**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

Um aplicativo *shiny* para modelos lineares generalizados

Cayan Atreio Portela Bárcena Saavedra

Dissertação apresentada para obtenção do título de
Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

**Piracicaba
2018**

Cayan Atreio Portela Bárcena Saavedra
Bacharel em Estatística

Um aplicativo *shiny* para modelos lineares generalizados

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:

Prof. Dr. **CRISTIAN MARCELO VILLEGAS LOBOS**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2018

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Saavedra, Cayan Atreio Portela Bárcena

Um aplicativo *shiny* para modelos lineares generalizados / Cayan Atreio Portela Bárcena Saavedra. -- versão revisada de acordo com a resolução CoPGr 6018 de 2011. -- Piracicaba, 2018 .
52 p.

Dissertação (Mestrado) -- USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Modelos lineares generalizados
2. Interatividade
3. Gráficos dinâmicos
4. *shiny* . I. Título.

SUMÁRIO

Resumo	4
Abstract	5
Lista de Figuras	6
Lista de Tabelas	7
1 Introdução	9
1.1 Revisão de Literatura	10
1.1.1 Componentes dos modelos lineares generalizados	10
1.1.2 Estimação dos parâmetros	11
1.2 Diagnósticos	12
1.2.1 Resíduos de Pearson	13
1.2.2 <i>Deviance</i>	13
1.2.3 <i>Leverage</i>	13
1.2.4 Distância de Cook	14
1.2.5 Gráficos normal e semi-normal de probabilidades	14
1.3 Gráficos Dinâmicos	14
1.3.1 Dashboard	17
1.4 Pacote shiny	17
1.4.1 Estrutura do shiny	17
1.4.2 Exemplo de um aplicativo shiny	18
1.5 Pacote ggplot2	19
1.5.1 Exemplo de gráfico utilizando ggplot	19
1.6 Pacote plotly	20
1.6.1 Exemplo de gráfico interativo utilizando plotly	20
2 Material e Métodos	23
2.1 Dados utilizados	23
2.1.1 Dados salário	23
2.1.2 Dados rotenone	23
2.1.3 Dados de chegadas de clientes	24
2.2 Aplicativo shiny para modelos lineares generalizados	25
2.2.1 Introdução	26
2.2.2 Leitura de dados	26
2.2.3 Análise descritiva	27
2.2.4 Ajuste de modelos	30
2.2.5 Análise de diagnósticos	33
2.3 Ensino da estatística	34
3 Resultados	37
3.1 Dados salário	37
3.2 Dados rotenone	41
3.3 Dados de chegadas de clientes	45
4 Discussão	49
Referências	51

RESUMO

Um aplicativo *shiny* para modelos lineares generalizados

Recentes avanços tecnológicos e computacionais trouxeram alternativas que acarretaram em mudanças na forma com que se faz análises e visualizações de dados. Uma dessas mudanças caracteriza-se no uso de plataformas interativas e gráficos dinâmicos para a realização de tais análises. Desta maneira, análises e visualizações de dados não se limitam mais a um ambiente estático, de modo que, explorar a interatividade pode possibilitar um maior leque na investigação e apresentação dos dados. O presente trabalho tem como objetivo propor um aplicativo interativo, de fácil uso e interface amigável, que viabilize estudos, análises descritivas e ajustes de modelos lineares generalizados. Este aplicativo é feito utilizando o pacote shiny no ambiente R de computação estatística com a proposta de atuar como ferramenta de apoio para a pesquisa e ensino da estatística. Usuários sem afinidade em programação podem explorar os dados e realizar o ajuste de modelos lineares generalizados sem digitar uma linha código. Em relação ao ensino, a dinâmica e interatividade do aplicativo proporcionam ao aluno uma investigação descomplicada de métodos envolvidos, tornando mais fácil a assimilação de conceitos relacionados ao tema.

Palavras-chave: Modelos lineares generalizados; Interatividade; Gráficos dinâmicos; *shiny*.

ABSTRACT

A shiny app to perform generalized linear models

Recent technological and computational advances have brought alternatives that have led to changes in the way data analyzes and visualizations are done. One of these changes is characterized by the use of interactive platforms and dynamic graphics to carry out such analyzes. In this way, data analyzes and visualizations are no longer limited to a static environment, so exploring this dynamic interactivity can enable a wider range of data exploration and presentation. The present work aims to propose an interactive application, easy to use and with user-friendly interface, which enables studies and descriptive analysis and fit generalized linear models. This application is made using the shiny package in the R environment of statistical computing. The purpose of the application is to act as a support tool for statistical research and teaching. Users with no familiarity in programming can explore the data and perform the fit of generalized linear models without typing a single code line. Regarding teaching, the dynamics and interactivity of the application gives the student an uncomplicated way to investigate the methods involved, making it easier to assimilate concepts related to the subject.

Keywords: Generalized linear models; Interactivity; Dynamic graphics; *shiny*.

LISTA DE FIGURAS

1.1	Fluxograma sobre os três processos em análises de dados: transformar, visualizar e modelar (WICKHAM, 2011).	15
1.2	Aplicativo em Java, Javascript e Flash, desenvolvido por Allan J. e Beth L.	16
1.3	Exemplo de aplicativo shiny para criação de histogramas.	19
1.4	Gráfico de dispersão para duas variáveis do dataset iris destacando as espécies, utilizando o pacote ggplot2.	20
1.5	Gráfico de dispersão interativo utilizando o pacote plotly.	21
2.1	Página inicial do aplicativo.	26
2.2	Carregando os dados no aplicativo com o mouse.	27
2.3	Aplicativo com os dados “salarios” carregados.	28
2.4	Gráficos dispostos na aba “Análise Descritiva”.	28
2.5	Gráficos de dispersão com diferentes transformações para as variáveis Exp e Salário, feitos no aplicativo.	29
2.6	Medidas descritivas das variáveis do conjunto de dados “salarios”	30
2.7	Box-plot e gráfico de dispersão (com opção para gráfico de barras), na aba “Análise Descritiva”.	30
2.8	Histograma com 5, 10, 15 e 20 classes, na aba “Análise Descritiva”.	31
2.9	Exemplo de ajuste de modelo linear simples com diferentes funções de ligação utilizando o aplicativo.	33
2.10	Gráficos de diagnósticos, disponibilizados na aba “Resíduos”.	34
2.11	Gráficos de diagnósticos sem a observação #30, dispostos pelo aplicativo.	35
2.12	Dados fictícios propostos por ANSCOMBE (1973).	36
3.1	Análises descritivas dos dados de salários de executivos e executivas.	37
3.2	Gráfico de dispersão entre as variáveis salário e posição, por sexo.	37
3.3	Estimativas do modelo linear gaussiano para os dados de salário.	38
3.4	Estimativas do modelo linear gaussiano com interação para os dados de salário.	39
3.5	Estimativas do modelo final disponibilizadas na aba “Ajuste de Modelos”.	39
3.6	Gráficos de diagnósticos, disponibilizados na aba “Análise de diagnósticos”.	40
3.7	Estimativas obtidas incluindo (esquerda) e não incluindo (direita) a observação 30.	40
3.8	Estimativas para o modelo de proporção logito.	42
3.9	Curvas de ajuste para o modelo binomial por três funções de ligação.	42
3.10	Dose letal estimada a partir do modelo logito, que mata 50% e 90% dos insetos.	43
3.11	Dose letal estimada a partir do modelo probito, que mata 50% e 90% dos insetos.	43
3.12	Dose letal estimada a partir do modelo complemento log-log, que mata 50% e 90% dos insetos.	44
3.13	Envelope simulado para o modelo com função de ligação logito.	44
3.14	Envelope simulado para o modelo com função de ligação probito.	44
3.15	Envelope simulado para o modelo com função de ligação complemento log-log.	45
3.16	Gráficos de dispersão relacionando a variável de interesse “Número de Clientes”.	46
3.17	Estimativas obtidas a partir do modelo log-linear Poisson.	47
3.18	Resíduos obtidos a partir do ajuste do modelo log-linear Poisson.	47
3.19	À esquerda, uso da interatividade <i>brushing</i> para selecionar os pontos e à direita os pontos já selecionados.	48
3.20	Estimativas obtidas a partir do modelo log-linear Poisson com todas as observações.	48
3.21	Estimativas obtidas a partir do modelo log-linear Poisson sem as observações #20 e #43.	48

LISTA DE TABELAS

1.1	Algumas distribuições da família exponencial.	11
2.1	Proporção p_i do número y_i de insetos mortos entre n_i insetos que receberam a dose d_i de rotenone.	24
2.2	Dados de pesquisa realizada em uma loja, apresentados em NETER <i>et al.</i> (1996).	25
2.3	Dados fictícios propostos por ANSCOMBE (1973).	36

1 INTRODUÇÃO

A possibilidade de visualizar e explorar conjuntos de dados, de forma dinâmica e interativa, remete ao âmbito de um proveitoso recurso em análises de dados. Com o auxílio de novas ferramentas, que vêm sendo desenvolvidas continuamente, a análise, comunicação de resultados e *insights* podem ser providos de maneiras mais eficientes e criativas, facilitando o entendimento. A análise de modelos de regressão, por exemplo, é uma área que pode usufruir vantajosamente destes recursos.

Análises estatísticas que visam estudar relações e efeitos que certas covariáveis (independentes) possam vir a ter em uma variável de interesse (dependente), frequentemente remetem ao cenário em que se utilizam classes de modelos de regressão. Segundo CORDEIRO e DEMÉTRIO (2008), no caso mais rudimentar, abordando um modelo clássico de regressão linear, dentre outros pressupostos, assume-se distribuição normal, com média zero e variância constante para o erro aleatório. Todavia, nem sempre essa suposição é satisfeita, uma vez que a variável dependente pode pertencer a outra distribuição, pela própria natureza dos dados. Sendo assim, é necessária a busca por outros modelos para abordar tais problemas.

NELDER e WEDDERBURN (1972) propõem uma classe de modelos de regressão, uma teoria acoplando uma série de ferramentas estatísticas que eram estudadas separadamente. Tal teoria, denomina-se modelos lineares generalizados (MLG) (do inglês *generalized linear models*), sendo uma extensão dos modelos clássicos de regressão.

Essa generalização foi de grande ganho, pois permite uma maior flexibilização quanto à distribuição da variável dependente, de modo que o pressuposto de sua distribuição não se restringe apenas à distribuição normal, mas às distribuições pertencentes à família exponencial. Assim, o modelo de regressão linear clássico é apenas um caso específico dos modelos lineares generalizados.

O presente trabalho teve como objetivo o desenvolvimento de um aplicativo interativo que permitisse análises exploratórias além do estudo e ajuste de modelos lineares generalizados. Este aplicativo pretende contribuir com a democratização do conhecimento, sendo uma ferramenta de auxílio para o ensino de conceitos relacionados ao tema e desempenhando um papel de conector entre as informações contidas nos dados e resultados estatísticos, preenchendo uma possível lacuna referente à programação computacional. Todavia, também propõe-se a auxiliar usuários com conhecimento em programação, ao propiciar análises interativas simultâneas. Estas análises tendem a impulsionar o poder de investigação de dados e ajudar na percepção de relações múltiplas.

O aplicativo é desenvolvido com uso do pacote **shiny** (CHANG *et al.*, 2017), no ambiente R de computação estatística (R CORE TEAM, 2017), que se caracteriza como *software* livre de código aberto. O pacote shiny tem sido amplamente difundido entre a comunidade acadêmica e científica, por sua fácil utilização na criação de aplicativos interativos por usuários que possuem experiência de programação em R, explorando visualização de dados interativa (WOJCIECHOWSKI *et al.*, 2015). Deste modo, o aplicativo propõe-se a ser uma ferramenta de apoio para estudos de caso, visualização de dados e ensino de conceitos relacionados ao assunto.

Como mencionado anteriormente, os modelos lineares generalizados, propostos por NELDER e WEDDERBURN (1972), formam uma classe de modelos de regressão que ampliam as possíveis distribuições a serem assumidas pelo componente aleatório, em que se assumia apenas distribuição normal. A generalização, por sua vez, fornece suporte para aquelas pertencentes à família exponencial de distribuições. O interesse em assumir diferentes distribuições para a variável resposta vem da própria natureza dos dados, pois, nas mais diversas áreas, as variáveis de interesse dos estudos podem estar definidas em diferentes espaços amostrais.

