X	X
	Ford	X	-	-	X	X	X	X
	Peugeot	X	X	X	X	X	X	X
	Honda	X	X	X	X	X	X	X
	Renault	X	-	-	X	X	X	X
	Citroen	X	X	X	X	X	X	X
	Toyota	X	X	X	X	X	X	X
	Hyundai	X	-	-	X	X	X	X
	Nissan	X	X	X	X	X	X	X
	Kia Motors	X	-	-	X	X	X	X
	Jac Motors	X	-	-	X	X	X	X
	Mitsubishi	X	-	-	X	X	X	X
	Seat Ibiza	X	-	-	X	X	X	X
Combinações	Sex_Fem*Separada	-	-	X	-	-	X	-
	Sex_Fem*Viúva	-	-	X	-	-	X	-
	Sex_Fem*Solteira	-	-	X	-	-	X	-
	Sex_Fem*Não Info	-	-	X	-	-	X	-
	Sex_Fem*Fx_18_23	-	-	-	X	-	-	X
	Sex_Fem*Fx_24_29	-	-	-	X	-	-	X
	Sex_Fem*Fx_36_41	-	-	-	X	-	-	X
	Sex_Fem*Fx_42_47	-	-	-	X	-	-	X
	Sex_Fem*Fx_48_53	-	-	-	X	-	-	X
	Sex_Fem*Fx_54_59	-	-	-	X	-	-	X
	Sex_Fem*Fx_60_65	-	-	-	X	-	-	X
	Sex_Fem*Fx_66_71	-	-	-	X	-	-	X
	Sex_Fem*Fx_72_77	-	-	-	X	-	-	X
	Sex_Fem*Fx_>=78	-	-	-	X	-	-	X

Como critério de comparação de modelos, foram utilizados AIC, que leva em conta tanto a qualidade do ajuste quanto parcimônia de cada modelo, bem como erros absoluto e quadrático, tendo sido tais erros avaliados sobre as 15.000 observações reservadas para análise da capacidade preditiva. Considerando-se conjuntamente tais critérios, bem como interpretação de resultados, conclui-se que o **Modelo 2** seria o melhor a ser utilizado, por vias de capacidade preditiva, já que, de acordo com o critério AIC não há grande diferença de ajuste entre este modelo e o modelo completo.

Em seguida, a *Tabela 3.5* apresenta o critério AIC, os erros relativos, absolutos e quadráticos médios dos modelos propostos:

Tabela 3.5 - Comparativo dos tipos de erros

Modelo Selecionado	AIC	ERRO ABSOLUTO	ERRO QUADRÁTICO MÉDIO	ERRO ABSOLUTO RELATIVO
Completo	1103782	0,887580	1,548235	0,568071
Modelo 1	1103834	0,889834	1,557431	0,567596
Modelo 2	1103818	0,801964	2,097605	0,287273
Modelo 3	1103806	0,890036	1,557209	0,569463
Modelo 4	1104224	0,888738	1,551283	0,568934
Modelo 5	1103981	0,887877	1,549319	0,568198
Modelo 6	1103877	0,887639	1,549222	0,568171

Basicamente o que pesou na decisão pelo Modelo 2 foi o fato dele ter o menor erro absoluto relativo, razoavelmente menor que a maioria dos outros modelos que apresentaram erro relativo duas vezes maior.

Passamos a descrever os resultados referentes ao modelo selecionado.

3.3) Resultados Obtidos

A tabela 3.6 exibe a saída do R referente ao Modelo 2, o escolhido.

