

4 INTRODUÇÃO AOS MODELOS LINEARES GENERALIZADOS - MLGs

Nos capítulos iniciais, foi apresentado o modelo *clássico* de análise de regressão supondo que a variável resposta, ou melhor, o erro do modelo é uma variável aleatória com distribuição normal com média zero e variância constante. Como consequência desta suposição, a variável resposta também apresenta uma distribuição normal com média condicionada ao valor da variável preditora e variância constante. Entretanto, nem sempre esta suposição pode ser aplicada, ou adaptada à realidade do problema.

Por exemplo, sejam consideradas as seguintes situações:

1. Deseja-se prever as chances de sobrevivência de um paciente quando submetido a uma cirurgia a partir de dados clínicos do pré-operatório do paciente.
2. Deseja-se prever o número esperado de pessoas acometidas pela Dengue em um municípios a partir de informações sócio-econômicas do município.
3. Deseja-se estimar o valor de venda de um imóvel a partir de suas características intrínsecas e de localização.

No primeiro caso, a variável resposta observada é categórica e pode assumir dois valores possíveis: (S) o paciente sobreviveu e (N) o paciente não sobreviveu. Numericamente, define-se a variável aleatória $Y = 1$ se o paciente sobreviveu ou $Y = 0$ se o paciente morreu. Obviamente, a variável resposta não apresenta uma distribuição normal, nem pode ser transformada, pois mesmo em virtude de uma transformação a variável irá assumir dois possíveis valores. Neste caso, a distribuição de probabilidade conveniente para a variável resposta de interesse é a distribuição de *Bernoulli*.

No segundo caso, a variável resposta observada é o número de pessoas acometidas pela Dengue em um município. Por definição, o valor mínimo possível para esta variável é zero e o valor máximo é o número total da habitantes do município. Ou seja, no cenário otimista, nenhuma pessoa apresenta a doença e, no cenário pessimista, todas as pessoas do município apresentam a doença. Por definição, esta variável apresenta um comportamento de uma distribuição *Binomial*(N, p) onde N é a população do município e p é a probabilidade de um indivíduo contrair a doença.

No terceiro caso, a variável resposta é contínua e positiva. Ou seja, não existe nenhum imóvel com preço negativo (alguém oferecendo pagamento para se ver livre do bem). Portanto, com exceção da restrição de ser estritamente positiva, a variável aparenta ter similaridade com a distribuição normal. Entretanto, valores de imóveis são extremamente heterocedásticos, ou seja, a variabilidade do preço de um imóvel não é constante e está vinculada ao valor médio. Do ponto de vista prático, quanto maior o valor do imóvel, maior a faixa de negociação do bem e vice versa: quanto menor o valor do imóvel, menor a faixa de negociação do bem. Portanto, a característica da variável resposta, além de ser estritamente positiva, é possuir uma variância que depende da média. Considerando essas características da variável resposta, uma possível distribuição de

probabilidade é a distribuição *Gama*. Ou, neste caso, seria possível aplicar uma transformação logarítmica com o objetivo de tornar a dispersão da variável resposta homocedástica (constante) e permitir valores negativos e positivos.

Em todas as situações indicadas e em várias situações reais, a suposição de normalidade do erro ou da variável resposta e o uso do modelo de regressão *clássico* pode gerar resultados inconsistentes, mesmo que medidas de ajuste como o coeficiente de determinação do modelo (R^2) seja elevado. Uma escolha adequada da distribuição de probabilidade e, consequentemente, das suposições do comportamento probabilístico da variável resposta tornam o ajuste do modelo coerente e com interpretações relevantes e pertinentes.

Neste capítulo, iremos tratar dos principais modelos de regressão vinculados a uma classe de distribuições de probabilidade conhecida como *família exponencial*. Não é objetivo exaurir este amplo assunto, mas apresentar as principais aplicações desses modelos e suas interpretações.

4.1 A FAMÍLIA EXPONENCIAL

Uma função de probabilidade $f_Y(y)$, ou densidade de probabilidade, é dita pertencer à família exponencial se puder ser escrita na forma:

$$f_Y(y | \theta, \phi) = \exp \left\{ \frac{[y\theta - b(\theta)]}{a(\phi)} + c(y, \phi) \right\} \quad (4.1)$$

onde $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções conhecidas; $\phi > 0$ é o parâmetro de dispersão e θ é parâmetro canônico. Vamos supor, por exemplo a distribuição normal, $Y \sim Normal(\mu; \sigma^2)$. Neste caso, a densidade de probabilidade da distribuição normal pode ser escrita na forma da família exponencial como:

$$\begin{aligned} f_Y(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2}(y - \mu)^2 \right\} \\ &= (2\pi\sigma^2)^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2}(y - \mu)^2 \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} + \ln(2\pi\sigma^2)^{-1/2} \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2) - \frac{1}{2}\ln(2\pi\sigma^2) \right\} \\ &= \exp \left\{ \frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2} \left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2) \right) \right\} \end{aligned} \quad (4.2)$$

(4.3)

Em geral, é possível associar o parâmetro θ à média (e/ou à variância) da variável resposta. O parâmetro ϕ está associado, unicamente, à variância da resposta. Portanto, a partir da decomposição apresentação na equação 4.3 é possível identificar que: $\theta = \mu$, $\phi = \sigma^2$ e $a(\phi) = \phi$. Como consequência: $b(\theta) = \frac{\theta^2}{2}$ e $c(y, \phi) = -\frac{1}{2} \left\{ \frac{y^2}{\sigma^2} + \ln(2\pi\phi) \right\}$.

A maneira mais simples de inserir uma equação de regressão linear simples, $\eta = \beta_0 + \beta_1 x$, na família exponencial consiste em vincular o parâmetro canônico à família exponencial: $\theta = \eta = \beta_0 + \beta_1 x$.

4.2 PROPRIEDADES DA FAMÍLIA EXPONENCIAL

Sob certas condições de regularidade, descritas em McCullagh e Nelder (1989) - Apêndice A, a derivada da log-verossimilhança satisfaz as seguintes identidades:

1. $E(Y) = b'(\theta)$

$$2. \text{ } Var(Y) = a(\phi) \cdot b''(\theta)$$

Aplicando as propriedades 1 e 2 no exemplo da distribuição normal, $Y \sim Normal(\mu; \sigma^2)$:

1. $E(Y) = b'(\theta) = \frac{2\theta^{2-1}}{2} = \theta = \mu$
2. $Var(Y) = a(\phi) \cdot b''(\theta) = \sigma^2 \cdot \theta^{1-1} = \sigma^2$

Portanto, no caso de distribuições pertencentes à família exponencial, os parâmetros de média, $E(Y)$, e variância, $Var(Y)$, podem ser obtidos utilizando as propriedades 1 e 2; desde que os elementos da família exponencial θ , ϕ , $b(\theta)$ e $a(\phi)$ sejam corretamente identificados.

4.3 ALGORITMO DE ESTIMAÇÃO PARA A FAMÍLIA EXPONENCIAL, SUPONDO LIGAÇÃO CANÔNICA

Será demonstrado o algoritmo de estimação utilizando a ligação canônica. O caso genérico pode ser encontrado em McCullagh e Nelder (1989). O objetivo do algoritmo é maximizar a função de verossimilhança ou, de forma equivalente, o logarítmico da função de verossimilhança. Por se tratar de um modelo não-linear, o algoritmo é iterativo.