O presente trabalho está organizado de forma que, nesta mesma seção 1, se destacam as vantagens e usos de ferramentas como o aplicativo aqui proposto, descrevendo brevemente os principais

recursos utilizados. Na seção 2, são descritos os conjuntos de dados analisados e o aplicativo é apresentado detalhadamente. Na seção 3, são apresentados os resultados das análises desenvolvidas com o uso da plataforma proposta. Na seção 4, são destacados considerações finais e projetos para trabalhos futuros.

1.1 Revisão de Literatura

1.1.1 Componentes dos modelos lineares generalizados

Segundo CORDEIRO e DEMÉTRIO (2008), os modelos lineares generalizados envolvem três componentes que podem ser descritos como componente aleatório, componente sistemático e função de ligação. O componente aleatório é representado por um vetor de variáveis aleatórias independentes. O componente sistemático configura-se como uma soma linear de efeitos, que carregam as informações das covariáveis. A função de ligação relaciona a média do componente aleatório ao preditor linear. Aqui, tais conceitos são descritos mais detalhadamente.

1. Componente aleatório

Vetor de variáveis aleatórias independentes $\mathbf{Y} = (y_1, \dots, y_n)^\top$ sendo cada uma delas com distribuição pertencente à mesma família exponencial, com média μ_i e parâmetro de dispersão $\phi > 0$, e cuja função densidade de probabilidade possa ser escrita como:

$$f(y_i; \theta_i, \phi) = \exp \{ \phi^{-1} [y_i \theta_i - b(\theta_i)] + c(y_i, \phi) \},$$

em que θ_i é o parâmetro canônico; $b(\cdot)$ e $c(\cdot)$ são funções conhecidas. Assim, denota-se $Y_i \sim \text{FE}(\mu_i, \phi)$, com $i = 1, 2, \dots, n$. A média e variância de Y_i são dadas, respectivamente, por:

$$E(Y_i) = b'(\theta_i) = \mu_i$$

$$\text{Var}(Y_i) = \phi b''(\theta_i) = \phi V(\mu_i),$$

em que $b'(\theta_i) = \frac{db(\theta_i)}{d\theta_i}$ e $b''(\theta_i) = \frac{d^2b(\theta_i)}{d\theta_i^2}$ e $V(\mu_i)$ é a função de variância, que depende unicamente de μ_i . Na Tabela 1.1, encontram-se algumas distribuições pertencentes à família exponencial, explicitando seus componentes.

2. Componente sistemático

Preditor linear, composto pelas variáveis explicativas, que entram na forma de uma soma linear de efeitos. Assim, tem-se que:

$$\eta_i = \sum_{r=1}^p x_{ir} \beta_r = x_{i1} \beta_1 + \dots + x_{ip} \beta_p = \mathbf{x}_i^\top \boldsymbol{\beta},$$

em que $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ é um vetor de dimensão $p \times 1$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ é o vetor de parâmetros desconhecidos a serem estimados, de dimensão $p \times 1$; e $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top$ é o preditor linear, de dimensão $n \times 1$.

3. Função de ligação

Função que relaciona a média de Y_i ao preditor linear.

$$g(\mu_i) = \eta_i,$$

em que $g(\cdot)$ é uma função monótona e diferenciável. Como $E(Y_i) = \mu_i$, temos que a função $g(\cdot)$ estabelece uma relação entre a média e as variáveis explicativas do modelo.

Uma vez definidos os componentes a serem ajustados em um MLG, a estimação de parâmetros se dá de maneira única, qualificando-se como uma vantagem no uso de tais modelos.

Tabela 1.1. Algumas distribuições da família exponencial.

Distribuição	ϕ	θ	$b(\theta)$	$c(y, \phi)$
Normal	σ^2	μ	$\frac{\theta^2}{2}$	$-\frac{1}{2} \left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right]$
Poisson	1	$\log(\mu)$	e^θ	$-\log(y!)$
Binomial	1	$\log(\frac{\mu}{m-\mu})$	$m\log(1 + e^\theta)$	$\log\binom{m}{y}$
Gama	ν^{-1}	$-\frac{1}{\mu}$	$-\log(-\theta)$	$\nu\log(\nu y) - \log(y) - \log\Gamma(\nu)$
Normal Inversa	σ^2	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$	$-\frac{1}{2} [\log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2} y]$

Fonte: CORDEIRO e DEMÉTRIO (2008).

1.1.2 Estimação dos parâmetros

O método mais usual para estimação dos parâmetros em modelos lineares generalizados é o da máxima verossimilhança, por possuir propriedades ótimas assintoticamente e serem consistentes (CORDEIRO e DEMÉTRIO, 2008). As estimativas de máxima verossimilhança dos parâmetros de um modelo linear generalizado são obtidas ao igualar o vetor escore a zero e resolver as equações. O logaritmo da função de verossimilhança de uma distribuição pertencente à família exponencial é dado por:

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

O vetor escore $U_\beta = \frac{\partial l(\beta)}{\partial \beta}$ de dimensão p , pode ser calculado pela regra da cadeia, cujo r -ésimo elemento é dado por:

$$U_{\beta_r} = \frac{\partial l(\beta)}{\partial \beta_r} = \sum_{i=1}^n \frac{dl(\beta)}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_r}, \quad r = 1, \dots, p.$$

Nota-se que as derivadas parciais dos parâmetros presentes no algoritmo se dá pelas seguintes relações:

1. $l(\beta) = f(\theta_1, \dots, \theta_i, \dots, \theta_n)$
2. $\theta_i = \int V_i^{-1} d\mu_i = q(\mu_i)$
3. $\mu_i = g^{-1}(\eta_i)$
4. $\eta_i = \sum_{r=1}^p x_{ir} \beta_r$

Pelas relações $\mu_i = b'(\theta_i)$ e $\frac{d\mu_i}{d\theta_i} = V_i = b''(\theta_i)$, temos que:

$$U_{\beta_r} = \phi^{-1} \sum_{i=1}^n (y_i - \mu_i) \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} x_{ir}, \quad r = 1, \dots, p.$$

Todavia, as equações encontradas ao fazer $U_{\beta_r} = 0$ são não lineares em β , fazendo-se necessário o uso de métodos numéricos para serem resolvidas. Uma alternativa é o uso do algoritmo de Newton-Raphson, que para o caso univariado, é dado por:

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})}, \quad m = 0, 1, 2, \dots$$

em que $x^{(m)}$ corresponde ao valor de x no passo m , $f(x^{(m)})$ é a função avaliada em $x^{(m)}$ e $f'(x^{(m)})$ é a derivada da função avaliada em $x^{(m)}$.

Utilizando o processo acima em sua forma multivariada, a estimativa de máxima verossimilhança de β é obtida ao expandir, em série de Taylor de 2ª ordem, a função escore U_β em torno de um valor inicial $\beta^{(0)}$ (PAULA, 2004), de maneira que:

$$U_\beta \cong U_\beta^{(0)} + U_\beta'^{(0)}(\beta - \beta^{(0)}),$$

em que U_β' é a primeira derivada de U com respeito a $\beta^\top$ e $U_\beta'^{(0)}$ e $U_\beta^{(0)}$ são os respectivos valores aplicados em $\beta^{(0)}$. À vista disso, o processo iterativo para obter a estimativa de máxima verossimilhança do vetor β é dado por:

$$\beta^{(m+1)} = \beta^{(m)} + [(-U_\beta')^{-1}]^{(m)} U_\beta^{(m)}, \quad m = 0, 1, 2, \dots$$

Segundo PAULA (2004), aplicar o método escore de Fisher, ao substituir a matriz $-U_\beta'$ (matriz de informação observada) pela matriz $K_{\beta\beta}$ (matriz de informação esperada) pode ser mais conveniente, pois a matriz $-U_\beta'$ pode não ser positiva definida. Desta maneira, o processo iterativo fica:

$$\beta^{(m+1)} = \beta^{(m)} + [K_{\beta\beta}^{-1}]^{(m)} U_\beta^{(m)}, \quad m = 0, 1, 2, \dots$$

em que $K_{\beta\beta}$ possui elementos dados por

$$k_{r,s} = E \left[-\frac{\partial^2 l(\beta)}{\partial \beta_r \partial \beta_s} \right] = E \left[\frac{\partial l(\beta)}{\partial \beta_r} \frac{\partial l(\beta)}{\partial \beta_s} \right].$$

Com algumas manipulações algébricas em CORDEIRO e DEMÉTRIO (2008), mostra-se que a solução das estimativas de máxima verossimilhança equivale a um processo de mínimos quadrados iterativos reponderados, sobre uma variável dependente ajustada z . O algoritmo é dado por:

$$\beta^{(m+1)} = (X^\top W^{(m)} X)^{-1} X^\top W^{(m)} z^{(m)}, \quad m = 0, 1, 2, \dots$$

em que $W = \text{diag}[w_1, \dots, w_n]$, com $w_i = V_i^{-1} \left(\frac{d\mu_i}{d\eta_i} \right)^2$; e z é um vetor $n \times 1$ com i -ésimo elemento dado por $z_i^{(m)} = \eta_i^{(m)} + (y_i - \mu_i^{(m)}) g'(\mu_i^{(m)})$. Em geral, estipula-se um erro representado por uma medida de diferença entre estimativas de passos consecutivos como critério de convergência e também é usual iniciar o algoritmo com $\eta^{(0)} = g(y)$. Nota-se que o algoritmo iterativo não depende de ϕ , ou seja, o parâmetro de dispersão não interfere nas estimativas pontuais do vetor β .

1.2 Diagnósticos

A definição dos componentes descritos em 1.1.1 representa uma das principais etapas na escolha de um modelo linear generalizado. Todavia, mesmo que esta etapa tenha sido realizada de maneira cautelosa, é possível que os resultados gerados após o ajuste sejam insatisfatórios. Razões para que isso ocorra podem envolver pontos discrepantes ou desvios sistemáticos entre observações (CORDEIRO e DEMÉTRIO, 2008). Então, outra etapa importante se dá na análise de adequabilidade do modelo e de possíveis afastamentos das suposições citadas anteriormente, em especial, para os componentes aleatório e sistemático e na observação de pontos influentes e discrepantes (PAULA, 2004). Para isto, são realizados diagnósticos a partir dos resíduos do modelo. Resíduos são medidas de discrepância entre a observação y_i e seu valor ajustado $\hat{\mu}_i$. Para verificar a qualidade do ajuste podem ser utilizadas ferramentas visuais e testes formais.

De maneira geral, diferentes tipos de resíduos (e suas análises) trazem diferentes informações. Em modelos lineares generalizados, usualmente se faz necessário o uso algumas funções ou escalonamentos para que os resíduos tenham forma similar aos de modelos de regressão clássico. Isto se deve ao fato de que a maioria dos MLGs não possuem variância constante (FARAWAY, 2016). Aqui serão descritos alguns tipos de resíduos e medidas de diagnóstico utilizados neste trabalho.

1.2.1 Resíduos de Pearson

O resíduo de Pearson é dado pela diferença entre o valor observado e ajustado, padronizado pela função de variância, tal que:

$$r_{pi} = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)^{1/2}},$$

em que $V(\hat{\mu}_i)$ é a função de variância sob o modelo ajustado. Segundo CORDEIRO e DEMÉTRIO (2008), este resíduo está relacionado com a estatística de Pearson generalizada X_p^2 , que é dada por:

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

e representa uma medida de discrepância de ajuste de um modelo.

1.2.2 Deviance

Medidas de informação que expressam a importância de uma variável são muito importantes para seleção de modelos. No processo de ajuste de um modelo, também deve-se considerar qual o número de termos necessários para uma representação razoável do processo gerador dos dados. Segundo CORDEIRO e DEMÉTRIO (2008), um número grande de variáveis explicativas pode levar a um modelo muito complexo, enquanto um número pequeno pode acarretar em um ajuste pobre. Desta maneira, deve-se ater ao princípio de parcimônia. Basicamente, em modelos com o mesmo poder explicativo, este princípio indica à escolha do modelo com menos parâmetros.