Tabela 3.6 - Saída do modelo 2 testado, via glm

	Estimate	Std.	Error	z	value	Pr(>|z|)
(Intercept)	0,536040	0,003554	150.823	<	2,00E-16	***
I_SEPARA	0,026520	0,009377	2.828	0,004680	**	
I_VIUVO	-0,012954	0,019320	-0,671	0,502537		
I_SOLT	0,030463	0,004240	7.185	0,000000	***	
I_NPREN	0,000631	0,011909	0,053	0,957743		
I_FEM	-0,008010	0,003192	-2.510	0,012086	*	
VHL_6_11	0,084707	0,002816	30.076	<	2,00E-16	***
VHL_12_17	0,160912	0,005401	29.794	<	2,00E-16	***
VHL_18_25	0,145673	0,012354	11.791	<	2,00E-16	***
I_GASOL	0,024966	0,003704	6.741	0,000000	***	
I_DISEL	0,136257	0,018417	7.398	0,000000	***	
ID_18_23	0,005225	0,009890	0,528	0,597265		
ID_24_29	-0,003008	0,004619	-0,651	0,514777		
ID_36_41	-0,004173	0,003989	-1.046	0,295556		
ID_42_47	-0,012244	0,004423	-2.768	0,005641	**	
ID_48_53	-0,020274	0,004579	-4.428	0,000010	***	
ID_54_59	-0,027277	0,004857	-5.616	0,000000	***	
ID_60_65	-0,052959	0,005629	-9.408	<	2,00E-16	***
ID_66_71	-0,063546	0,007214	-8.809	<	2,00E-16	***
ID_72_77	-0,082670	0,010010	-8.259	<	2,00E-16	***
ID_.78	-0,078411	0,013593	-5.769	0,000000	***	
I_FIA	-0,024024	0,003276	-7.334	0,000000	***	
I_PEU	0,042418	0,005545	7.650	0,000000	***	
I_HON	-0,040706	0,005729	-7.105	0,000000	***	
I_CIT	0,038804	0,007056	5.499	0,000000	***	
I_TOY	-0,078597	0,009433	-8.332	<	2,00E-16	***
I_NIS	-0,046135	0,012537	-3.680	0,000233	***	
MOD_MV	0,045574	0,004279	10.650	<	2,00E-16	***
MOD_SW	0,037058	0,006483	5.716	0,000000	***	
MOD_FV	0,132595	0,022835	5.807	0,000000	***	
I_SEPARA(FEM*)	0,033345	0,011712	2.847	0,004411	**	
I_VIUVO(FEM*)	0,075500	0,021937	3.442	0,000578	***	
I_SOLT(FEM*)	-0,009473	0,005942	-1.594	0,110890		
I_NPREN(FEM*)	0,012619	0,037784	0,334	0,738388		

**Possui carateristica da variável indicadora, sendo do sexo feminino.*

Observa-se que grande parte das variáveis foram significativas neste modelo, exceto as dummies: viúvo, não preencheu, faixas de idade do condutor de: 18-23 anos, 24-29, 36-41 anos, interação solteiro e sexo feminino, interação não preencheu e sexo feminino. A seguir, apresentam-se intervalos de confiança para o efeito de cada categoria. Pela propriedade de invariância dos intervalos de confiança, se um intervalo para θ é dado por $(L.I < \theta < L.S)$, então o intervalo de confiança para $\exp(\theta)$, que é transformação crescente do parâmetro original, é dado por $(\exp(\mathrm{L.I}) < \exp(\theta) < \exp(\mathrm{L.S.}))$.

Intervalos ao nível de confiança 95% para os parâmetros do preditor linear foram obtidos diretamente no pacote e a tabela 3.7 apresenta tais intervalos com limites exponenciados. Assim, tem-se 95% de confiança de que a utilização média na categoria de base (casados, homens, idades de 30 a 35, veículo Chevrolet, biocombustível, idade do veículo 0 a 5 anos, modelo hatch) seja algo entre 1,697 e 1,721. Indivíduos separados teriam utilização entre 0,8% e 4,6% maior que os indivíduos da categoria de base. Ainda solteiros apresentariam uma frequência esperada de utilização do serviço 2,24% a 3,95% maior que a categoria de base, mantendo-se fixas as demais características.