Seja a função log-verossimilhança da família exponencial definida como

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

Vamos definir $\eta_i = \mathbf{x}_i \beta$ e $\theta_i = \eta_i = \mathbf{x}_i \beta$ (ligação canônica). Definem-se também a função escore, $\mathbf{U}(\beta)$, e a matriz de informação de Fisher, $\mathbf{K}(\beta)$:

a. Função Escore, $\mathbf{U}(\beta)$:

$$\begin{aligned} \mathbf{U}(\beta) &= \frac{\partial l(\beta)}{\partial \beta} = \frac{1}{a(\phi)} \sum_{i=1}^n \{y_i - b'(\theta_i)\} \frac{\partial \theta_i}{\partial \beta} \\ &= a(\phi)^{-1} \sum_{i=1}^n \mathbf{x}_i^T (y_i - \mu_i) \\ &= a(\phi)^{-1} \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu}) \end{aligned}$$

b. Matriz de Informação de Fisher, $\mathbf{K}(\beta)$:

$$\begin{aligned} \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} &= a(\phi)^{-1} \cdot \frac{\partial}{\partial \beta^T} \left[\sum_{i=1}^n \mathbf{x}_i^T (y_i - b(\theta_i)) \right] \\ &= a(\phi)^{-1} \cdot \left[\sum_{i=1}^n \mathbf{x}_i^T (-b'(\theta_i)) \mathbf{x}_i \right] \\ &= -a(\phi)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} \end{aligned} \tag{4.4}$$

onde $[\mathbf{W}]_{ii} = b''(\theta_i)$.

$$\mathbf{K}(\beta) = E \left\{ -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} \right\} = a(\phi)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X}$$

4.3.1 O ALGORITMO DE ESTIMAÇÃO

Em sua forma geral, o algoritmo de estimação pode ser escrito na forma:

$$\beta^{(k+1)} = \beta^{(k)} + [\mathbf{K}^{(k)}]^{-1} \mathbf{U}(\beta) \quad (4.5)$$

multiplicando ambos os lados por $\mathbf{K}^{(k)}$ temos:

$$\mathbf{K}^{(k)} \beta^{(k+1)} = \mathbf{K}^{(k)} \beta^{(k)} + \mathbf{U}(\beta)$$

que pode ser escrito como:

$$\frac{1}{\phi} \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X} \beta^{(k+1)} = \frac{1}{\phi} \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X} \beta^{(k)} + \frac{1}{\phi} \mathbf{X}^T (\mathbf{Y} - \mu^{(k)})$$

onde $a(\phi) = \phi$. Fazendo $\mathbf{Z}^{(k)} = \mathbf{X} \beta^{(k)} + [\mathbf{W}^{(k)}]^{-1} (\mathbf{Y} - \mu^{(k)})$ temos:

$$\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X} \beta^{(k+1)} = \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{Z}^{(k)}$$

cuja solução é:

$$\beta^{(k+1)} = (\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{Z}^{(k)} \quad (4.6)$$

É interessante observar que a equação 4.6 tem a forma da solução de mínimos quadrados ponderados:

$$\begin{aligned} \tilde{\beta} &= \min_{\beta} (\mathbf{Y} - \mathbf{X} \beta)^T \mathbf{W} (\mathbf{Y} - \mathbf{X} \beta) \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \end{aligned}$$

ou $(\mathbf{X}^T \mathbf{W} \mathbf{X}) \tilde{\beta} = \mathbf{X}^T \mathbf{W} \mathbf{Y}$.

4.3.2 PROPRIEDADES DO ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA

Por ser um estimador de máxima verossimilhança, sob certas condições de regularidade e assumindo uma amostra de tamanho infinito, é possível assumir que a distribuição dos estimadores é assintoticamente normal:

$$\hat{\beta} \underset{n \rightarrow \infty}{\sim} \text{Normal}(\beta, \mathbf{K}^{-1}) \quad (4.7)$$

Portanto, a partir da equação 4.7 é possível realizar testes de hipótese e inferência nos estimadores do modelo de regressão da família exponencial.

4.4 MÉDIDAS DE AJUSTE DO MODELO DE REGRESSÃO DA FAMÍLIA EXPONENCIAL

As estimativas de máxima verossimilhança dos MLGs também podem ser obtidas, de forma equivalente, a partir da minimização da função *Deviance* ou *desvio escalonado*.

$$D(y; \mu) = 2 [l(y; y) - l(\mu; y)] \quad (4.8)$$

onde $\mu = E(Y)$ e $l(\mu; y)$ é a função log-verossimilhança. É intuitivo observar que a função *Deviance* é proporcional ao negativo da função log-verossimilhança. Portanto, maximizar a função log-verossimilhança é o mesmo que minimizar a função *Deviance*. O termo $l(y; y)$ é conhecido como a função log-verossimilhança saturada. Este termo representa uma constante,

calculada utilizando somente os valores observados da variável resposta e representa a função log-verossimilhança quando o parâmetro de esperança da variável resposta é associado ao valor observado: $\mu_i = E(Y_i) = y_i$.

Por exemplo, a função log-verossimilhança assumindo $Y_i \sim Normal(\mu; \sigma^2)$ é:

$$l(\mu; y) = \sum_i -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - \mu_i)^2}{2\sigma^2}$$

Para o cálculo da função log-verossimilhança saturada, faz-se $E(Y_i) = y_i$ ou, neste caso, $\mu_i = y_i$:

$$l(y; y) = -\frac{n}{2} \log(2\pi\sigma^2)$$

Então,

$$D(y; \mu) = 2[l(y; y) - l(\mu; y)] = \frac{\sum_i (y_i - \mu_i)^2}{\sigma^2} \quad (4.9)$$

Como pode ser visto na equação 4.9, no caso da distribuição normal, a função Deviance é proporcional à soma dos quadrados dos erros.

A Tabela 4.1 apresenta as formas da função Deviance para as principais distribuições da família exponencial, que serão estudadas. Maiores detalhes podem ser encontrados em Nelder e Wedderburn (1972).

Tabela 4.1: Tabela da função Deviance para algumas distribuições de probabilidade pertencentes à família exponencial ($\mu_i = E(Y_i)$).

Modelo		Deviance
Normal	$Y_i \sim Normal(\mu_i; \sigma^2)$	$\sum_{i=1}^n (y_i - \hat{\mu}_i)^2 / \sigma^2$
Poisson	$Y_i \sim Poisson(\mu_i)$	$2 \sum_{i=1}^n y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)$
Binomial	$Y_i \sim Binomial(m_i; p_i)$	$2 \sum_{i=1}^n y_i \log(y_i/\hat{\mu}_i) + (m_i - y_i) \log[(m_i - y_i) / (m_i - \hat{\mu}_i)]$
Gama	$Y_i \sim Gama(\mu_i; v)$	$\sum_{i=1}^n (y_i - \hat{\mu}_i)^2 / (\hat{\mu}_i^2 y_i)$

É usual a comparação do valor observado da função Deviance com a distribuição Qui-quadrado com $n - p$ graus de liberdade, onde n é o tamanho da amostra e p é o número de parâmetros do modelo (incluindo o intercepto). Esta comparação é atribuída ao fato de que a Deviance é um estatística de razão de verossimilhanças ou, de forma equivalente, é uma estatística que utiliza a diferença do logaritmo de funções de log-verossimilhança. Portanto, assintoticamente, a estatística D (Deviance) sob a hipótese de que o modelo de probabilidade escolhido é compatível com o comportamento observado nos dados, é comparada à distribuição Chi-Quadrado (χ^2) com $n - p$ graus de liberdade:

$$D|H_0 \sim \chi^2_{n-p} \quad (4.10)$$

4.5 ANÁLISE DE RESÍDUOS NA FAMÍLIA EXPONENCIAL

O conceito do resíduo de um modelo de regressão está, historicamente, vinculado à estrutura do modelo de regressão linear normal, $Y_i \sim Normal(\mu_i; \sigma^2)$, que pode ser escrito na forma $Y_i = \mu_i + \epsilon_i$ onde $\epsilon_i \sim Normal(0; \sigma^2)$. Como consequência, o resíduo do modelo (r_i) é uma estimativa da variável aleatória erro, ϵ_i ou $r_i = \hat{\epsilon}_i = y_i - \hat{\mu}_i$. Em outras distribuições de probabilidade da família exponencial, como a Poisson, Bernoulli, Binomial ou Gama; não existe uma equação aditiva de uma componente determinística de média e uma variável aleatória. Neste sentido, a análise de resíduos na família exponencial poderia ser um tanto equivocado.