Desta maneira se faz necessária uma medida comparativa para modelos que possuem mais parâmetros. A função *deviance* proposta por NELDER e WEDDERBURN (1972), representa uma medida de qualidade de ajuste entre dois modelos, um com mais parâmetros (modelo completo) e outro com menos parâmetros (modelo reduzido). Assim, é possível obter uma noção de contribuição marginal desses parâmetros extras. A função *deviance* é dada por:

$$S_p = 2(\hat{l}_{p1} - \hat{l}_{p2}),$$

em que $\hat{l}_{p1}$ e $\hat{l}_{p2}$ representam o máximo do logaritmo da função de verossimilhança para os modelos com mais parâmetros e menos parâmetros, respectivamente. O modelo com mais parâmetros pode ser caracterizado pelo modelo saturado e o modelo com menos parâmetros sendo o modelo corrente. À medida que modelos mais complexos são ajustados, o valor da *deviance* decresce, como o resíduo das somas de quadrados e no caso em que $\hat{\mu}_i = y_i$ a deviance se reduz à zero (McCULLAGH, 1984).

Uma indicativa de ajuste também é fornecida ao comparar o modelo sob pesquisa com o modelo saturado, que corresponde ao ajuste de n parâmetros especificados pelas médias $\mu_1, \dots, \mu_n$, ou seja, um modelo que apenas reproduz os dados.

1.2.3 Leverage

Para os modelos lineares generalizados, a matriz de projeção $\mathbf{H}$ depende das variáveis explicativas, da função de ligação e da função de variância (CORDEIRO e DEMÉTRIO, 2008) e é dada por:

$$\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2},$$

em que $\mathbf{W}$ é a matriz de pesos e $\mathbf{X}$ a matriz de delineamento do modelo. A matriz $\mathbf{H}$ possui algumas propriedades, como:

- i) Idempotente: $\mathbf{H}^2 = \mathbf{H} \times \mathbf{H} = \mathbf{H}$,

- ii) Simétrica: $\mathbf{H} = \mathbf{H}^\top$,
- iii) $\text{tr}(\mathbf{H}) = \sum_{i=1}^n h_{ii} = p$ e $0 \leq h_{ii} \leq 1$.

O *leverage* é extraído a partir da diagonal de $\mathbf{H}$. De acordo com PAULA (2004), na suposição de que todos os pontos exerçam a mesma influência sobre as estimativas, espera-se que o valor de h_{ii} fique próximo de $\frac{p}{n}$. Assim, por exemplo, cabe a análise de pontos que apresentem valor de *leverage* $h_{ii} > \frac{2p}{n}$, podendo ser classificados como pontos de alavanca. Valores altos para h_{ii} indicam que o ajuste pode ser sensível à resposta no i -ésimo caso, e comumente, que aquele valor predito é de alguma forma incomum (FARAWAY, 2016).

1.2.4 Distância de Cook

A distância de Cook (COOK e WEISBERG, 1982) é uma medida que expressa a mudança nos parâmetros do modelo, caso aquela observação fosse excluída, revelando quais pontos são mais influentes na regressão (STEVENS, 1984). Segundo WILLIAMS (1987) uma medida equivalente para o modelo linear generalizado é $2p^{-1}[\mathbf{L}(\hat{\beta}) - \mathbf{L}(\hat{\beta}_{(i)})]$ em que $\mathbf{L}(\hat{\beta})$ é a função de verossimilhança calculada com o vetor de parâmetros estimados $\hat{\beta}$, que pode ser aproximada por:

$$C_i = p^{-1}h_{ii}(1 - h_{ii})^{-1}r_{pi}^2,$$

em que h_{ii} é o valor da observação na diagonal da matriz $\mathbf{H}$ e r_{pi} o resíduo de Pearson. Desta maneira, a distância de Cook é uma medida importante para detecção de pontos influentes. É indicado a análise de pontos que apresentem valores substancialmente maiores que os demais, em que um dos critérios sugeridos é $C_i > 4/n$, em que n é o número de observações (BOLLEN e JACKMAN, 1985).

1.2.5 Gráficos normal e semi-normal de probabilidades

De acordo com MORAL *et al.* (2017), no ajuste de modelos lineares generalizados, uma maneira de se observar a adequabilidade do modelo é observar o comportamento indicado em um gráfico, construído a partir do valor absoluto dos resíduos (*deviance*, Pearson, *studentizados*, entre outros) do modelo contra os valores esperados ordenados de uma distribuição *half*-normal, que pode ser aproximada por:

$$\Phi^{-1} \left(\frac{i + n - 1/8}{2n + 1/2} \right),$$

em que i é a i -ésima estatística de ordem. Para o gráfico normal são utilizados os valores ordenados ajustados contra os valores ordenados esperados de uma distribuição normal padrão, aproximado por:

$$\Phi^{-1} \left(\frac{i - 3/8}{n + 1/4} \right).$$

No gráfico mencionado, espera-se que os pontos se comportem conforme uma bissetriz. ATKINSON (1985) propõe o acréscimo de um envelope simulado no gráfico, com objetivo de mitigar a subjetividade na interpretação do comportamento dos pontos. A fuga deste comportamento desejado pode indicar falhas na especificação da distribuição do componente aleatório ou função de ligação, inadequabilidades no preditor linear etc. Desta maneira, para um modelo bem ajustado, os pontos são propensos a se apresentarem dentro do envelope (MORAL *et al.*, 2017).

1.3 Gráficos Dinâmicos

Recentes avanços tecnológicos, automatizações e esta era de grandes informações tem sido responsáveis por diversas mudanças em análises e visualizações de dados. Uma dessas mudanças, caracteriza-

se no uso de plataformas interativas e dinâmicas para tais análises. Segundo POTTER (2013), o crescimento no uso de gráficos dinâmicos e interativos baseados em *web pages*, demonstra haver um interesse significativo na criação dessas plataformas. Um dos motivos para tais crescimentos, é a maneira em que são feitas visualizações e análises, que não se limitam mais a um ambiente estático, possibilitando um maior leque na maneira de apresentação e investigação dos dados. Com respeito à análise de dados, um ponto importante ao utilizar abordagens e ferramentas do tipo é destacado por WICKHAM (2011), apontando o motivo pelo qual são feitas visualizações mais sofisticadas: não se trata de criar gráficos elegantes e bonitos, mas sim, em entender melhor os dados. Segundo WICKHAM (2011), o processo de exploração e análise de dados consiste na arte de observar os dados, gerar hipóteses e testá-las, repetidamente. Neste processo, que consiste em “gerar respostas”, a visualização é parte fundamental. Na Figura 1.1, está representado o fluxograma descrito por WICKHAM (2011) sobre os três processos na análise de dados: transformar (seleção de variáveis, medidas agregadas e sumarizadas, filtros, etc), visualizar (gráficos e análises visuais) e modelar (ajuste de modelos e teste de hipóteses).

Figura 1.1. Fluxograma sobre os três processos em análises de dados: transformar, visualizar e modelar (WICKHAM, 2011).

Ferramentas dinâmicas, responsivas e interativas ampliam o leque de possibilidades na ciência de dados, oferecendo opções que trazem ganho em diversos aspectos. DA SILVA *et al.* (2017), utilizam ggplot2, plotly e shiny para diagnosticar e explorar múltiplas árvores de decisão. Os autores apontam que análises visuais a partir de gráficos interativos podem proporcionar *insights* em estruturas de classes de alta dimensão. Esta interatividade contribui de maneira expressiva, pois proporciona ao analista uma maior eficiência em observar relações multivariadas e realizar diagnósticos de modelos (DA SILVA *et al.*, 2017). O aplicativo web para explorar modelos *random forest* é criado utilizando shiny (CHANG *et al.*, 2017), e a interatividade é dada por *mouse-over* (passando o mouse por cima das informações), *mouse-click selection* (clcando e selecionando) e *brushing* (selecionando vários pontos).

KIST (2017) cria aplicativos shiny para o estudo de modelos logísticos, com o intuito de prover um melhor entendimento em como os coeficientes do modelo influenciam na probabilidade de uma resposta. O aplicativo é descrito detalhadamente, com fotos e funções/ações.

POTTER *et al.* (2016) abordam uma outra contribuição desta área dinâmica: o ganho no ensino, em específico, da Estatística. Segundo os autores, a computação pode elevar o rendimento do ensino da Estatística, ao ilustrar conceitos fundamentais, como aleatoriedade, amostra, estimação, variabilidade e convergência por meio de simulações e visualizações gráficas. Desta maneira, o avanço tecnológico citado anteriormente, também pode ter uma contribuição relevante na melhora da performance de alu-

nos. Aplicações feitas a partir de ferramentas como shiny, oferecem vantagens como interatividade, uso amigável, bom apelo visual e são acessíveis via web, de maneira que alunos podem utilizar tais aplicações para um melhor entendimento de conceitos da estatística (POTTER *et al.*, 2016). Esta perspectiva representa o componente motivador essencial para o desenvolvimento deste trabalho.

Segundo POTTER *et al.* (2016), apesar da grande quantidade de softwares disponíveis que atuam nesta área, achar uma plataforma que sirva bem para um propósito específico, não é uma tarefa fácil, tornando atrativo a criação de ferramentas próprias de ensino. Porém, isto pode demandar um conhecimento de linguagens como Java, Javascript, HTML, CSS e PHP, que por muitas vezes é custoso de se adquirir, gerando uma barreira na construção do instrumento. Uma alternativa viável, então, é o uso de ferramentas como shiny (CHANG *et al.*, 2017), que requer apenas o conhecimento do ambiente de programação estatística R (R CORE TEAM, 2017), que já é bastante difundido entre pesquisadores da área.

BIEHLER (1993) enfatiza que a tecnologia permite a visualização de conceitos estatísticos, demonstração de idéias abstratas. CHANCE *et al.* (2007) destacam a importância de utilizar estas ferramentas da melhor maneira possível, visando o aprendizado do aluno. Outro ponto importante, é a combinação da ferramenta para um objetivo específico, tendo em vista que diferentes ferramentas possuem diferentes objetivos. Os autores destacam o ganho para alunos em melhor explorar e visualizar idéias estatísticas e para professores, que podem melhor apresentá-las.

Allan J. Rossman e Beth L. Chance desenvolvem uma coleção de aplicativos (*applet collection*, desenvolvidos em Java, Javascript e Flash e disponível em <http://www.rossmanchance.com/applets/index.html>) para o propósito educacional, de diversos temas como análise descritiva, modelos de regressão, correlação, simulação etc. Um exemplo pode ser dado utilizando o aplicativo “Least Squares Regression”, em que é possível observar dinamicamente, como uma observação influencia na reta de regressão, como mostra a Figura 1.2. Neste caso, percebe-se que a deleção do ponto em destaque no gráfico à esquerda, acarreta em uma mudança significativa na inclinação da curva (gráfico a direita).

Figura 1.2. Aplicativo em Java, Javascript e Flash, desenvolvido por Allan J. e Beth L.

Softwares e ferramentas agregadas que viabilizam o trabalho com gráficos dinâmicos vêm sendo criadas, dando uma importante contribuição e continuidade ao avanço da análise e visualização de dados. O aplicativo aqui proposto é designado neste âmbito, em colaboração com o ensino e disseminação do conhecimento em análise de dados. O presente trabalho faz o uso de três importantes pacotes disponíveis no ambiente R de computação estatística para sua funcionalidade: shiny (CHANG *et al.*, 2017), ggplot2 (WICKHAM, 2016) e plotly (SIEVERT *et al.*, 2017).

1.3.1 Dashboard

Dashboard é um termo comum em visualização de dados, sendo a mais recente fronteira da visualização de dados. Consiste em um painel de informações (dispostas em gráficos ou tabelas) que permite uma visualização simultânea de diferentes métricas, como indicadores e estatísticas, visando facilitar a compreensão de informações. Esta ferramenta é muito comum em *Business Intelligence* em empresas e organizações, pois permite um monitoramento de maneira bem ilustrada, de indicadores necessários para tomada de decisão. Desta maneira, este conceito possui boa contribuição em análise de dados em geral, desde análises descritivas à análises de resíduos, por exemplo.

Este trabalho faz o uso dessa ferramenta por meio do ambiente **shiny** utilizando o pacote **shinydashboard**, que permite de maneira fácil e intuitiva, implementar um aplicativo shiny na idéia de um *dashboard*.