Tabela 3.7 Intervalo de Confiança para a exponencial do parâmetro θ

Nível de Confiança α=2.5%		
Covariável	**LimInf**	**LimSup**
(Intercept)	1,697360	1,721173
I_SEPARA	1,008175	1,045922
I_VIUVO	0,950448	1,025226
I_SOLT	1,022400	1,039534
I_NPREN	0,977545	1,024263
I_FEM	0,985836	0,998247
VHL_6_11	1,082407	1,094423
VHL_12_17	1,162139	1,187081
VHL_18_25	1,129144	1,185171
I_GASOL	1,017865	1,032750
I_DISEL	1,105348	1,188099
ID_18_23	0,985941	1,024914
ID_24_29	0,988012	1,006062
ID_36_41	0,988080	1,003653
ID_42_47	0,979304	0,996432
ID_48_53	0,971175	0,988763
ID_54_59	0,963872	0,982399
ID_60_65	0,938012	0,958941
ID_66_71	0,925256	0,951794
ID_72_77	0,902768	0,938896
ID_>=78	0,900278	0,949547
I_FIA	0,970015	0,982550
I_PEU	1,032053	1,054732
I_HON	0,949391	0,970953
I_CIT	1,025289	1,054044
I_TOY	0,907478	0,941663
I_NIS	0,931734	0,978668
MOD_MV	1,037887	1,055444
MOD_SW	1,024650	1,051024
MOD_FV	1,091813	1,194051
I_SEPARA(FEM*)	1,010445	1,057914
I_VIUVO(FEM*)	1,033038	1,125802
I_SOLT(FEM*)	0,979102	1,002176
I_NPREN(FEM*)	0,940414	1,090541

**Possui caraterística da variável indicadora, sendo do sexo feminino.*

Para as variáveis cujo intervalo de confiança contém a unidade, a interpretação é de que a frequência esperada de utilização não difere significativamente da categoria de base.

Por fim, segue abaixo a composição do modelo 2, escrito na sua forma completa, com todas as estimativas dos parâmetros β´s (conforme saída do R)

$$
\begin{aligned}
log\,(\hat{\mu}_i) = {} & 0{,}536040 + 0{,}026520\ \boldsymbol{xI_{SEPARA}}\ \ 0{,}012954\ \ \boldsymbol{xI_{VIUVO}} + 0{,}030463 \\
& \boldsymbol{xI_{SOLT}} + 0{,}000631\ \ \boldsymbol{xI_{NPREN}}\ \ 0.008010\ \ \boldsymbol{xI_{FEM}} + 0{,}084707 \\
& \boldsymbol{xVHL_{6_{11}}} + 0{,}160912\ \ \boldsymbol{xVHL_{12_{17}}} + 0{,}145673\ \ \boldsymbol{xVHL_{18_{25}}} + 0{,}024966 \\
& \boldsymbol{xI_{GASOL}} + 0{,}136257\ \ \boldsymbol{xI_{DISEL}} + 0{,}005225\ \ \boldsymbol{xID_{18_{23}}}\ \ 0{,}003008 \\
& \boldsymbol{xID_{24_{29}}}\ \ 0{,}004173\ \ \boldsymbol{xID_{36_{41}}}\ \ 0{,}012244\ \ \boldsymbol{xID_{42_{47}}}\ \ 0{,}020274 \\
& \boldsymbol{xID_{48_{53}}}\ \ 0{,}027277\ \ \boldsymbol{xID_{54_{59}}}\ \ 0{,}052959\ \ \boldsymbol{xID_{60_{65}}}\ \ 0{,}063546 \\
& \boldsymbol{xID_{66_{71}}}\ \ 0{,}082670\ \ \boldsymbol{xID_{72_{77}}}\ \ 0{,}078411\ \ \boldsymbol{xID_{.78}}\ \ 0{,}024024\ \ \boldsymbol{xI_{FIA}} \\
& + 0{,}042418\ \ \boldsymbol{xI_{PEU}}\ \ 0{,}040706\ \ \boldsymbol{xI_{HON}} + 0{,}038804\ \ \boldsymbol{xI_{CIT}}\ \ 0{,}078597 \\
& \boldsymbol{xI_{TOY}}\ \ 0{,}046135\ \ \boldsymbol{xI_{NIS}} + 0{,}045574\ \ \boldsymbol{xMOD_{MV}} + 0{,}037058 \\
& \boldsymbol{xMOD_{SW}} + 0{,}132595\ \ \boldsymbol{xMOD_{FV}} + 0{,}033345\ \ \boldsymbol{xI_{SEPARA}}\boldsymbol{:I_{FEM}} \\
& + 0{,}075500\ \ \boldsymbol{xI_{VIUVO}}\boldsymbol{:I_{FEM}}\ \ 0{,}009473\ \ \boldsymbol{xI_{SOLT}}\boldsymbol{:I_{FEM}} + 0{,}012619 \\
& \boldsymbol{xI_{NPREN}}\boldsymbol{:I_{FEM}}
\end{aligned}
$$