Por outro lado, é pertinente investigar se as observações utilizadas no ajuste do modelo estão coerentes com a distribuição de probabilidade selecionada. Ou seja, deseja-se investigar o comportamento das observações em relação ao modelo estatístico/probabilístico ajustado.

Existem diversas propostas para a *análise de resíduos* em modelos lineares generalizados. Não há um consenso na literatura em relação ao melhor método de análise de resíduos. A seguir, será apresentado o *Desvio residual*, que apresenta uma definição compatível com o conceito da medida de ajuste *Deviance*, definida na seção anterior.

Como definido na Tabela 4.1, a função *Deviance* pode ser definida na forma genérica como:

$$\text{Deviance} = \sum_{i=1}^n d_i$$

onde $d_i = 2 [l(y_i; y_i) - l(\mu_i; y_i)]$ é a componente da *Deviance* associada a cada observação i . Caso o modelo linear generalizado seja *compatível* com os dados observados é possível comparar o valor *Deviance* com uma distribuição Qui-Quadrado com $n - p$ graus de liberdade.

Seja definido o contexto no qual $n \gg p$, tal que seria possível assumir que n é suficientemente grande para comparar a *Deviance* não mais com χ^2_{n-p} mas com χ^2_n . Então $\sum d_i \approx \chi^2_n$. Considerando que $\chi^2_n = \sum_i \chi^2_1$ é possível identificar que $d_i \sim \chi^2_1$. Ou seja para um modelo com um número grande de observações e um número reduzido de parâmetros a componente da deviance, d_i pode ser comparada a uma distribuição Qui-Quadrado com um grau de liberdade. Considerando o percentil 95 da referida distribuição, $\chi^2_{1,0.95} = 3.8415$, então uma observação é dita como *atípica* se o valor da componente da deviance $d_i > \chi^2_{1,0.95}$.

Por outro lado, a distribuição χ^2_1 é obtida a partir da distribuição normal padrão: se $r \sim N(0; 1)$ então $r^2 \sim \chi^2_1$. É possível, então, induzir uma transformação que permita a geração de um suposto *resíduo* comparável com a distribuição normal padrão:

$$r_{D_i} = \text{sinal}(y_i - \hat{\mu}_i) \cdot \sqrt{d_i} \quad (4.11)$$

A equação 4.11 permite especificar uma função para cada distribuição de probabilidade pertencente à família exponencial de forma que, para cada observação i , caso o valor seja maior que 3, ou menor que -3 (valores esses considerando um intervalo bilateral de 99% para a distribuição normal padronizada), então suspeita-se que a observação é imcompatível com as suposições do modelo de probabilidade escolhido.

É importante salientar que a análise da função *Deviance* e dos *desvios residuais* tem como objetivo avaliar se a estrutura estatística/probabilística escolhida é compatível com os dados observados. Ou, se a estrutura de média e variância especificadas no modelo são compatíveis com o comportamento observado nos dados. Esta análise será posteriormente avaliada para as principais distribuições de probabilidade utilizadas neste livro.

4.5.1 FORMAS ALTERNATIVAS PARA MEDIDA DO AJUSTE E ANÁLISE DOS RESÍDUOS EM MODELOS LINEARES GENERALIZADOS

O próprio R permite o uso de pelo menos dois tipos de resíduos para os modelos lineares generalizados, como exemplificado a seguir.

```
resíduos ← residuals(modelo, type="deviance")
resíduos ← residuals(modelo, type="pearson")
```

Os resíduos de Pearson (DAVISON; SNELL, 1991) são definidos na forma:

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i} \quad (4.12)$$

onde $\mu_i = E(Y_i)$ e $\sigma_i^2 = Var(Y_i)$. Adaptando a equação 4.12 para a família exponencial, é possível escrever:

$$r_{P_i} = \frac{y_i - b'(\theta_i)}{a(\phi) \cdot b''(\theta_i)} \quad (4.13)$$

Basicamente, os resíduos de Pearson representam a padronização da variável aleatória resposta, Y_i , com relação à sua média e respectivo desvio padrão. Como consequência, r_{P_i} possui média zero e desvio padrão unitário. É também usual a comparação dos resíduos de Pearson com os percentis de uma distribuição normal padrão. Caso o valor absoluto do resíduo de Pearson seja maior que 3, os resíduos são considerados *suspeitos*.

A partir dos resíduos de Pearson, é possível calcular a estatística de Pearson como:

$$X = \sum_{i=1}^n r_{P_i} \quad (4.14)$$

A estatística de Pearson é uma alternativa à Deviance na análise do ajuste dos modelos lineares generalizados.

Semelhante aos resíduos deletados no modelo de regressão linear múltipla, também é possível definir resíduos padronizados *deletados* para os *desvios residuais* e para os resíduos de Pearson. Para isso, é necessário definir a matriz de projeção H para o modelo linear generalizado.

A partir dos resultados das seções 4.3.1 e 4.3.2, considerando a convergência do algoritmo de estimação, é possível definir a matriz de projeção na forma:

$$H = W^{1/2} X \left(X^T W X \right)^{-1} X^T W^{1/2} \quad (4.15)$$

onde $[W^{1/2}]_{ii} = \sqrt{w_{ii}}$. Então, utilizando os elementos da diagonal da matriz H , os *desvios residuais* e os resíduos de Pearson padronizados podem ser calculados como:

$$r_{D_i}^* = \frac{r_{D_i}}{\sqrt{1 - h_{ii}}} \quad (4.16)$$

$$r_{P_i}^* = \frac{r_{P_i}}{\sqrt{1 - h_{ii}}} \quad (4.17)$$

É importante destacar que a matriz H no caso dos modelos lineares generalizados representa uma aproximação. Não sendo, portanto, garantida a propriedade dos resíduos deletados, como avaliado para o modelo de regressão linear múltipla. Por outro lado, tal característica merece ser testada podendo gerar bons resultados para validação cruzada do tipo *leave-one-out* nos modelos lineares generalizados.

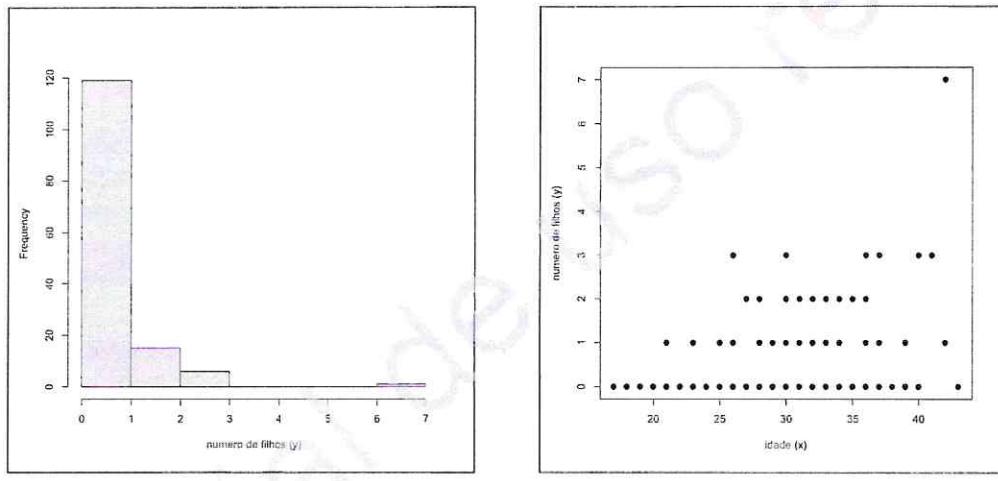
4.6 O MODELO DE REGRESSÃO DE POISSON

Para demonstrar o uso da família exponencial no ajuste de modelos de regressão cuja distribuição de probabilidade da variável resposta é não-normal, será utilizado um exemplo no qual a resposta representa dados de contagem.