1.4 Pacote shiny

Shiny é um pacote no ambiente R de computação estatística que facilita a criação de aplicações web (aplicativos ou app) diretamente do R. Desenvolvido pela RStudio, teve sua primeira versão lançada em dezembro de 2012, sendo o resultado do trabalho de vários pesquisadores e colaboradores, em que destacam-se os autores Winston Chang, Joe Cheng, Joseph J. Allaire, Yihui Xie e Jonathan McPheson, que trabalham na RStudio, Inc (Boston, Massachusetts, EUA).

Como se trata de uma estrutura que permite a criação de *web apps* diretamente em R, shiny foi desenvolvido para estatísticos e cientistas de dados em geral, uma vez que não há necessidade prévia de conhecimentos em linguagens como HTML, JavaScript ou CSS. Deste modo, desenvolvedores tem a possibilidade de implementar todas as ferramentas estatísticas do R em um *app*, sem se preocupar com estruturas de linguagem de desenvolvimento web. Isto se deve ao fato de que o shiny faz uma espécie de conversão de linguagem, de R para HTML, ou seja, um código de um aplicativo shiny desenvolvido em R, é lido pela máquina como HTML.

A ferramenta trabalha, em geral, com funções do tipo `reactive{() }`, ou seja, funções que geram uma reação, dependente de *inputs* e *outputs*, permitindo uma interatividade nas análises. De forma geral, a função recebe um valor (*input*), reage, e retorna o resultado (*output*) de maneira automática.

Devido à sua facilidade de uso e compartilhamento, shiny tem sido bastante usado para análises interativas (exploratórias e inferenciais) e vem se difundindo cada vez mais entre as comunidades científicas e acadêmicas. No Brasil, há responsáveis pelo desenvolvimento de alguns projetos, como o professor da UFPR, Walmes Zeviani (<http://shiny.leg.ufpr.br/walmes/>) e o professor da ESALQ/USP, Cristian Villegas (<https://github.com/clobos>). Internacionalmente, também existem diversos sites/-materiais sobre shiny e Estatística, como por exemplo, o próprio site da companhia RStudio (<https://shiny.rstudio.com/>), *Art of Stat* dos autores Alan Agresti, Christine A. Franklin e Bernhard Klingenberg (<http://www.artofstat.com/>) e eBooks, como *Web Application Development with R Using Shiny* (BEELEY, 2016).

1.4.1 Estrutura do shiny

A estrutura de programação no pacote shiny, divide-se basicamente em dois componentes: *UI* (*User Interface*) e *Server*. Estes dois componentes são duas funções, que podem ser escritos em dois arquivos texto R (*scripts*) diferentes - por questões de organização de código - ou em um único *script*, declarando as duas funções. Desta maneira, estes dois componentes são responsáveis pelos seguintes quesitos:

i *UI (User Interface)* - responsável pelo *layout* do aplicativo, ou seja, a aparência, o que o usuário vê. Onde o usuário insere informações desejadas e resultados são mostrados.

i *Server* - responsável por toda a estatística computacional executada pelo R, em que todos os possíveis cenários do *app* estão pré-programados. As tarefas a serem executadas (análises, gráficos, etc) estão prontas para serem realizadas a partir da inserção de informações por parte do usuário.

A comunicação entre estes dois componentes se dá através de *inputs* e *outputs* que representam, respectivamente, informações inseridas pelo usuário e informações geradas pelo aplicativo. Deste modo, a interatividade se dá na transmissão de dados entre os dois componentes, mantendo um ciclo ativo de conexão.

1.4.2 Exemplo de um aplicativo shiny

Um aplicativo simples pode ser construído com apenas poucas linhas de código. Neste exemplo, o usuário define o número de valores aleatórios a serem gerados por uma distribuição normal padrão e a cor de um histograma construído a partir desses valores. Seguindo a linha do que foi dito anteriormente, neste caso, o *ui* é responsável por três componentes: o número de valores gerados (inserido pelo usuário), a cor do histograma gerado (definida pelo usuário) e o gráfico plotado (gerado automaticamente). O *server*, por sua vez, recebe as informações inseridas pelo usuário, e as utiliza como argumentos para gerar os valores e os gráficos, a partir das funções `rnorm()` e `hist()`.

```

1 # definindo objetos
2 cores <- c("white","blue","green","red","yellow")
3 names(cores) <- c("branco","azul","verde","vermelho","amarelo")
4
5 # User interface.
6 ui <- bootstrapPage(
7   numericInput('n', 'Numero de observacoes', value = 100),
8   selectInput('cor', 'Selecione a cor desejada' , choices = cores ),
9   plotOutput('plot')
10 )
11
12 # Server.
13 server <- function(input, output) {
14   output$plot <- renderPlot({
15     hist(rnorm(input$n) , col = input$cor , main = "" , xlab = "" , ylab = "Frequência")
16   })
17 }
18
19 # Roda o aplicativo.
20 shinyApp(ui = ui, server = server)

```

Figura 1.3. Exemplo de aplicativo shiny para criação de histogramas.

1.5 Pacote ggplot2

O pacote `ggplot2` (WICKHAM, 2016) para visualização de dados, é baseado na gramática dos gráficos, proposto por WILKINSON (2006). Esta gramática consiste na criação dos gráficos por camadas, composto por um conjunto de componentes independentes (WICKHAM, 2016). Segundo KAHLE e WICKHAM (2013), tais camadas são compostas por cinco componentes:

1. Conjunto de dados com mapeamentos estéticos.
2. Uma ou mais camadas, cada uma com objetos geométricos (“geom”), transformações estatísticas (“stat”) e um conjunto de dados com mapeamentos estéticos.
3. Escala para cada mapeamento estético (que pode ser gerado automaticamente).
4. Sistema de coordenadas.
5. Especificação de facetas, que consistem em pequenas partes de um todo.

Desta maneira, o pacote propicia uma maior flexibilidade na criação de gráficos e incentiva a criatividade do analista de dados.

1.5.1 Exemplo de gráfico utilizando ggplot

Um gráfico feito a partir do pacote `ggplot2`, é feito “por partes”. Desta maneira, as camadas (ou partes) do gráfico, são concatenadas utilizando o sinal “+”. Neste exemplo, é feito um gráfico de dispersão com as variáveis ‘Sepal.Length’ (Tamanho da Sépala) e ‘Petal.Length’ (Tamanho da Pétala) da base de dados *iris* (disponível no R). Os mapeamentos estéticos (eixo x e y) são dados no primeiro elemento (por meio do uso do argumento `aes()`), e o segundo elemento é responsável pelo modo como este gráfico será plotado, no caso, um gráfico de pontos.

Figura 1.4. Gráfico de dispersão para duas variáveis do dataset iris destacando as espécies, utilizando o pacote ggplot2.

```

1 library(ggplot2)
2 data(iris)
3 names(iris)[5] <- "Espécies"
4
5 ggplot( data = iris , aes( x = Sepal.Length ,
6                           y = Petal.Length ,
7                           colour = Espécies ) ) +
8   geom_point() +
9   ylab("Comprimento da Pétala") +
10  xlab("Comprimento da Sépala")

```

1.6 Pacote plotly

Desenvolvido por SIEVERT *et al.* (2017), o pacote **plotly** é uma interface de alto nível em R para a biblioteca gráfica de código aberto plotly.js, em JavaScript (SIEVERT *et al.*, 2017). O pacote proporciona uma interatividade de alta responsividade, pois com manipulações do usuário, o gráfico não é redesenhado, mas apenas alterado instantaneamente. Funções do pacote **plotly** automaticamente transformam objetos estáticos ggplot2 em interativo, ao extrair suas informações e alocando-as em um JavaScript Object Notation (JSON) (DA SILVA *et al.*, 2017).

Dentre diversos trabalhos desenvolvidos nesta esfera inovadora da interatividade, não foi encontrado um aplicativo para este propósito específico. Assim sendo, este trabalho usufrui da potente combinação dos pacotes **ggplot2**, **plotly** e **shiny** para a construção de uma plataforma *web* para modelos lineares generalizados.

1.6.1 Exemplo de gráfico interativo utilizando plotly

A interatividade de gráficos dinâmicos feitos a partir do **plotly**, pode ser realizada utilizando uma ferramenta de comunicação entre o **plotly** e objetos criados pelo **ggplot**, de extrema usabilidade.

Assim, uma vez criado um objeto `ggplot`, para torná-lo interativo, basta utilizar a função `ggplotly()` do pacote `plotly`, como mostra o exemplo a seguir.

```
1 library(plotly)
2 library(ggplot2)
3
4 data(iris)
5 names(iris)[5] <- "Espécies"
6 names(iris)[1] <- "Comprimento da Sépala"
7 names(iris)[3] <- "Comprimento da Pétala"
8
9 p <- ggplot( data = iris , aes( x = `Comprimento da Sépala` ,
10                               y = `Comprimento da Pétala` ,
11                               colour = Espécies) ) +
12   geom_point()
13
14 ggplotly(p)
```

Figura 1.5. Gráfico de dispersão interativo utilizando o pacote `plotly`.

2 MATERIAL E MÉTODOS

2.1 Dados utilizados

O presente trabalho irá analisar três diferentes bases de dados, utilizando as funcionalidades do aplicativo para realizar esta tarefa. Esses conjuntos de dados, bem como objetivos de cada estudo, são aqui descritos. A análise dos dados referentes ao salário de executivos será feita nesta seção, servindo como exemplo para uma demonstração geral das funcionalidades do aplicativo proposto. O estudo dos demais conjunto de dados é feito na seção 3.

2.1.1 Dados salário

Os dados são referentes ao salário anual (em milhares de dólares) de uma amostra aleatória de 220 executivos, em que 145 são homens e 75 são mulheres. Neste conjunto de dados, estão dispostas as seguintes informações: salário, sexo, anos de experiência no cargo e posição na empresa (1 a 9, quanto maior o valor, mais alta é a posição). Os dados em questão foram retirados do livro texto de Gilberto de Paula (disponível em <https://www.ime.usp.br/~giapaula/textoregressao.htm>) e estão disponível em seu endereço eletrônico, nomeados como “salary.dat”.

Nesta aplicação, o interesse se dá em analisar a variável “Salário” como dependente, ou seja, o estudo consiste em encontrar relações funcionais como, covariáveis preditoras, distribuição de probabilidade assumida pelo componente aleatório e função de ligação adequada, que melhor expliquem a variável de interesse. Assim, modelos lineares generalizados serão utilizados para o desenvolvimento deste caso. Mais especificamente, é ajustado um modelo linear com distribuição normal e função de ligação identidade, sendo feito inteiramente no aplicativo. Desta maneira, o modelo inicial é dado por:

$$\text{Salário}_i = \beta_0 + \beta_1 \text{Sexo}_i + \beta_2 \text{Exp}_i + \beta_3 \text{Posicao}_i + \epsilon_i$$

em que Salário_i representa o salário da i -ésima pessoa e $\epsilon_i \sim N(0, \sigma^2)$.

2.1.2 Dados rotenone

Os dados analisados nesta seção correspondem a um ensaio de toxicidade de rotenone (Martin, 1942) de um delineamento completamente casualizado. Os dados, apresentados na Tabela 2.1, foram retirados do livro de CORDEIRO e DEMÉTRIO (2008). Neste delineamento, doses de inseticida foram aplicadas a grupos com n_i insetos (*Macrosiphoniella sanborni*, pulgão do crisântemo) e o número de insetos mortos em cada grupo foi observado.

Segundo CORDEIRO e DEMÉTRIO (2008), experimentos deste tipo frequentemente possuem objetivo de identificar um nível de tolerância em relação às doses aplicadas, que pode ser estimado pela probabilidade de morte nas diferentes doses. Desta maneira, a probabilidade de sucesso (morte) é usualmente nula para baixos níveis de dose e praticamente 1 para doses muito altas. Neste experimento, o interesse do pesquisador se dava na identificação de doses letais que matam 50% ou 90% dos insetos.