Pode-se testar esse modelo na previsão selecionando um indivíduo com alguns desses atributos. Cada característica, caso se faça presente no modelo, significa que a indicadora relacionada àquela estimativa valerá 1(um), caso contrário 0(zero), conforme mencionado anteriormente. A partir de então, foram calculadas estimativas para a frequência média de utilização do serviço de assistência 24 horas, que permitiram a obtenção dos erros absoluto e quadrático já apresentados.

O código R utilizado para ajuste de todos os modelos encontra-se no Apêndice.

Capítulo 4 – Conclusão

Devido à importância do seguro de automóveis para a sociedade e para esse ramo de mercado, este projeto teve como foco principal analisar a frequência de utilização do produto de assistência 24h, numa carteira de segurados da empresa P&R Seguros e para esse fim estudar e estruturar um modelo linear generalizado a fim de avaliar características que discriminem o perfil de utilização do produto e predizer a utilização por segurado, de acordo com seu perfil.

Inicialmente, foram testadas as variáveis explicativas, observando quais foram as mais significativas, via modelagem por GLM. Possíveis modelos utilizando combinações dessas variáveis explicativas também foram analisados, verificando qual desses apresentava melhor capacidade preditiva, através de critérios específicos.

Chegou-se à escolha do modelo 2, que demonstrou melhor capacidade preditiva, apresentando erro relativo médio em torno de 28%, para esta base de dados, em que a frequência média de utilização, por segurado, foi em torno de 1,8. Foi visto, nesse modelo, que as variáveis explicativas que mais influenciaram a resposta foram: idade do condutor, idade do veículo, tipo de combustível e marca. De tal modo, é plausível interpretar que uma pessoa mais jovem e solteira se "arriscaria" mais em certas situações, podendo levar à maior necessidade de utilização dos serviços-perfil de risco do seguro.

A idade está diretamente ligada à responsabilidade com manutenção do veículo. Em geral, quanto mais "madura" a pessoa mais experiente e consciente ela é. Esse dado não é apenas uma constatação por observação, os números de acidentes, roubos e colisões também estão ligados à idade e à forma como cada pessoa usa o veículo.

O modelo ajustado apontou um aumento na frequência esperada de utilização para veículos mais velhos, a diesel, e modelos furgão e van, que costumam ser utilizados para prestação de serviços, gerando maior desgaste. Por outro lado, a variável marca do veículo demonstrou risco reduzido de utilização associado a montadoras que reconhecidamente produzem veículos com boa mecânica.

Apêndice

Comandos no R

```
setwd("K:\\")
dados<-read.table("PROJET00_prev_Completo.csv",header=TRUE,sep=";")
head(dados)

hist(dados[,1],main="",xlab="Frequência",ylab="Probabilidade")
mean(dados[,1])
var(dados[,1])
```

Análise - Variável "Estado Civil"

```
modelo1<-glm(FREQ~I_SEPARA+I_VIUVO+I_SOLT+I_NPREN, data=dados, family=poisson)
summary(modelo1)
```

Análise - Variável "Sexo"

```
modelo2<-glm(FREQ~I_FEM, data=dados, family=poisson)
summary(modelo2)
```

Análise - Variável "Idade do Veículo"

```
modelo3<-glm(FREQ~VHL_6_11+VHL_12_17+VHL_18_25, data=dados, family=poisson)
summary(modelo3)
```

Análise - Variável "Idade do Condutor"

```
modelo4<glm(FREQ~ID_18_23+ID_24_29+ID_36_41+ID_42_47+ID_48_53+ID_54_59+ID_60_65+ID
_66_71+ID_72_77+ID_.78, data=dados, family=poisson)
summary(modelo4)
```

Análise - Variável "Combustível Utilizado"

```
modelo5<-glm(FREQ~I_GASOL+I_DISEL+I_TRIC+I_ALCOOL, data=dados, family=poisson)
summary(modelo5)
```