A Figura 4.1 mostra uma base de dados composta por 141 mulheres grávidas. A variável preditora (x) é a idade das mulheres e o eixo y mostra o respectivo número de filhos já nascidos (variável resposta, Y). O objetivo é identificar se a idade das mulheres grávidas está associada ao número atual de filhos. Este problema é apresentado em Jong, Heller et al. (2008). A Figura 4.1 (a) mostra o histograma da variável resposta, que consiste em uma variável discreta com valor mínimo igual a zero e valor máximo igual a 7. Teoricamente, o número máximo do número de filhos é infinito. Portanto, o suporte da variável resposta é zero a infinito. A distribuição de Poisson caracteriza variáveis discretas positivas e será, em um primeiro instante, associada ao número de filhos. Portanto será investigado o seguinte modelo:

$$Y \sim \text{Poisson}(\mu) \quad (4.18)$$

onde $f_Y(y) = \frac{e^{-\mu} \mu^y}{y!}$, $E(Y) = \mu$ e $\text{Var}(Y) = \mu$.



(a) Histograma do número de filhos.

(b) Gráfico de dispersão do número de filhos em função da idade.

Figura 4.1: Análise exploratória da distribuição do número de filhos e do gráfico de dispersão em função da idade.

A distribuição de Poisson na forma da família exponencial é definida por:

$$f_Y(y) = \exp \left\{ \frac{y \cdot \log \mu - \mu}{1} - \log y! \right\} \quad (4.19)$$

Comparando as equações 4.19 e 4.1, verifica-se que $\theta = \log \mu$, ou $\mu = e^\theta$, $a(\phi) = \phi = 1$ e $c(y, \phi) = -\log y!$. Por consequência, $b(\theta) = \mu_{(\theta)} = e^\theta$. Além disso, as propriedades de esperança e variância da família exponencial podem ser verificadas:

$$\begin{aligned} E(Y) &= b'(\theta) = e^\theta \\ \text{Var}(Y) &= b''(\theta) = e^\theta \end{aligned}$$

Assumindo o modelo de regressão da família exponencial com ligação canônica, $\theta = \beta_0 + \beta_1 \cdot x$, a relação entre o parâmetro da média e a equação de regressão é na forma:

$$\mu_{(x)} = e^{\beta_0 + \beta_1 \cdot x} \quad (4.20)$$

O ajuste do modelo é mostrado a seguir.

```
> modelo <- glm(children ~ age, family = poisson, data = dados)
> summary(modelo)

Call:
glm(formula = children ~ age, family = poisson,
     data = dt)

Deviance Residuals:
    Min      1Q   Median      3Q      Max 
-2.0753 -0.9960 -0.7510  0.5358  2.8532 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.08955   0.71361 -5.731   1e-08 ***  
age          0.11295   0.02121  5.326   1e-07 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 194.42 on 140 degrees of freedom
Residual deviance: 165.01 on 139 degrees of freedom
AIC: 289.98

Number of Fisher Scoring iterations: 5
```

Os resultados mostram que as estimativas para os parâmetros de regressão são $\hat{\beta}_0 = -4.08955$ e $\hat{\beta}_1 = 0.11295$. Intervalos de confiança para os parâmetros podem ser obtidos a partir dos comandos:

```
> confint(modelo)
Waiting for profiling to be done...
              2.5 %    97.5 %
(Intercept) -5.52183430 -2.7229142
age         0.07168062  0.1548595
```

e a análise do ajuste do modelo, utilizando a estatística *Deviance* é obtida na forma:

```
> 1 - pchisq(modelo$deviance, modelo$df.residual)
0.06531977
```

Como $E(Y) = \mu = e^{\beta_0 + \beta_1 \cdot x}$, a interpretação do modelo é a seguinte: para cada aumento de uma unidade na variável regressora, $x + 1$, a média da variável resposta será multiplicada por e^{β_1} . Ou seja, se $x + 1$, então $e^{\beta_0 + \beta_1 \cdot (x+1)} = e^{\beta_1} \times e^{\beta_0 + \beta_1 \cdot x}$. Fazendo $\mu_{(x)} = e^{\beta_0 + \beta_1 \cdot x}$, segue que: $\mu_{(x+1)} = e^{\beta_1} \times \mu_{(x)}$. Por esta razão, o modelo de regressão de Poisson é dito possuir as características de um modelo multiplicativo. Para este modelo em particular, é de interesse verificar as propriedades estatísticas de e^{β_1} como, por exemplo, intervalos de confiança:

```
> exp(confint(modelo))
Waiting for profiling to be done...
              2.5 %    97.5 %
(Intercept) 0.003998507 0.06568306
age        1.074312175 1.16749396
```

No exemplo proposto $e^{\hat{\beta}_1} = e^{0.11295} = 1.119576$, ou seja, para cada aumento de um ano na idade das mulheres grávidas, o número esperado de filhos é multiplicado por 1,1195. Ou ainda, há um aumento de 11,95% no número esperado de filhos para cada aumento de um ano na idade das mulheres grávidas.

Os valores estimados: $\hat{\beta}_0$ e $\hat{\beta}_1$ podem ser utilizados para definir intervalos de predição, assumindo que $Y|x \sim Poisson(\hat{\mu}_{(x)})$, onde $\hat{\mu}_{(x)} = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x}$. A Figura 4.2 (a) mostra a média estimada pelo modelo de regressão de Poisson e os limites preditivos superior e inferior, utilizando 95% de confiança.

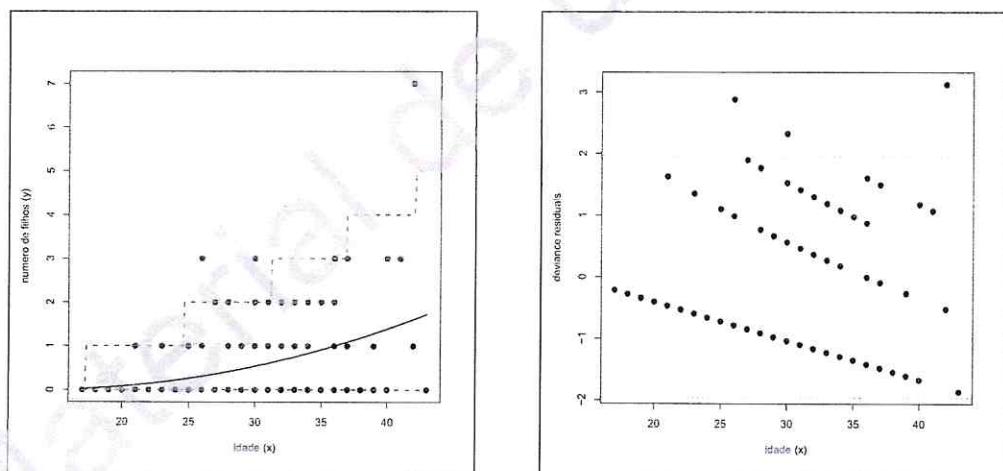
```
plot(children ~ age, data=dt, pch=19, cex=1.2, col="blue",
      xlab="idade (x)", ylab="numero de filhos (y)")
grid()
novos.dados     ← data.frame( age = seq(min(dt$age), max(dt$age), by=0.1) )
novos.dados$fit ← predict(modelo, newdata=novos.dados, type="response")
lines(fit ~ age, data=novos.dados, lwd=2)

limite.inferior ← qpois(0.025, novos.dados$fit)
limite.superior ← qpois(0.975, novos.dados$fit)

lines(novos.dados$age, limite.inferior, lty=2, col="red", lwd=2)
lines(novos.dados$age, limite.superior, lty=2, col="red", lwd=2)
```