Devido à natureza dos dados, para esta aplicação, é ajustado um modelo de proporção, em que se especifica distribuição binomial com probabilidade π_i de ocorrência do evento (morte do inseto) para o componente aleatório, ou seja, $Y_i \sim \text{Binomial}(n_i, \pi_i)$. Alguns exemplos de funções de ligação para distribuição binomial são:

- logito: $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$
- probito: $g(\pi_i) = \Phi^{-1}(\pi_i) = \eta_i$
- complemento log-log: $g(\pi_i) = \log[-\log(1 - \pi_i)] = \eta_i$.

Tabela 2.1. Proporção p_i do número y_i de insetos mortos entre n_i insetos que receberam a dose d_i de rotenone.

d_i	n_i	y_i	p_i
0,0	49	0	0,0
2,6	50	6	0,12
3,8	48	16	0,33
5,1	46	24	0,52
7,7	49	42	0,86
10,2	50	44	0,88

Fonte: CORDEIRO e DEMÉTRIO (2008).

Então, para estes três casos, a relação entre o preditor linear e a probabilidade de sucesso (morte) estimada para o i -ésimo inseto é dada pelas seguintes relações, em que d_j representa os níveis de dose:

- logito: $\mu_i = n_i \left(\frac{e^{\beta_0 + \beta_1 d_j}}{1 + e^{\beta_0 + \beta_1 d_j}} \right)$
- probito: $\mu_i = n_i \Phi(\beta_0 + \beta_1 d_j)$
- complemento log-log: $\mu_i = n_i \{1 - \exp[-\exp(\beta_0 + \beta_1 d_j)]\}$.

Desta maneira, as estimativas de doses efetivas que causam a morte de 100p% das observações (que é uma das perguntas a serem respondidas pelo modelo) são dadas pelas seguintes relações:

$$\hat{\theta}_p = \frac{F^{-1}(p) - \hat{\beta}_0}{\hat{\beta}_1} = g(\hat{\beta}_0, \hat{\beta}_1),$$

que para os modelos citados, são:

- logístico:

$$\text{logito}(\hat{\pi}) = \log \left(\frac{\hat{\pi}}{1 - \hat{\pi}} \right) = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} \left[\log \left(\frac{\hat{\pi}}{1 - \hat{\pi}} \right) - \hat{\beta}_0 \right],$$

- probito:

$$\text{probit}(\hat{\pi}) = \Phi^{-1}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} [\Phi^{-1}(\hat{\pi}) - \hat{\beta}_0],$$

- complemento log-log:

$$\log[-\log(1 - \hat{\pi})] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} \{\log[-\log(1 - \hat{\pi})] - \hat{\beta}_0\}.$$

2.1.3 Dados de chegadas de clientes

Este conjunto de dados, apresentados em NETER *et al.* (1996) e também utilizados em PAULA (2004) (disponível em seu endereço eletrônico, em <https://www.ime.usp.br/giapaula/textoregressao.htm>), refere-se à uma grande loja varejista de madeira e pintura.

Uma pesquisa interna, realizada durante 2 semanas de um período representativo, foi conduzida e obtiveram-se os endereços dos clientes. Estes endereços foram utilizados para identificar a região censitária ao qual um dado cliente reside. Ao final da pesquisa, foi considerado o número total de clientes que visitaram a loja, residentes em regiões censitárias que estivessem dentro de um raio de 10 milhas, bem como informações demográficas relevantes de tais regiões. Desta maneira, a região censitária caracteriza-se como unidade experimental. Assim, deseja-se estudar a relação do número de clientes esperados em cada região, com suas características. Tais características consistem em: número de domicílios (em milhares), renda média anual (em milhares de dólares), idade média dos domicílios (em anos), distância ao concorrente mais próximo (em milhas) e distância à própria loja (em milhas). O conjunto de dados consiste em 110 regiões censitárias.

Tabela 2.2. Dados de pesquisa realizada em uma loja, apresentados em NETER *et al.* (1996).

# Clientes	# Domicílios	Renda	Idade	Dist. Concorrente	Dist. Loja
9	606	41.393	3	3,04	6,32
6	641	23.635	18	1,95	8,89
28	505	55.475	27	6,54	2,05
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
6	817	54.429	47	1,90	9,90
4	268	34.022	54	1,20	9,51
6	519	52.850	43	2,92	8,62

O interesse se dá em estimar o número esperado de clientes de uma i -ésima região censitária. Denotando Y_i como o número de clientes da i -ésima região censitária que visitaram a loja, é natural supor que $Y_i \sim \text{Poisson}(\mu_i)$, pois trata-se de uma contagem. Com o componente aleatório assim definido, para esta aplicação, estima-se um modelo log-linear (que utiliza a função de ligação logarítmica), com estrutura sistemática representada da seguinte maneira:

$$\log(\mu_i) = \beta_0 + \beta_1 \text{Domicílio}_i + \beta_2 \text{Renda}_i + \beta_3 \text{Idade}_i + \beta_4 \text{Distância1}_i + \beta_5 \text{Distância2}_i.$$

2.2 Aplicativo shiny para modelos lineares generalizados

O aplicativo *web* em formato de *dashboard* (painel com apresentação visual de informações importantes) desenvolvido neste trabalho, conta com um ambiente organizado em 6 abas. O acesso para cada aba é indicado na barra lateral por um ícone específico e seus devidos nomes, quais sejam, “Introdução”, “Leitura de Dados”, “Análise Descritiva”, “Ajuste de Modelos”, “Análise de Diagnósticos” e “Livro”. As funcionalidades de cada aba, assim como o aplicativo em geral, serão aqui apresentados em imagens e explicados nesta seção.

A dinâmica e interação propiciada pelo utilitário proposto, vem com objetivo de tornar mais funcional a comunicação entre o usuário e as informações contidas nos dados. Esta interatividade se faz presente em três formas:

- i *mouse-over*: ao deslizar o ponteiro por cima do ponto, informações são exibidas.
- ii *mouse-click*: seleção de algum ponto ou filtragem de informação, a partir de um clique.
- iii *brushing*: seleção de um ou mais pontos ao traçar um retângulo com o ponteiro, de maneira que, os pontos dentro do retângulo são selecionados.

As abas “Leitura de Dados”, “Análise Descritiva”, “Ajuste de Modelos” e “Análise de Diagnósticos” possuem nomes auto-explicativos e são responsáveis pelo processo de leitura, análise e modelagem

dos dados. Estas funcionalidades serão descritas mais detalhadamente durante aplicações e ajustes de modelos lineares generalizados, que serão desenvolvidos inteiramente no aplicativo, servindo como um tutorial para o usuário.

2.2.1 Introdução

Conforme Figura 2.1, a aba “Introdução” funciona como uma *home-page*, com o título PPGEAA (Programa de Pós-Graduação em Estatística e Experimentação Agronômica). Nesta aba é explicado o propósito do aplicativo, autores responsáveis, objetivos específicos e é indicado o local onde foi desenvolvido, isto é, o departamento de Ciências Exatas (LCE) da Universidade de São Paulo em Piracicaba. Também conta com 4 imagens no inferior da página, correspondentes à Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Universidade de São Paulo, RStudio e shiny.

Figura 2.1. Página inicial do aplicativo.

Logo abaixo dos ícones correspondente às abas, encontram-se os nomes dos autores responsáveis com respectivas fotos, que também age como link para o currículo lattes de cada um, para divulgação e maior conhecimento dos responsáveis. O ícone correspondente à um envelope de carta, no canto superior direito da tela, fornece link para envio de *e-mail* direto ao endereço cayanportela@hotmail.com; endereço eletrônico pessoal do desenvolvedor.

2.2.2 Leitura de dados

Como mencionado anteriormente, a análise dos conjuntos de dados neste trabalho serão feitas inteiramente no aplicativo e cada etapa descrita detalhadamente. A título de ilustração, serão carregados os dados referentes aos salários de uma amostra executivos. Estes dados foram carregados a partir de um arquivo em extensão “txt”, porém, outras extensões com outros tipos de separadores também são viáveis, como arquivos “csv”, bastando para isso escolher a opção necessário na caixa de seleção “Separador”.

Para carregar o banco de dados, é necessário clicar em “Browse” que abrirá uma janela para indicar o local do arquivo, conforme mostra Figura 2.2.

Figura 2.2. Carregando os dados no aplicativo com o mouse.

Após carregados os dados, o usuário deve preencher as opções “Cabeçalho” e “stringAsFactors” de acordo com sua necessidade, que se referem ao fato de que a primeira linha seja o nome das variáveis (Cabeçalho) e a conversão de strings em fatores (stringAsFactor). Neste caso, ambas as opções são selecionadas, pois os dados utilizados possuem a primeira linha referente ao nome das variáveis e deseja-se que a variável “Sexo” seja interpretada como fator.

Ao carregar os dados, os valores são disponibilizados na tela, possibilitando ao usuário uma visualização tabular bem como ordenação crescente ou decrescente pela variável desejada (clitando na setas ao lado do nome da variável). Outro ponto importante é identificação da estrutura das variáveis dos dados, disponível no canto superior direito. Essa noção de estrutura é fundamental para conhecimento do tipo de variáveis e checar se os dados foram carregados corretamente. Neste caso, vimos que se trata de uma tabela com 220 observações e 4 variáveis, onde o nome e tipo de estrutura de cada variável é dado por: Salario (inteiro), Sexo (fator com dois níveis), Posicao (inteiro) e Exp (numérico), conforme pode-se observar na Figura 2.3.

2.2.3 Análise descritiva

O próximo passo se dá na investigação exploratória dos dados, que é realizada na aba “Análise Descritiva” permitindo uma facilidade em visualizações gráficas exploratórias. A aba em questão divide-se em quatro quadrantes, em que basicamente são referentes à medidas descritivas das variáveis e visualizações gráficas com diferentes flexibilizações, conforme Figura 2.4.

Durante o uso do aplicativo, os gráficos são dispostos simultaneamente em um *dashboard*. Essa visualização dinâmica e simultânea amplifica o poder de percepção de relações e investigação dos dados por parte do usuário. É disposto então, um exemplo da aba com os quatro quadrantes preenchidos a partir dos dados referente aos salários de executivos.

No quadrante superior esquerdo da Figura 2.4, o *input* é dado quando o usuário insere o texto referente à variável desejada nos eixos horizontais (X) ou verticais (Y) (nota-se que o texto inserido

Figura 2.3. Aplicativo com os dados “salarios” carregados.

Figura 2.4. Gráficos dispostos na aba “Análise Descritiva”.

deve ser idêntico ao nome da variável no conjunto de dados). Este tipo de *input* permite uma boa flexibilidade, possibilitando observar relações das variáveis em suas escalas originais ou realizando alguma transformação matemática na variável, como por exemplo, logaritmo, potência e divisão. A opção Grupo que corresponde a algum agrupamento desejado, é fornecida selecionando uma das opções disponíveis. As interações nos gráficos são realizadas por *mouse-over* e *mouse-click*.

A Figura 2.5 mostra dois gráficos com a variável Experiência (denotada por Exp) no eixo horizontal e a variável Salário no eixo vertical. Por exemplo, o gráfico posicionado em cima e à esquerda, é representado na escala original com os dois grupos, enquanto o gráfico posicionado em cima e à direita possui Exp^2 no eixo x e $\log(\text{Salário})$ no eixo y, e o grupo Masculino foi retirado. Analogamente, os demais gráficos possuem outras manipulações como exemplo.

Figura 2.5. Gráficos de dispersão com diferentes transformações para as variáveis Exp e Salário , feitos no aplicativo.

O quadrante superior direito da Figura 2.4 é responsável por fornecer medidas descritivas. Para variáveis quantitativas, são calculadas média, quartis, mínimo e máximo, enquanto para variáveis qualitativas (fatores) é calculada a frequência absoluta de cada nível do fator. As opções são dadas pelo nome das variáveis (colunas) e podem ser escolhidas múltiplas colunas simultaneamente.