Análise - Variável "Marca do Veículo"

```
modelo6<glm(FREQ~I_FIA+I_VOL+I_FOR+I_PEU+I_HON+I_REN+I_CIT+I_TOY+I_HYU+I_NIS+I
_KIA+I_JAC+I_MIT+I_SEA, data=dados, family=poisson)
summary(modelo6)
```

Análise - Variável "Modelo do Veículo"

```
modelo7<-glm(FREQ~MOD_MV+MOD_SW+MOD_PK+MOD_BR+MOD_SE+MOD_FV+MOD_JI,
data=dados, family=poisson)
summary(modelo7)
```

Especificação do modelo completo (todas as variáveis)

```
modeloCompleto<glm(FREQ~I_SEPARA+I_VIUVO+I_SOLT+I_NPREN+I_FEM+VHL_6_11+VHL_
12_17+VHL_18_25+ID_18_23 + ID_24_29 + ID_36_41 + ID_42_47 +ID_48_53 + ID_54_59 + ID_60_65
+ ID_66_71 + ID_72_77 + ID_.78+I_GASOL + I_DISEL + I_TRIC + I_ALCOOL+I_FIA + I_VOL +
I_FOR + I_PEU + I_HON + I_REN + I_CIT + I_TOY + I_HYU + I_NIS + I_KIA + I_JAC + I_MIT
+I_SEA+MOD_MV + MOD_SW + MOD_PK + MOD_BR + MOD_SE +MOD_FV + MOD_JI,
data=dados, family=poisson)
summary (modeloCompleto)
exp(confint.default(modeloCompleto))
```

Especificação do modelo 1

```
modelo1<glm(FREQ~I_SEPARA+I_VIUVO+I_SOLT+I_NPREN+I_FEM+VHL_6_11+VHL_12_17+V
HL_18_25+I_GASOL+I_DISEL+ID_18_23 + ID_24_29 + ID_36_41 + ID_42_47 +ID_48_53 +
ID_54_59+ID_60_65+ID_66_71+ID_72_77+ID_.78+I_FIA+I_PEU+I_HON+I_CIT+I_TOY+I_NIS+M
OD_MV+MO     D_SW+MOD_FV, data=dados, family=poisson)
summary(modelo1)
exp(confint.default(modelo1))
```

Especificação do modelo 2

```
modelo2<glm(FREQ~I_SEPARA+I_VIUVO+I_SOLT+I_NPREN+I_FEM+VHL_6_11+VHL_12_17+V
HL_18_25+I_GASOL+I_DISEL+ID_18_23 + ID_24_29 + ID_36_41 + ID_42_47 +ID_48_53 +
ID_54_59+ID_60_65+ID_66_71+ID_72_77+ID_.78+I_FIA+I_PEU+I_HON+I_CIT+I_TOY+I_NIS+M
OD_MV+MOD_SW+MOD_FV+I_FEM*I_SEPARA+I_FEM*I_VIUVO+I_FEM*I_SOLT+I_FEM*I_
NPREN , data=dados, family=poisson)
summary(modelo2)
exp(confint.default(modelo2))
```

Especificação do modelo 3

```
modelo3<glm(FREQ~I_SEPARA+I_VIUVO+I_SOLT+I_NPREN+I_FEM+VHL_6_11+VHL_12_17+V
HL_18_25+I_GASOL+I_DISEL+ID_18_23 + ID_24_29 + ID_36_41 + ID_42_47 +ID_48_53 +
ID_54_59+ID_60_65+ID_66_71+ID_72_77+ID_.78+I_FIA+I_PEU+I_HON+I_CIT+I_TOY+I_NIS+M
OD_MV+MOD_SW+MOD_FV+I_FEM*ID_18_23+I_FEM*ID_24_29+I_FEM*ID_36_41+I_FEM*ID
_42_47+I_FEM*ID_48_53+I_FEM*ID_54_59+I_FEM*ID_60_65+I_FEM*ID_66_71+I_FEM*ID_72_
77+I_FEM*ID_.78 , data=dados, family=poisson)
summary(modelo3)
exp(confint.default(modelo3))
```