Os resultados mostram que existem três pontos fora dos intervalos preditivos. É importante destacar que o modelo de regressão de Poisson assume que $Var(Y) = E(Y)$, ou seja, para valores médios elevados a dispersão dos dados será elevada. No problema, isso significa que quanto maior o número esperado (médio) de filhos, maior será a amplitude (variabilidade) das observações. A análise da Deviance residual permite identificar as observações para as quais a suposição do modelo, $Var(Y) = E(Y)$, é limitada. A Figura 4.2 (b) mostra a análise da deviance residual em função da idade. A Figura mostra que existem três observações discrepantes: $(x = 26, y = 3, r.deviance = 2.85)$, $(x = 42, y = 7, r.deviance = 2.82)$ e $(x = 42, y = 0, r.deviance = -2.08)$.

```
> dados$deviance ← residuals(modelo, type="deviance")
> plot(deviance ~ age, ylab="deviance residuals", xlab="idade (x)",
+       data=dados, pch=19, col="blue", cex=1.2); grid()
> abline(h = c(qnorm(0.025), qnorm(0.975)), col="red", lty=2)
```



(a) Número de filhos em função da idade.

(b) Deviance residual.

Figura 4.2: Resultado do ajuste do modelo de regressão de Poisson (a) e análise da Deviance residual (b).

4.6.1 ESTUDO DE SIMULAÇÃO PARA O MÓDELO DE POISSON

Como mostrado na equação 4.7, os estimadores de máxima verossimilhança para a família exponencial convergem para a distribuição normal quando o tamanho de amostra tende para o infinito. Na prática, assume-se que, se o tamanho amostral for suficientemente grande, os

estimadores apresentam um comportamento segundo uma distribuição normal. Entretanto, é subjetivo indicar qual o tamanho mínimo da amostra para o qual é possível assumir que o *infinito* foi alcançado. Para avaliar a suposição de que o tamanho amostral é suficientemente grande, podemos realizar um estudo de simulação com os próprios dados ajustados.

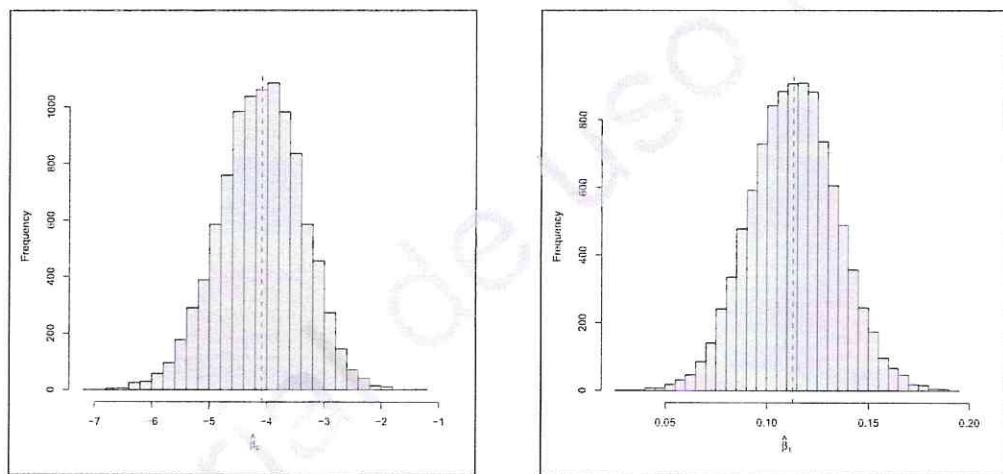
No caso do número de filhos, vamos fixar os parâmetros estimados e simular várias bases de dados de tamanho $n = 141$, que é o tamanho original da base de dados. A simulação será feita somente para a variável resposta Y_i . O modelo simulado é, então, na forma:

$$Y_i \sim \text{Poisson} \left(\mu_i = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i} \right) \quad (4.21)$$

A seguir é apresentado o código em R referente ao estudo de simulação e os resultados das estimativas. Foram realizadas 10.000 simulações.

```
modelo <- glm(children ~ age, family = poisson, data = dados)
beta0 <- modelo$coef[1]
beta1 <- modelo$coef[2]
mu     <- exp(beta0 + beta1 * dados$age)
vbetas <- c()

for(s in 1:10000){
  dados$ysim <- rpois(length(mu), mu)
  modelo.sim <- glm(ysim ~ age, family = poisson, data = dados)
  vbetas      <- rbind(vbetas, coef(modelo.sim))
}
```



(a) Estudo de simulação para $\hat{\beta}_0$.

(b) Estudo de simulação para $\hat{\beta}_1$.

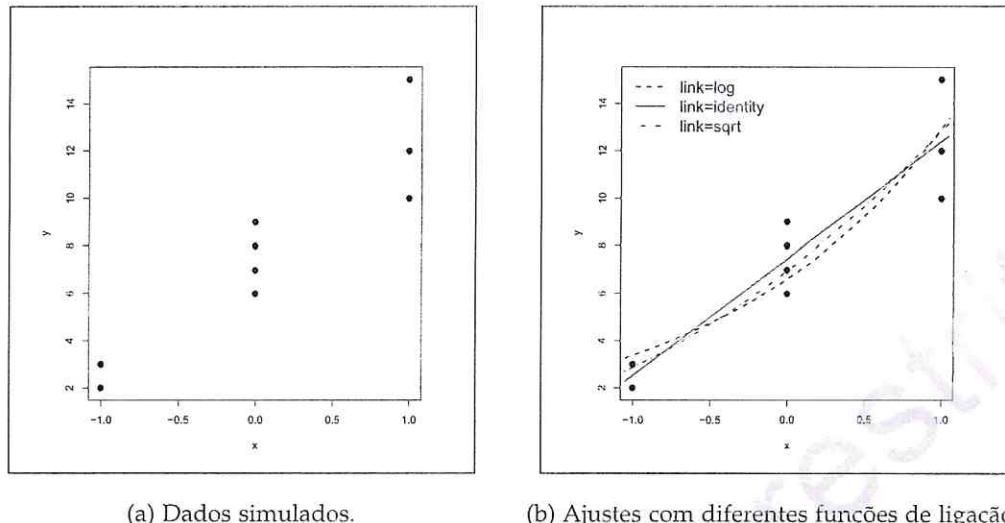
Figura 4.3: Estudo de simulação dos estimadores, $\hat{\beta}_0$ e $\hat{\beta}_1$, para tamanho de amostra $n = 141$. A linha tracejada vertical indica os valores utilizados para a simulação: $\beta_0 = -4.0895$ e $\beta_1 = 0.1129$.

Os resultados apresentados na Figura 4.3 mostram que, para uma amostra de tamanho $n = 141$, os estimadores apresentam uma distribuição levemente assimétrica. Particularmente este fato é característico na Figura 4.3 (a). Por outro lado, os histogramas das simulações estão centralizados nos valores reais dos coeficientes (β_0 e β_1). Este resultado demonstra que as estimativas de máxima verossimilhança para amostras limitadas são, em geral, não viesadas para o modelo de Poisson.