Figura 2.6. Medidas descritivas das variáveis do conjunto de dados “salarios”

O quadrante inferior esquerdo da Figura 2.4 proporciona visualizações gráficas bastante similares ao que foi mostrado na Figura 2.5 e com a mesma interatividade. Porém, difere no tipo de *input* (uma vez que não é possível fazer manipulações nas variáveis) e no tipo de gráfico, disponibilizando três opções: gráficos de dispersão, gráficos de barras e *boxplots*, como mostra a Figura 2.7. Por fim, o quadrante inferior direito da Figura 2.4 realiza uma análise visual univariada a partir de um histograma. Muitas vezes deseja-se saber o comportamento de frequência de uma variável e analisar visualmente o histograma, que é um estimador para a densidade de probabilidade (SCOTT, 2015). Dado isso, torna-se importante tal visualização gráfica. No aplicativo, o usuário pode selecionar a variável e escolher o número de classes, selecionando o valor na barra disponível, como mostra a Figura 2.8. A seguir, exemplos de gráficos localizados na parte inferior do *dashboard*, o *boxplot* mencionado, gráfico de dispersão com as opções mostradas, e um histograma feito com a variável salário, com 5, 10, 15 e 20 classes.

Figura 2.7. Box-plot e gráfico de dispersão (com opção para gráfico de barras), na aba “Análise Descritiva”.

2.2.4 Ajuste de modelos

A seção “Ajuste de Modelos” é a responsável pelo ajuste de modelos lineares generalizados. Naturalmente, possui opções em que é necessário informar a variável resposta e sua distribuição de probabilidade assumida, a função de ligação e as variáveis pertencentes ao preditor linear.

Figura 2.8. Histograma com 5, 10, 15 e 20 classes, na aba “Análise Descritiva”.

O ajuste de modelos lineares generalizados pelo aplicativo proposto baseia-se na função `glm()` do *software R*, de tal modo que as opções disponíveis em cada *input* correspondem às opções dos respectivos argumentos da função. Assim, o argumento “Distribuições”, assume aquelas pertencentes à família exponencial citadas anteriormente, bem como estruturas de relação dadas por quase-verossimilhança (quase-binomial e quase-Poisson), utilizadas para dados com superdispersão.

Para a modelagem de dados de proporção (distribuição binomial) é necessário marcar o *checkbox* “Proporção Binomial” para que se declare duas variáveis como variável resposta: número de sucessos ocorridos e número total de indivíduos sujeitos ao evento. Isto se dá pelo fato de que a função `glm()` recebe dois vetores como variável resposta: o número de sucesso e o número de fracassos. Internamente, a função calcula a proporção para o ajuste de modelos. O aplicativo recebe o número de sucessos e o número total ao invés dos números de sucessos e fracassos, por acreditar ser a forma mais comum a ser coletada por pesquisadores, tornando seu uso menos custoso, já que em geral, as variáveis disponíveis já estarão nesse formato.

Esta aba possui cinco *inputs* principais, que são as listas de seleção descritas a seguir:

- i Variável resposta: seleciona a variável dependente do modelo.
- ii Variáveis preditoras: seleciona variáveis preditoras que entram em forma de efeito aditivo.
- iii Adicionar interação: seleciona variáveis preditoras que entram em forma de interação.
- iv Distribuição: seleciona a distribuição da variável resposta.
- v Função de ligação: seleciona a função de ligação entre o preditor linear e a variável resposta.

Além destas opções e do *checkbox* “Proporção Binomial” mencionado anteriormente, outra opção importante é o botão “Adicionar interação”. Como as variáveis inseridas na opção “Variáveis preditoras” entram apenas com efeito aditivo, caso o usuário deseje modelar interação entre alguma variável, basta clicar em “Adicionar interação”, o que irá gerar uma nova lista de seleções com as variáveis dos dados. As variáveis que forem selecionadas nesta nova lista, entrarão no modelo em forma de interação.

Uma vez inseridas todas as informações desejadas corretamente, basta clicar em “Rodar novo modelo” para ajustar o modelo. Feito isso, duas tabelas com resultados são fornecidas. A primeira se refere às estimativas dos parâmetros, erros padrão, estatísticas de teste e p-valores. A segunda conta com graus de liberdade, valores de deviances e resultados da análise de variância do modelo medindo a contribuição marginal da variável, ou seja, o efeito da variável dado que todas as outras estão no modelo, fornecendo graus de liberdade e valores de deviances.

Abaixo das tabelas de resultados, é disponibilizado um gráfico com a variável resposta no eixo Y e a primeira variável selecionada em “Variáveis Preditoras” no eixo X. Este gráfico conta com a opção “Função de Ligação” que permite múltiplas escolhas. Ao selecionar uma função de ligação, sua curva correspondente é plotada. Para problemas em que o preditor linear é composto por apenas uma variável, isto se torna bastante eficiente, permitindo uma análise visual confortável a respeito da relação entre os componentes aleatório e sistemático, por diferentes funções de ligação.

A título de exemplificação, a Figura 2.9 ilustra o ajuste de um modelo simples (apenas uma covariável) com estimativas e gráfico de função de ligação. Neste modelo, as variáveis dependentes e independentes são, salário e experiência, respectivamente. O modelo foi ajustado a partir da distribuição normal e função de ligação identidade.

Figura 2.9. Exemplo de ajuste de modelo linear simples com diferentes funções de ligação utilizando o aplicativo.

2.2.5 Análise de diagnósticos

Segundo MORAL *et al.* (2017), a avaliação gráfica para diagnóstico de modelos com suposição de normalidade pode ser mais satisfatória do que o uso de testes formais, como o teste para normalidade de resíduos Shapiro-Wilk (SHAPIRO e WILK, 1965) e o teste para homogeneidade de variâncias de Bartlett (BARTLETT, 1937), pois estes podem apresentar resultados insatisfatórios, como por exemplo, em casos com amostra pequena. As análises gráficas de resíduos e medidas de diagnóstico disponibilizadas nesta seção são:

- *Half-Normal plot*: para verificar o ajuste do modelo, utilizando o pacote hnp (MORAL *et al.*, 2017).
- Resíduos Padronizados vs Valores Ajustados: para verificação de homogeneidade e independência.
- Distância de Cook: para verificação de pontos influentes.
- Leverage: para verificação de pontos de alavanca.

Assim sendo, esta seção é fundamental para validação das análises feitas, sendo responsável pelo diagnóstico dos resíduos. Uma vez ajustado o modelo na aba anterior, os gráficos de resíduos são gerados automaticamente. Similar ao painel de análises descritivas, este campo divide-se em quatro quadrantes em que são apresentados quatro diferentes gráficos. Todavia, as análises visuais aqui disponibilizadas estão conectadas entre si, oferecendo uma maior interatividade na análise residual. Tal dinâmica tem capacidade de desempenhar um papel expressivo quanto ao ensino da Estatística, ao permitir a visualização simultânea de um mesmo ponto em diferentes gráficos. Esta dinâmica simultânea pode contribuir para o entendimento e interpretação dos resíduos dispostos.

Para ilustrar esta aba, a figura 2.10 demonstra os resíduos do modelo final a ser ajustado no capítulo 3, utilizando os dados “salario”. Percebe-se a facilidade propiciada pela interatividade *mouse-over* na identificação de pontos; ao observar um possível ponto de alavanca ou outlier, basta passar o mouse no ponto para obter seu valor residual e sua identificação em relação a base dados. Neste caso, nota-se que a observação 30 demonstra indícios de um possível ponto de alavanca, por apresentar uma medida de distância de Cook substancialmente maior que os demais.

Figura 2.10. Gráficos de diagnósticos, disponibilizados na aba “Resíduos”.

A seleção de uma observação em qualquer um dos gráficos, se destaca ao trocar sua cor e estilo. Este evidenciamento se caracteriza nas seguintes mudanças: os pontos mudam de preto preenchido para cinza não preenchido. Este realce dinâmico ocorre nos quatros gráficos simultaneamente, pois, como dito anteriormente, eles estão conectados. Esta interatividade promove uma maneira conveniente para identificação de pontos e valores residuais, apresentando-se como um método vantajoso tanto para o ensino e quanto para análises. A Figura 2.11 mostra o *dashboard* de resíduos após a seleção da observação 30 (identificada como possível ponto de alavanca).

Uma vez selecionado o ponto, esta informação fica salva temporariamente, e é construída uma nova base de dados na qual o ponto é deletado. Isto possibilita o ajuste de um novo modelo sem a observação selecionada, e por conseguinte, a averiguação de sua influência em estimativas e resultados. Em vista disso, uma vez que o ponto foi selecionado, basta voltar para a aba “Ajuste de Modelos” e clicar em “Rodar novo modelo” para obter os novos resultados.

O mesmo pode ser feito selecionando mais de um ponto, pelo uso da interatividade *brushing*, que consiste em traçar um retângulo com o ponteiro, de maneira que, os pontos dentro do retângulo serão selecionados.

2.3 Ensino da estatística

Um objetivo importante que motivou o desenvolvimento do aplicativo, refere-se à questão do ensino da estatística. Plataformas como esta proposta, com gráficos dinâmicos e análises interativas, permitem ao aluno, um envolvimento maior e de maneira mais fácil nas análises e estudos. Nesse sentido,

Figura 2.11. Gráficos de diagnósticos sem a observação #30, dispostos pelo aplicativo.

a facilidade proporcionada por estas ferramentas, permite uma investigação descomplicada de conceitos e análises em análises exploratórias e inferenciais por parte do aluno. Desta maneira o aluno torna-se protagonista do próprio aprendizado, por meio de benefícios obtidos que o ajudam a assimilar conceitos relativos ao tema.

Esta contribuição ao ensino, pode ser exemplificada utilizando os dados propostos por ANSCOMBE (1973), que utiliza dados fictícios para ilustrar a importância da visualização gráfica para o entendimento de relações estatísticas.

ANSCOMBE (1973) propõe quatro modelos de regressão simples. Três desses modelos compartilham os mesmos valores para a variável regressora (x) e diferem nos valores da variável dependente (y). O último modelo, difere em ambas as variáveis. Os dados estão dispostos na Tabela 2.3.

Mesmo com dados diferentes, os quatro modelos de regressão estimam o mesmo valor para os coeficientes da reta e demais estatísticas usuais, como soma de quadrados e erro quadrático médio, como pode ser visto em ANSCOMBE (1973). Na Figura 2.12, estão dispostas as quatro retas ajustadas utilizando os diferentes dados. Os dados acima foram carregados no aplicativo através de um arquivo com extensão .txt e os ajustes foram realizados inteiramente no aplicativo, na aba “Ajuste de Modelos”, em que é possível ajustar uma reta de regressão simples em um gráfico de dispersão, facilitando a visualização de cenários propostos por cada reta.

A curva ajustada no cenário (1), corresponde às expectativas de um modelo linear bem ajustado. Porém, nos outros quadros, é possível notar a necessidade de um olhar mais atento ao observar a relação entre as variáveis. No cenário (2), por exemplo, um efeito quadrático pode acarretar em um melhor ajuste. No cenário (3), um ponto apresenta-se substancialmente diferente em relação aos demais, indicando a necessidade de uma análise mais profunda naquele ponto. E no cenário (4), nota-se que uma única observação (que se apresenta completamente afastada das outras) é responsável pela inclinação da curva.

Com este exemplo didático, percebe-se a importância da análise visual para o entendimento de relações estatísticas e ajustes de modelos. Assim, o aplicativo também pretende contribuir como ferramenta instrutiva para casos como este, que são de grande valia para o ensino da estatística.

Tabela 2.3. Dados fictícios propostos por ANSCOMBE (1973).

X1.3	X4	Y1	Y2	Y3	Y4
10	8	8.04	9.14	7.46	6.58
8	8	6.95	8.14	6.77	5.76
13	8	7.58	8.74	12.74	7.71
9	8	8.81	8.77	7.11	8.84
11	8	8.33	9.26	7.81	8.47
14	8	9.96	8.10	8.84	7.04
6	8	7.24	6.13	6.08	5.25
4	19	4.26	3.10	5.39	12.50
12	8	10.84	9.13	8.15	5.56
7	8	4.82	7.26	6.42	7.91
5	8	5.68	4.74	5.73	6.89

Fonte: ANSCOMBE (1973).

Figura 2.12. Dados fictícios propostos por ANSCOMBE (1973).

3 RESULTADOS

3.1 Dados salário

O objetivo está na investigação da relação do salário dos executivos e executivas com as demais variáveis. Em um primeiro momento, para investigar o comportamento dessas relações, realiza-se uma análise exploratória associando a variável de interesse com as variáveis experiência, posição e sexo; um histograma do salário e um resumo dessas variáveis, como demonstram as Figuras 3.1 e 3.2.