Especificação do modelo 4

```
modelo4<-glm(FREQ~VHL_6_11+VHL_12_17+VHL_18_25+I_GASOL + I_DISEL + I_TRIC +
I_ALCOOL+I_FIA + I_VOL + I_FOR + I_PEU + I_HON + I_REN + I_CIT + I_TOY + I_HYU + I_NIS
+ I_KIA + I_JAC + I_MIT + I_SEA+ MOD_MV + MOD_SW + MOD_PK + MOD_BR + MOD_SE +
MOD_FV + MOD_JI, data=dados, family=poisson)
summary(modelo4)
exp(confint.default(modelo4))
```

Especificação do modelo 5

```
modelo5<-glm(FREQ~VHL_6_11+VHL_12_17+VHL_18_25+I_GASOL + I_DISEL + I_TRIC + I_ALCOOL+I_FIA + I_VOL + I_FOR + I_PEU + I_HON + I_REN + I_CIT + I_TOY + I_HYU + I_NIS + I_KIA + I_JAC + I_MIT +I_SEA+ MOD_MV + MOD_SW + MOD_PK + MOD_BR + MOD_SE + MOD_FV + MOD_JI+I_FEM*I_SEPARA+I_FEM*I_VIUVO+I_FEM*I_SOLT+I_FEM*I_NPREN, data=dados, family=poisson)
summary(modelo5)
exp(confint.default(modelo5))
```

Especificação do modelo 6

```
modelo6<glm(FREQ~VHL_6_11+VHL_12_17+VHL_18_25+I_GASOL+I_DISEL+I_TRIC+I_ALCOOL+I_FIA + I_VOL + I_FOR + I_PEU + I_HON + I_REN + I_CIT + I_TOY + I_HYU + I_NIS + I_KIA + I_JAC + I_MIT +I_SEA+ MOD_MV + MOD_SW + MOD_PK + MOD_BR + MOD_SE +MOD_FV + MOD_JI+I_FEM*ID_18_23 + I_FEM*ID_24_29 + I_FEM*ID_36_41 + I_FEM*ID_42_47 + I_FEM*ID_48_53 + I_FEM*ID_54_59 + I_FEM*ID_60_65 + I_FEM*ID_66_71 + I_FEM*ID_72_77 + I_FEM*ID_.78, data=dados, family=poisson)
summary(modelo6)
exp(confint.default(modelo6))
```

Referências Bibliográficas

Andrews, D. F. e Herzberg, A. M. (1985) Data: A Collection of Problems from Many Fields for the Student and Research Worker, Springer Verlag, New York.

Akaike, H. (1974). A new look at statistical model identification. IEEE Transactions on Automatic Control AU-19 716-722.

Dobson, Anette.J. (2002). An Introduction to Generalized Linear Models. Second Edition. Chapman & Hall/CRC.

KPMG (2013). Situação Atual e Perspectiva do Mercado de Distribuição de Seguros no Brasil.

McCullagh, P. e Nelder, J.A. (1989). Generalized Linear Models. Second ed. London: Chapman and Hall.

Nelder, John A. e Wedderburn, Robert W (1972). "Generalized linear models". *Journal of the Royal Statistical Society*, Series A 135 (3): 370–384.

Turkman, M.A. e Silva, G. (2000). Modelos Lineares Generalizados – da Teoria à Prática, Edições SPE, Lisboa.

Websites auxiliares

http://data.princeton.edu/R/gettingStarted.html
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html
http://wiki.icmc.usp.br/images/e/e8/Glm-2012-ADC.pdf
http://www.leg.ufpr.br/Rpira/Rpira.pdf
http://www.stat.columbia.edu/~martin/W2024/R11.pdf
http://web.stanford.edu/class/stats306a/RforGLM.pdf
http://www.tudosobreseguros.org.br/sws/portal/pagina.php?c=1215#
http://www.theanalysisfactor.com/r-glm-model-fit/
http://www.icmc.usp.br/~ehlers/inf/cap2.pdf
http://stats.stackexchange.com/
http://www.leg.ufpr.br/~paulojus/CE210/ce210/node3.html