4.6.2 CASOS ESPECIAIS: LIGAÇÃO NÃO CANÔNICA

A ligação canônica, $\theta = \beta_0 + \beta_1 \cdot x$, ou $\mu = e^\theta$, é a forma natural de incorporar a equação de regressão em uma distribuição da família exponencial. Entretanto, em algumas circunstâncias, é

possível explorar mais de uma opção. A Figura 4.4 (a) mostra um conjunto de dados simulados que representam contagens. É possível perceber que, além de serem dados de contagem, a dispersão dos dados está aumentando à medida que os valores da resposta aumentam. Esses resultados indicam o ajuste de um modelo de regressão de Poisson.



(a) Dados simulados.

(b) Ajustes com diferentes funções de ligação.

Figura 4.4: Ajustes da regressão de Poisson utilizando diferentes funções de ligação.

A Figura 4.4 (b) apresenta o ajuste de três modelos de Poisson utilizando diferentes funções de ligação. A função de ligação canônica é do tipo $\mu(x) = e^{\beta_0 + \beta_1 \cdot x}$. Para esta função de ligação, é possível especificar no código R a componente `link="log"`, como ilustra o código a seguir.

```
> y      ← c(2, 3, 6, 7, 8, 9, 10, 12, 15)
> x      ← c(-1, -1, 0, 0, 0, 1, 1, 1)
> dados  ← data.frame(x=x, y=y)
> modelo ← glm(y ~ x, family=poisson(link="log"), data = dados)
> summary(modelo)

Call:
glm(formula = y ~ x, family = poisson(link = "log"), data = dados)

Deviance Residuals:
Min     1Q   Median     3Q    Max 
-0.8472 -0.2601 -0.2137  0.5214  0.8788 

Coefficients:
Estimate Std. Error z value Pr(>|z|)    
(Intercept) 1.8893     0.1421 13.294 < 2e-16 ***
x          0.6698     0.1787  3.748 0.000178 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 18.4206  on 8 degrees of freedom
Residual deviance: 2.9387  on 7 degrees of freedom
AIC: 41.052
---
```

A função de ligação do tipo *log* também é a opção padrão para a regressão de Poisson.

A função de ligação do tipo $\mu(x) = \beta_0 + \beta_1 \cdot x$ é conhecida como função de ligação identidade (*identity*). Neste caso, é possível especificar no código R a componente `link="identity"`, como ilustra o código a seguir.

```

> modelo <- glm(y ~ x, family=poisson(link="identity"), data = dados)
> summary(modelo)

Call:
glm(formula = y ~ x, family = poisson(link = "identity"), data = dados)

Deviance Residuals:
Min      1Q   Median      3Q      Max 
-0.7019 -0.3377 -0.1105  0.2958  0.7184 

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 7.4516     0.8841    8.426 < 2e-16 ***
x           4.9353     1.0892    4.531 5.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 18.4206  on 8 degrees of freedom
Residual deviance: 1.8947  on 7 degrees of freedom
AIC: 40.008

```

A função de ligação identidade assume a reta de regressão, ou seja, um modelo linear para representar o comportamento médio dos dados. É importante destacar dois pontos importantes: (A) escolher a função identidade não gera, necessariamente, o mesmo resultado do ajuste do modelo de regressão linear clássico. Isso porque a função de otimização para o ajuste dos parâmetros do modelo de regressão é diferente em ambos os casos. No modelo de Poisson, os parâmetros são estimados utilizando a maximização da verossimilhança de Poisson. No modelo de regressão linear, os parâmetros são estimados utilizando a minimização da soma dos quadrados dos erros. (B) Utilizando a função identidade é possível gerar valores médios de contagem negativos. Dessa forma, há uma violação conceitual no modelo.

Entretanto, ao observar o ajuste dos modelos, a função identidade apresentou um melhor ajuste. Este fato pode ser também comprovado pelo menor valor da Deviance do modelo (1.8947). Portanto, no exemplo, a função identidade pode ser considerada uma boa aproximação local para o comportamento médio dos dados e, portanto, após o seu ajuste, o analista deve ter consciência de possíveis violações e suas implicações. Por outro lado, a função identidade também apresenta uma interpretação intuitiva para o modelo. No exemplo, para cada aumento de uma unidade na variável preditora, a variável resposta terá um aumento médio de 4.9353.

A função de ligação do tipo $\mu(x) = (\beta_0 + \beta_1 \cdot x)^2$ é conhecida como função de ligação raiz quadrada (*sqrt*). Neste caso, é possível especificar no código R a componente *link="sqrt"*, como ilustra o código a seguir

```

> modelo <- glm(y ~ x, family=poisson(link="sqrt"), data = dados)
> summary(modelo)

Call:
glm(formula = y ~ x, family = poisson(link = "sqrt"), data = dados)

Deviance Residuals:
Min      1Q   Median      3Q      Max 
-0.82315 -0.36827  0.01977  0.38968  0.74429 

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.6359     0.1685   15.639 < 2e-16 ***
x           0.9465     0.2261    4.186 2.84e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```