Figura 3.1. Análises descritivas dos dados de salários de executivos e executivas.

Figura 3.2. Gráfico de dispersão entre as variáveis salário e posição, por sexo.

Nota-se uma relação linear positiva entre o tempo de experiência e a posição exercida na empresa com o salário, para ambos os sexos, ou seja, pessoas que ocupam cargos com maior posição e possuem

maior tempo de experiência, tendem a apresentar um salário mais elevado. Na amostra, estão presentes 75 executivas e 145 executivos. Mesmo parecendo não haver uma diferença acentuada entre o salário de homens e mulheres, a mediana do salário dos executivos é superior à mediana do salário das executivas. Porém o maior salário na amostra pertence a uma pessoa do sexo feminino.

Ajustar-se-á inicialmente um modelo linear gaussiano com função de ligação identidade, com a variável “Salário” como dependente e as demais (Experiência, Posição e Sexo) como variáveis explicativas de efeito aditivo. Desta maneira, temos que o modelo será dado por:

$$\text{Salário}_i = \beta_0 + \beta_1 \text{Sexo}_i + \beta_2 \text{Exp}_i + \beta_3 \text{Posicao}_i + \epsilon_i,$$

em que Salário_i representa o salário da i -ésima pessoa e $\epsilon_i \sim N(0, \sigma^2)$. Utilizando o aplicativo para ajustar o modelo proposto, observa-se na Figura 3.3 os resultados obtidos de estimativas, erros-padrão e testes de significância, bem como as opções selecionadas em cada argumento.

Figura 3.3. Estimativas do modelo linear gaussiano para os dados de salário.

Ao incluir uma interação (assim como variáveis isoladas), é útil verificar sua contribuição marginal na tabela de análise de variância. Uma interação com p-valor menor do que o nível de significância estipulado leva à conclusão de que possui contribuição significativa e portanto, permanece no modelo. Para que isto seja feito no aplicativo, deve-se clicar em “Adicionar Interação” para adicionar uma interação. Caso deseje-se adicionar duas ou mais interações, basta clicar no botão novamente e novas opções de *input* serão criadas. Desta maneira, adicionando as três possíveis interações, nota-se pela Figura 3.4 que a interação entre experiência e posição apresenta valor significativo (ao nível de 5%) e consequentemente deve ser incluída. Pela tabela de análise de variância na Figura 3.4, nota-se que apenas a interação entre a posição ocupada na empresa e o tempo de experiência do executivo ou da executiva, demonstra possuir uma contribuição significativa para o modelo. Sendo assim, esta será a única interação a permanecer no modelo.

Então, os resultados do modelo final ajustado é disposto na Figura 3.5.

Figura 3.4. Estimativas do modelo linear gaussiano com interação para os dados de salário.

Figura 3.5. Estimativas do modelo final disponibilizadas na aba “Ajuste de Modelos”.

Desta forma, o modelo final é dado por:

$$\text{Salário}_i = \beta_0 + \beta_1 \text{Sexo}_i + \beta_2 \text{Exp}_i + \beta_3 \text{Posição}_i + \beta_4 \text{Posição} * \text{Experiência} + \epsilon_i,$$

em que $\epsilon_i \sim N(0, \sigma^2)$. Como feito anteriormente, verifica-se a adequabilidade do ajuste da distribuição normal pelo componente aleatório, de modo a validar o modelo proposto. Em conjunto, uma análise residual visual é favorável para uma boa interpretação da qualidade do ajustes e dos valores preditos.

Na Figura 3.6 pode-se observar que a observação 30 apresenta um distância de Cook mais elevada que as demais observações. Ajustando então, outro modelo retirando esta observação, pode-se comparar as estimativas obtidas com a do modelo que a inclui:

Figura 3.6. Gráficos de diagnósticos, disponibilizados na aba “Análise de diagnósticos”.

Estimativa dos Parâmetros

	Estimativa	Erro Padrão	T-valor	P-valor
(Intercept)	108.04	2.96	36.48	0.00
Posicao	8.10	0.59	13.73	0.00
SexoMasculino	-2.81	1.09	-2.59	0.01
Exp	0.34	0.31	1.07	0.29
Posicao:Exp	-0.13	0.05	-2.76	0.01

Estimativa dos Parâmetros

	Estimativa	Erro Padrão	T-valor	P-valor
(Intercept)	109.37	3.04	36.04	0.00
Posicao	7.86	0.60	13.07	0.00
SexoMasculino	-2.50	1.09	-2.28	0.02
Exp	0.21	0.32	0.64	0.52
Posicao:Exp	-0.12	0.05	-2.33	0.02

Figura 3.7. Estimativas obtidas incluindo (esquerda) e não incluindo (direita) a observação 30.

Mesmo não havendo mudanças em relação às hipóteses testadas, observa-se que os pontos causam uma variação notável nas estimativas do modelo. Assim, recomenda-se um olhar mais atento à observações deste tipo. Trata-se pois, de um executivo com salário anual de \$110 mil, com posição 2 e aproximadamente 2 anos e meio de experiência, sendo o menor salário dentre os executivos analisados. Todavia, não se trata de um caso aberrante e sim de um possível cenário a ser observado quando ao analisar outros executivos. Decidindo-se manter todas as observações na estimação do modelo, chega-se ao modelo final:

$$\text{Salário}_i = 108,04 + 0,34\text{Experiência}_i - 2,81 * \text{Sexo}_i + 8,1\text{Posição}_i - 0,13\text{Posição} * \text{Experiência}.$$

Considerando todas as variáveis, os executivos demonstram receber um salário médio inferior ao salário médio das executivas. Verifica-se que executivos e executivas com posição alta, tem propensão a receber um salário menor caso tenha mais experiência, ou seja, estagnar em um mesmo cargo por muito tempo, pode não ser tão frutífero quanto ficar menos tempo no mesmo cargo.

Assim, o aplicativo demonstra-se eficiente ao possibilitar ajustes de modelos lineares generalizados e realizar inferências de estimativas com a facilidade da interatividade em uma interface amigável.

3.2 Dados rotenone

O conjunto de dados consiste em observar o número de insetos mortos pelo nível de dose. Pela natureza do experimento, é comum admitir distribuição binomial com probabilidade p_i de ocorrência do evento (morte do inseto), ou seja, $Y_i \sim \text{Binomial}(n_i, \pi_i)$. Ensaios do tipo dose-resposta como este apresentado, comumente deparam-se com o propósito de obter a curva sigmoidal que se ajuste bem aos dados, para então, calcularem-se as doses letais pretendidas. Ao assumir distribuição binomial para o componente aleatório, geram-se diferentes possibilidades de função de ligação. Consequentemente, diferentes funções de ligação geram diferentes curvas, pois a média μ_i é funcionalmente relacionada ao preditor linear (CORDEIRO e DEMÉTRIO, 2008), ou seja,

$$g\left(\frac{\mu_i}{m_i}\right) = g(\pi_i) = \eta_i = X_i^T \beta$$

Alguns exemplos de funções de ligação para distribuição binomial são:

- logito: $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$
- probito: $g(\pi_i) = \Phi^{-1}(\pi_i) = \eta_i$
- complemento log-log: $g(\pi_i) = \log[-\log(1 - \pi_i)] = \eta_i$.

Então, para estes três casos, a relação entre o preditor linear e a probabilidade de sucesso (morte) estimada para o i -ésimo inseto é dada por:

- logito: $\mu_i = n_i \left(\frac{e^{\beta_0 + \beta_1 d_i}}{1 + e^{\beta_0 + \beta_1 d_i}} \right)$
- probito: $\mu_i = n_i \Phi(\beta_0 + \beta_1 d_i)$
- complemento log-log: $\mu_i = n_i \{1 - \exp[-\exp(\beta_0 + \beta_1 d_i)]\}$.

Os dados foram carregados de acordo com o que foi mostrado na seção 2.2.2. Para o ajuste do modelo é necessário marcar o *checkbox* “Proporção binomial” (na aba “Ajuste de modelos”) e declarar as variáveis correspondentes ao número de sucesso (insetos mortos) e número total de insetos expostos, necessariamente nesta ordem. Feito isso, é selecionado a variável “Dose” como preditora. Utilizando a função de ligação logito para um ajuste inicial, os resultados são dispostos, demonstrando haver um efeito positivo de Dose, ou seja, quanto maior a dose, maior a probabilidade de morte do inseto. Após o ajuste utilizando a função de ligação logito, as estimativas dos parâmetros foram $\hat{\beta}_0 = -3.3225$ e $\hat{\beta}_1 = 0.6051$, ambos com nível de significância menor que 5%.

A comparação simultânea de diferentes funções de ligação de maneira visual, pode prover um melhor entendimento sobre suas relações, tornando mais fácil sua compreensão. Isto pode ser atingido pelo aplicativo, ao selecionar mais de uma opção no argumento “Funções de Ligação”, abaixo das tabelas de resultados. Na figura 3.9, são dispostas curvas ajustadas com as três funções de ligação mencionadas: logito, probito e complemento log-log.

Como dito anteriormente, o interesse do pesquisador no experimento estava na identificação de doses letais, que matassem 50% e 90% dos insetos. CORDEIRO e DEMÉTRIO (2008) demonstram que a estimativa da dose efetiva θ_p que causa a mortalidade de 100p% dos indivíduos é dada por:

$$\hat{\theta}_p = \frac{F^{-1}(p) - \hat{\beta}_0}{\hat{\beta}_1} = g(\hat{\beta}_0, \hat{\beta}_1),$$

que para os modelos citados, são:

Figura 3.8. Estimativas para o modelo de proporção logito.

Figura 3.9. Curvas de ajuste para o modelo binomial por três funções de ligação.

– logístico:

$$\text{logito}(\hat{\pi}) = \log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} \left[\log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) - \hat{\beta}_0 \right],$$

– probito:

$$\text{probit}(\hat{\pi}) = \Phi^{-1}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} [\Phi^{-1}(\hat{\pi}) - \hat{\beta}_0],$$

– complemento log-log:

$$\log[-\log(1 - \hat{\pi})] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} \{\log[-\log(1 - \hat{\pi})] - \hat{\beta}_0\}.$$

Utilizando as estimativas fornecidas pelo modelo logístico (Figura 3.8), chega-se às doses letais de 50% (D_{50}) e 90% (D_{90}) da seguinte maneira:

$$D_{50} = \frac{1}{0.6051} \left[\log \left(\frac{0.5}{1-0.5} \right) - (-3.2256) \right] = 5.330,$$

$$D_{90} = \frac{1}{0.6051} \left[\log \left(\frac{0.9}{1-0.9} \right) - (-3.2256) \right] = 8.961.$$

O mesmo pode ser feito usufruindo-se da interatividade *mouse-over* no gráfico do modelo ajustado. No gráfico o eixo Y é dado pela proporção de insetos mortos e o eixo X pelo nível de dose. Desta maneira, basta arrastar o ponteiro para a proporção desejada (altura do eixo Y) e observar o nível de dose informado, como mostra a Figura 3.10.

Figura 3.10. Dose letal estimada a partir do modelo logito, que mata 50% e 90% dos insetos.

De maneira análoga, podem-se obter as estimativas para os modelos ajustados com as funções de ligação probito e complemento log-log, pelo cálculo e analisando o gráfico interativo.

– probito:

$$D_{50} = \frac{1}{0.346} [\Phi^{-1}(0.5) - (-1.878)] = 5.42$$

$$D_{90} = \frac{1}{0.346} [\Phi^{-1}(0.9) - (-1.878)] = 9.13$$

Figura 3.11. Dose letal estimada a partir do modelo probito, que mata 50% e 90% dos insetos.

– complemento log-log:

$$D_{50} = \frac{1}{0.345} \{ \log[-\log(1-0.5)] - (-2.397) \} = 5.88$$

$$D_{90} = \frac{1}{0.345} \{ \log[-\log(1-0.9)] - (-2.397) \} = 9.36$$

Figura 3.12. Dose letal estimada a partir do modelo complemento log-log, que mata 50% e 90% dos insetos.

Na escolha do modelo a ser utilizado para responder às questões propostas, é importante analisar os resíduos de cada ajuste. Uma maneira de verificar a adequabilidade dos modelos propostos, é utilizar um envelope simulado a partir de uma distribuição *half-normal* com os valores de *deviance* absolutos. Verificando os modelos ajustados para as funções de ligação logito (Figura 3.13), probit (Figura 3.14) e complemento log-log (Figura 3.15):

Figura 3.13. Envelope simulado para o modelo com função de ligação logito.

Figura 3.14. Envelope simulado para o modelo com função de ligação probito.

Mesmo com poucos pontos, nota-se um melhor ajuste para o modelo que utiliza a função de ligação logito para relacionar o preditor linear e a proporção média de insetos, pois apresenta apenas um ponto fora do envelope simulado.

Figura 3.15. Envelope simulado para o modelo com função de ligação complemento log-log.

3.3 Dados de chegadas de clientes

Dado o objetivo do estudo, que é relacionar as variáveis de regiões censitárias com o número de clientes esperados na loja, sugere-se uma análise exploratória para observar a relação destas variáveis com a variável de interesse (número de clientes). Relacionando-a com as variáveis, Renda, Idade, Distância Concorrente e Distância Loja (denotadas por “Renda”, “Idade”, “Dist_con” e “Dist_loja”, respectivamente) nota-se um grau de relação mais acentuada com as duas distâncias, indicando uma relação positiva com a distância ao concorrente e negativa com a distância à loja. A variável renda parece demonstrar uma leve relação positiva com o número de clientes, enquanto a idade aparenta não ter relação, como demonstram os gráficos presentes na Figura 3.16.

Seja Y_i o número de clientes da i -ésima região censitária que visitaram a loja dentro do período avaliado. Tratando-se de uma contagem, é natural supor que $Y_i \sim \text{Poisson}(\mu_i)$. Com o componente aleatório definido, estima-se um modelo com log-linear (o qual utiliza a função de ligação logarítmica), com estrutura sistemática representada da seguinte maneira:

$$\log(\mu_i) = \beta_0 + \beta_1 \text{Domicílio}_i + \beta_2 \text{Renda}_i + \beta_3 \text{Idade}_i + \beta_4 \text{Distância1}_i + \beta_5 \text{Distância2}_i$$

Na Figura 3.17, é possível analisar os resultados obtidos a partir das estimativas do modelo especificado. Também é demonstrado as opções selecionadas no aplicativo, para o ajuste do modelo em questão. Nota-se que todas as variáveis incluídas no modelo possuem contribuição marginal significativa. Visto isso, conclui-se que todas as variáveis devem ser mantidas no modelo.

Analisando o sinal das estimativas, é possível inferir que o número esperado de clientes que visitam a loja, tende a diminuir com o aumento da renda, idade e distância das regiões censitárias. Nesta mesma linha, quanto mais domicílios médios possuir a região e quanto mais longe a região estiver do concorrente, espera-se ter um efeito positivo no número de clientes visitando a loja.

Pela aba dos resíduos do aplicativo é possível estudar os diagnósticos do modelo e verificar possíveis afastamentos dos pressupostos do modelo, bem como a presença de pontos aberrantes e/ou influentes. Pelo gráfico de envelope simulado (presente na Figura 3.18) pode-se concluir que as suposições da distribuição assumida para o componente aleatório não estão sendo violadas, mesmo havendo um ponto sobreposto ao envelope. Dado que a confiança da simulação do envelope é a baseada em 95%, um envelope que não englobe exatamente todas as observações, não é sinônimo de um ajuste pobre.

Visualmente, as observações #20 e #43 destacam-se por apresentar valores de distância de Cook relativamente altos. A distância de Cook passa a idéia do quão influente nas estimativas tal observação

Figura 3.16. Gráficos de dispersão relacionando a variável de interesse “Número de Clientes”.

possa vir a ser. É interessante então, ajustar um modelo sem estas duas observações. Para isso, se faz útil a interatividade *brushing* disponível na análise visual dos resíduos. Como mostra a Figura 3.19, é possível selecionar dois (ou mais) maiores valores no gráfico da distância de Cook.

Após selecionados os pontos, é ajustado um segundo modelo que não os inclui. Desta forma, pode-se comparar as estimativas, como mostram as Figuras 3.20 e 3.21.

As estimativas dos parâmetros não apresentam alterações bruscas. A significância dos parâmetros e contribuições marginais das variáveis não são alteradas, a um nível de 5%. Desta maneira, as conclusões não se alteram.

Variável resposta
N_Clientes

Distribuição
poisson

☒ Proporção Binomial
☒ Interação

Variáveis preditoras
N_Domiclios Renda Idade Dist_con
Dist_loja

Função de Ligação
log

Rodar novo modelo

Adicionar Interacao

Estimativa dos Parâmetros

	Estimativa	Erro Padrão	T-valor	P-valor
(Intercept)	2.942	0.207	14.198	0.000
N_Domiclios	0.001	0.000	4.262	0.000
Renda	-0.000	0.000	-5.534	0.000
Idade	-0.004	0.002	-2.091	0.037
Dist_con	0.168	0.026	6.534	0.000
Dist_loja	-0.129	0.016	-7.948	0.000

Tabela de Análise de Variância

	LR Chisq	Df	Pr(>Chisq)
N_Domiclios	18.20	1.00	0.00
Renda	31.79	1.00	0.00
Idade	4.38	1.00	0.04
Dist_con	41.66	1.00	0.00
Dist_loja	67.50	1.00	0.00

Figura 3.17. Estimativas obtidas a partir do modelo log-linear Poisson.

Figura 3.18. Resíduos obtidos a partir do ajuste do modelo log-linear Poisson.

Figura 3.19. À esquerda, uso da interatividade *brushing* para selecionar os pontos e à direita os pontos já selecionados.

Estimativa dos Parâmetros

	Estimativa	Erro Padrão	T-valor	P-valor
(Intercept)	2.942	0.207	14.198	0.000
N_Domiclios	0.001	0.000	4.262	0.000
Renda	-0.000	0.000	-5.534	0.000
Idade	-0.004	0.002	-2.091	0.037
Dist_con	0.168	0.026	6.534	0.000
Dist_loja	-0.129	0.016	-7.948	0.000

Tabela de Análise de Variância

	LR Chisq	Df	Pr(>Chisq)
N_Domiclios	18.20	1.00	0.00
Renda	31.79	1.00	0.00
Idade	4.38	1.00	0.04
Dist_con	41.66	1.00	0.00
Dist_loja	67.50	1.00	0.00

Figura 3.20. Estimativas obtidas a partir do modelo log-linear Poisson com todas as observações.

Estimativa dos Parâmetros

	Estimativa	Erro Padrão	T-valor	P-valor
(Intercept)	3.003	0.209	14.369	0.000
N_Domiclios	0.001	0.000	4.253	0.000
Renda	-0.000	0.000	-5.803	0.000
Idade	-0.005	0.002	-2.763	0.006
Dist_con	0.178	0.026	6.824	0.000
Dist_loja	-0.131	0.017	-7.939	0.000

Tabela de Análise de Variância

	LR Chisq	Df	Pr(>Chisq)
N_Domiclios	18.16	1.00	0.00
Renda	35.19	1.00	0.00
Idade	7.65	1.00	0.01
Dist_con	45.47	1.00	0.00
Dist_loja	67.06	1.00	0.00

Figura 3.21. Estimativas obtidas a partir do modelo log-linear Poisson sem as observações #20 e #43.

4 DISCUSSÃO

Neste trabalho foi desenvolvido um aplicativo para modelos lineares generalizados. O objetivo fundamental almejado é que o aplicativo ajude alunos, professores e pesquisadores em geral, tanto no ensino em sala de aula, quanto ao proporcionar uma interface amigável e de fácil manuseio para análises exploratórias e ajustes de modelos lineares generalizados.

Com divulgação, hospedagem na *web* (que está por vir) e de livre acesso à todos, espera-se que de fato, o aplicativo contribua de algum modo com a comunidade acadêmica de interesse

O desenvolvimento do aplicativo não para com a entrega deste trabalho, visto que críticas e sugestões sempre virão à tona, e as melhorias a serem feitas são incessantes. Assim, seu desenvolvimento é contínuo, como os mais diversos *softwares* que possuem suas versões desde 1.0 adiante, sempre atualizando e melhorando o produto fornecido.

Para trabalhos futuros, pode-se ampliar os métodos de análise disponíveis, como por exemplo, modelos lineares generalizados mistos e modelos com abordagem bayesiana.

REFERÊNCIAS

- ANSCOMBE, F. J., 1973 Graphs in statistical analysis. *The American Statistician* **27**: 17–21.
- ATKINSON, A., 1985 Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis. Technical report, Clarendon Press.
- BARTLETT, M. S., 1937 Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A* **160**: 268–282.
- BEELEY, C., 2016 *Web application development with R using Shiny*. Packt Publishing Ltd.
- BIEHLER, R., 1993 Software tools and mathematics education: The case of statistics. *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES* **121**: 68–68.
- BOLLEN, K. A. e R. W. JACKMAN, 1985 Regression diagnostics: An expository treatment of outliers and influential cases. *Sociological Methods & Research* **13**: 510–542.
- CHANCE, B., D. BEN-ZVI, J. GARFIELD, e E. MEDINA, 2007 The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education* **1**.
- CHANG, W., J. CHENG, J. ALLAIRE, Y. XIE, e J. MCPHERSON, 2017 *shiny: Web Application Framework for R*. R package version 1.0.5.
- COOK, R. D. e S. WEISBERG, 1982 *Residuals and influence in regression*. New York: Chapman e Hall.
- CORDEIRO, G. M. e C. DEMÉTRIO, 2008 Modelos lineares generalizados e extensões. São Paulo .
- DA SILVA, N., D. COOK, e E.-K. LEE, 2017 Interactive graphics for visually diagnosing forest classifiers in r. arXiv preprint arXiv:1704.02502 .
- FARAWAY, J. J., 2016 *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, volume 124. CRC press.
- KAHLE, D. e H. WICKHAM, 2013 ggmap: Spatial visualization with ggplot2. *R Journal* **5**.
- KIST, M. J., 2017 *Logistic Regression Slope Study*. Ph.D. thesis, Rochester Institute of Technology.
- MCCULLAGH, P., 1984 Generalized linear models. *European Journal of Operational Research* **16**: 285–292.
- MORAL, R., J. HINDE, e C. DEMÉTRIO, 2017 Half-normal plots and overdispersed models in r: The hnp package. *Journal of Statistical Software, Articles* **81**: 1–23.
- NELDER, J. e R. WEDDERBURN, 1972 Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* **135**: 370–384.
- NETER, J., M. H. KUTNER, C. J. NACHTSHEIM, e W. WASSERMAN, 1996 *Applied linear statistical models*, volume 4. Irwin Chicago.
- PAULA, G. A., 2004 *Modelos de regressão: com apoio computacional*. IME-USP São Paulo.
- POTTER, G., J. WONG, I. ALCARAZ, P. CHI, *et al.*, 2016 Web application teaching tools for statistics using r and shiny. *Technology Innovations in Statistics Education* **9**.
- POTTER, S., 2013 *Dynamic, Interactive and Reactive Statistical Graphics for the Web*. Ph.D. thesis, ResearchSpace@ Auckland.

- R CORE TEAM, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- SCOTT, D. W., 2015 *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.
- SHAPIRO, S. S. e M. B. WILK, 1965 An analysis of variance test for normality (complete samples). *Biometrika* **52**: 591–611.
- SIEVERT, C., C. PARMER, T. HOCKING, S. CHAMBERLAIN, K. RAM, M. CORVELLEC, e P. DESPOUY, 2017 *plotly: Create Interactive Web Graphics via 'plotly.js'*. R package version 4.7.1.
- STEVENS, J. P., 1984 Outliers and influential data points in regression analysis. *Psychological Bulletin* **95**: 334.
- WICKHAM, H., 2011 ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* **3**: 180—185.
- WICKHAM, H., 2016 *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- WILKINSON, L., 2006 *The grammar of graphics*. Springer Science & Business Media.
- WILLIAMS, D., 1987 Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* pp. 181–191.
- WOJCIECHOWSKI, J., A. HOPKINS, e R. UPTON, 2015 Interactive pharmacometric applications using R and the shiny package. *CPT: pharmacometrics & systems pharmacology* **4**: 146–159.