```
Null deviance: 18.4206 on 8 degrees of freedom
Residual deviance: 2.2145 on 7 degrees of freedom
AIC: 40.328
```

Na concepção deste autor, a ligação raiz quadrado procura estabelecer um *meio termo* entre a ligação identidade e a ligação do tipo log. Entretanto, a interpretação dos parâmetros do modelo é complicada. É possível escrever a equação da reta na forma $\sqrt{\mu(x)} = \beta_0 + \beta_1 \cdot x$, implicando que para o aumento de uma unidade na variável preditora, a raiz quadrada da variável resposta terá um aumento médio de 0.9465. Ou seja, esta interpretação não é intuitiva. Além disso, assumir a ligação raiz quadrada implica em uma equação do tipo parabólica para a resposta média do modelo. Ou seja, de um lado do ponto de mínimo haverá uma direção de crescimento/decrescimento da variável resposta e, do outro lado, terá a mesma direção de crescimento/decrescimento. Portanto, na concepção deste autor, a função de ligação raiz quadrada deve ser evitada.

4.7 O MODELO DE REGRESSÃO LOGÍSTICO (BERNOULLI)

Uma variável aleatória que assume somente dois possíveis valores, $Y \in \{0, 1\}$, pode ser modelada por uma distribuição de Bernoulli: $Y \sim Bernoulli(p)$, onde $P(Y = 1) = p$ e $P(Y = 0) = 1 - P(Y = 1) = p(1 - p)$. Uma representação sucinta da distribuição de probabilidade de Y é apresentada na equação 4.22.

$$Y = \begin{cases} 1, & \text{com probabilidade } p. \\ 0, & \text{com probabilidade } 1 - p. \end{cases} \quad (4.22)$$

A partir da equação 4.22, é possível mostrar que $E[Y] = p$ e $Var[Y] = p(1 - p)$.

Uma variável de Bernoulli pode ser criada a partir de variáveis categóricas: {Sim, Não}, {Verdadeiro, Falso}, {A, B}, etc. Em cada exemplo, é necessário definir qual o valor de referência: $Y_{(\text{Sim})} = 1$, $Y_{(\text{Verdadeiro})} = 1$, $Y_{(A)} = 1$. Esta escolha é arbitrária.

A distribuição de probabilidade de Bernoulli é também definida como:

$$P(Y = y) = p^y(1 - p)^{1-y} \quad (4.23)$$

Escrevendo a equação 4.23 na forma exponencial:

$$\begin{aligned} P(Y = y) &= \exp \{y \log p + (1 - y) \log (1 - p)\} \\ &= \exp \left\{ y \log \left(\frac{p}{1 - p} \right) + \log(1 - p) \right\} \end{aligned} \quad (4.24)$$

é possível identificar que:

1. $\theta = \log \left(\frac{p}{1-p} \right)$, ou $p = \frac{e^\theta}{1+e^\theta}$;
2. $b(\theta) = -\log \left(1 - p_{(\theta)} \right) = -\log \left(\frac{1}{1+e^\theta} \right) = +\log (1 + e^\theta)$.

Considerando a ligação canônica, $\eta = x\beta$, o modelo de regressão de Bernoulli pode ser escrito na forma:

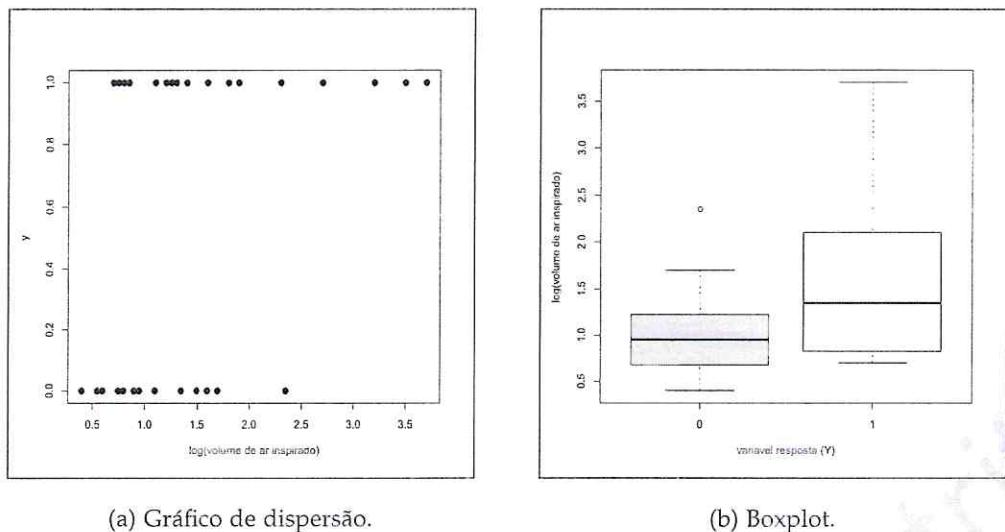
$$Y|x \sim Bernoulli \left(p = \frac{e^{x\beta}}{1 + e^{x\beta}} \right) \quad (4.25)$$

Considerando uma única variável preditora (x):

$$Y|x \sim Bernoulli \left(p = \frac{e^{\beta_0 + \beta_1 \cdot x}}{1 + e^{\beta_0 + \beta_1 \cdot x}} \right) \quad (4.26)$$

Para exemplificar o modelo de regressão canônica, considere o exemplo mostrado a seguir. A base de dados caracteriza a ocorrência ($y_i = 1$) ou não ($y_i = 0$) de vaso-constrição na pele dos dedos da mão a partir do logaritmo do volume de ar inspirado (x_i) em um conjunto de 39 indivíduos ($n = 39$).

A Figura 4.5 (a) procura demonstrar a relação entre a variável resposta e a variável preditora utilizando um gráfico de dispersão. Entretanto, como a variável resposta é dicotômica (somente dois valores), a abordagem mais apropriada é a representação em um boxplot, conforme Figura 4.5 (a). A Figura 4.5 mostra que a variável preditora (agora no eixo y) apresenta boxplot deslocado para valores superiores quando ocorre vaso-constrição ($y_i = 1$). Ou seja, quando ocorre vaso-constrição, em geral, os indivíduos apresentam valores elevados do logaritmo do volume de ar inspirado.



(a) Gráfico de dispersão.

(b) Boxplot.

Figura 4.5: Análise exploratória de dados para respostas binárias.

Assumindo o modelo de regressão da família exponencial com ligação canônica, a relação entre a média e a equação de regressão é na forma:

$$p(x) = \frac{e^{\beta_0 + \beta_1 \cdot x}}{1 + e^{\beta_0 + \beta_1 \cdot x}} \quad (4.27)$$

O ajuste do modelo é mostrado a seguir:

```
> modelo <- glm(y ~ logVolumeAr, family = binomial, data = dados)
> summary(modelo)

Call:
glm(formula = y ~ logVolumeAr, family = binomial, data = dados)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.8341 -1.0023  0.2719  1.1185  1.4978 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.6627    0.8123 -2.047   0.0407 *  
logVolumeAr  1.3357    0.6162  2.168   0.0302 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 54.040  on 38  degrees of freedom
Residual deviance: 46.989  on 37  degrees of freedom
AIC: 50.989

Number of Fisher Scoring iterations: 4

> c( modelo$deviance, 1-pchisq(modelo$deviance,modelo$df.residual) )
[1] 46.9893786 0.1257694
```

Os resultados mostram que existe uma relação estatisticamente significativa entre o logaritmo do volume de ar inspirado e a ocorrência de vaso-constricção $\hat{\beta}_1 = 1.3357$ (p-valor=0.0302). A média estimada para o modelo de Bernoulli também deve ser interpretada como a probabilidade estimada da ocorrência de vaso-constricção:

$$\hat{P}(Y_i = 1|x_i) = \hat{p}(x_i) = \frac{e^{-1.6627+1.3357 \cdot x_i}}{1 + e^{-1.6627+1.3357 \cdot x_i}} \quad (4.28)$$

A interpretação dos parâmetros do modelo de regressão Bernoulli é peculiar. Não é possível avaliar diretamente qual o efeito na resposta média do modelo em virtude do aumento de uma unidade na variável preditora ($x + 1$). Para definir uma interpretação para o coeficiente de regressão β_1 é necessário definir o conceito de razão de chance (O_R - Odds ratio):

$$\begin{aligned} O_R &= \frac{P(Y_i = 1)}{P(Y_i = 0)} = \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} \\ &= \frac{\frac{e^{\beta_0 + \beta_1 \cdot x}}{1 + e^{\beta_0 + \beta_1 \cdot x}}}{1 - \frac{e^{\beta_0 + \beta_1 \cdot x}}{1 + e^{\beta_0 + \beta_1 \cdot x}}} \\ &= e^{\beta_0 + \beta_1 \cdot x} \end{aligned} \quad (4.29)$$

Observe que, segundo a equação 4.29, a interpretação dos coeficientes do modelo de Bernoulli é muito semelhante à interpretação dos coeficientes do modelo de Poisson: a interpretação do modelo de regressão de Bernoulli é multiplicativa. Mas, avalia-se o efeito do estimador na razão de chance, $\frac{p(x_i)}{1-p(x_i)}$ e não na média do modelo ($E(Y|x_i) = p(x_i)$). Ou seja, para cada aumento de uma unidade na variável regressora ($x_i + 1$) a razão de chance será multiplicada por e^{β_1} :

$$\frac{p(x_i + 1)}{1 - p(x_i + 1)} = \frac{p(x_i)}{1 - p(x_i)} \times e^{\beta_1} \quad (4.30)$$

ou na forma:

$$\frac{P(Y = 1|x_i + 1)}{P(Y = 0|x_i + 1)} = \frac{P(Y = 1|x_i)}{P(Y = 0|x_i)} \times e^{\beta_1} \quad (4.31)$$

Portanto, verificando as Equações 4.30 e 4.31 é mais informativo, avaliar as propriedades estatísticas de $e^{\hat{\beta}_1}$ do que $\hat{\beta}_1$:

```
> exp(confint(modelo))
Waiting for profiling to be done...
      2.5 %    97.5 %
(Intercept) 0.03184434  0.8152995
logVolumeAr 1.35844788 15.9089104
```

Os resultados mostram que para cada aumento de uma unidade no logaritmo da razão de ar inspirado a razão de chance será multiplicada por 3.8026 ($e^{1.3357}$), o que representa um aumento de 380%.

É importante observar que o efeito do aumento de uma unidade no valor médio de $E(Y|x_i + 1)$, somente será calculável se o valor $\hat{E}(Y|x_i)$ for conhecido. Por exemplo, suponha que a probabilidade estimada seja de $\hat{E}(Y|x_i) = \hat{P}(Y = 1|x_i) = 0.6$. Se ocorrer um aumento de uma unidade na variável preditora, $x_i + 1$, então o novo valor da probabilidade será na forma:

$$\frac{\hat{P}(Y = 1|x_i + 1)}{\hat{P}(Y = 0|x_i + 1)} = \frac{0.6}{1 - 0.6} \times 3.8026 \quad (4.32)$$

fazendo $a = \frac{\hat{P}(Y=1|x_i)}{\hat{P}(Y=0|x_i)} \times e^{\hat{\beta}_1} = \frac{0.6}{1-0.6} \times 3.8026$, é possível demonstrar que:

$$\hat{P}(Y = 1|x_i + 1) = \frac{a}{a + 1} = 0.8508 \quad (4.33)$$

ou seja, para o modelo logístico, para cada aumento de uma unidade na variável preditora (x) a razão estimada entre a probabilidade de sucesso e a probabilidade de fracasso será multiplicada pela exponencial do coeficiente estimado.

A resposta do modelo de regressão logístico univariado é apresentado na Figura 4.6. É importante destacar que o modelo logístico estima a probabilidade do evento $Y = 1$ ocorrer. Por se tratar de uma medida de probabilidade, $0 < \hat{P}(Y = 1|x) < 1$, a partir do resultado estimado $\hat{P}(Y = 1|x)$ o usuário deve decidir se o evento ocorrerá ou não. Em geral, assume-se que quando $\hat{P}(Y = 1|x) \geq 0.5$ o evento $Y = 1$ irá ocorrer e quando $\hat{P}(Y = 1|x) < 0.5$ o evento $Y = 1$ não irá ocorrer, ou seja, o evento $Y = 0$ irá ocorrer. Entretanto, esta escolha é arbitrária e deve ser cuidadosamente analisada.

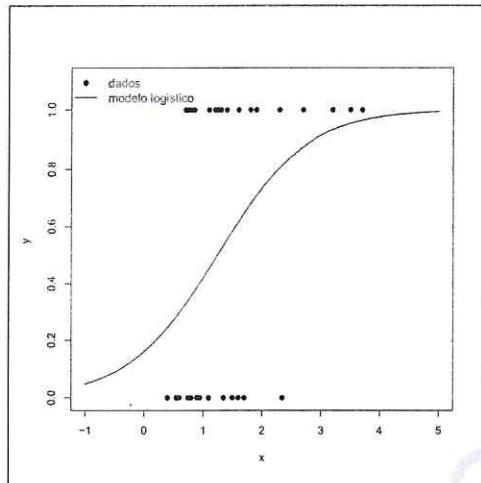


Figura 4.6: Resultado do ajuste do modelo logístico, que estima a probabilidade do evento de referência ($Y = 1$) ocorrer, $\hat{P}(Y = 1|x)$.

4.7.1 ESTUDO DE SIMULAÇÃO PARA O MODELO LOGÍSTICO (BERNOULLI)

É importante relembrar que a inferência estatística sobre os estimadores dos parâmetros da família exponencial assume que, sob certas condições de regularidade e assumindo uma amostra de tamanho infinito, é possível assumir uma distribuição assintoticamente normal, conforme a equação 4.7. Esta propriedade foi investigada para o modelo de Poisson, utilizando um estudo de simulação apresentado na seção 4.6.1. O mesmo estudo de simulação será utilizado para investigar as propriedades estatísticas do estimador de máxima verossimilhança do modelo logístico, conforme código apresentado a seguir.

```

beta0 ← -1.67
beta1 ← +1.34

nsim ← 1000 #número de simulações
nsmple ← 40 # tamanho da amostra
betas ← matrix(0,nsim,2)
y ← c()

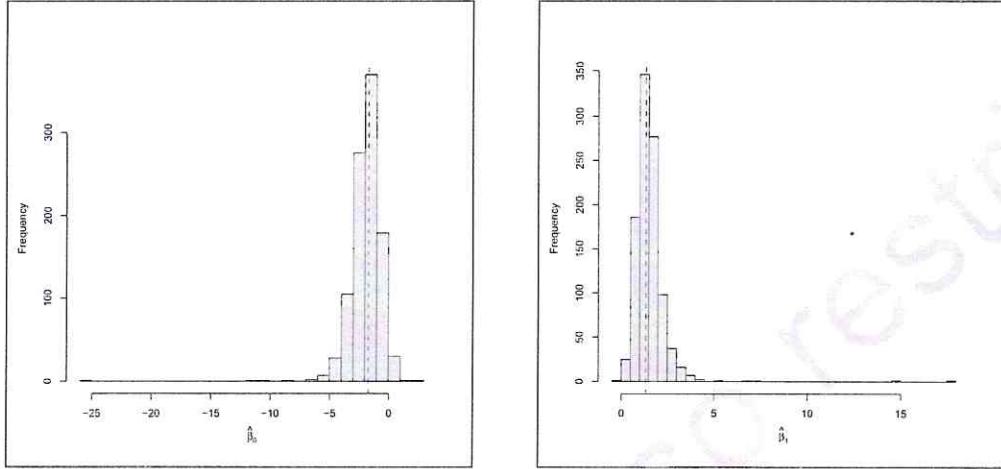
for(j in 1:nsim){
  x ← runif(nsmple, 0.4, 3.7)
  linear ← beta0 + beta1*x
  p ← exp(linear)/(1 + exp(linear))
  y ← rbinom(n=length(p), size=1, p)
  dados ← list(y=y,x=x);
  modelo ← glm(y ~ x, family = binomial, data = dados)
  betas[j,] ← modelo$coef
}

```

O modelo de simulação é definido como:

$$Y_i \sim \text{Bernoulli} \left(p_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})} \right) \quad (4.34)$$

onde os valores de β_0 e β_1 foram obtidos no exemplo logístico apresentado anteriormente. A Figura 4.7 mostra os resultados do estudo de simulação para tamanho de amostra $n = 40$. Os resultados mostram que, para pequenas amostras, os estimadores apresentam distribuição assimétrica com valores extremos. Portanto, para amostras pequenas, as inferências estatísticas para o modelo logístico não são confiáveis.



(a) Estudo de simulação para $\hat{\beta}_0$.

(b) Estudo de simulação para $\hat{\beta}_1$.

Figura 4.7: Estudo de simulação dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ para tamanho de amostra $n = 40$. A linha tracejada vertical indica os valores utilizados para a simulação: $\beta_0 = -1.67$ e $\beta_1 = +1.34$.

A Figura 4.8 apresenta os resultados do estudo de simulação para tamanho de amostra $n = 100$. Os resultados mostram um comportamento mais consistente com a distribuição normal se comparado aos resultados utilizando $n = 40$. Embora seja possível verificar uma leve assimetria na distribuição dos estimadores, os valores simulados estão próximos dos verdadeiros valores. Portanto, é importante destacar a sensibilidade do modelo logístico em relação ao tamanho da amostra. Caso haja a suspeita de que as propriedades assintóticas não podem ser aplicadas, a alternativa é realizar um estudo de simulação.

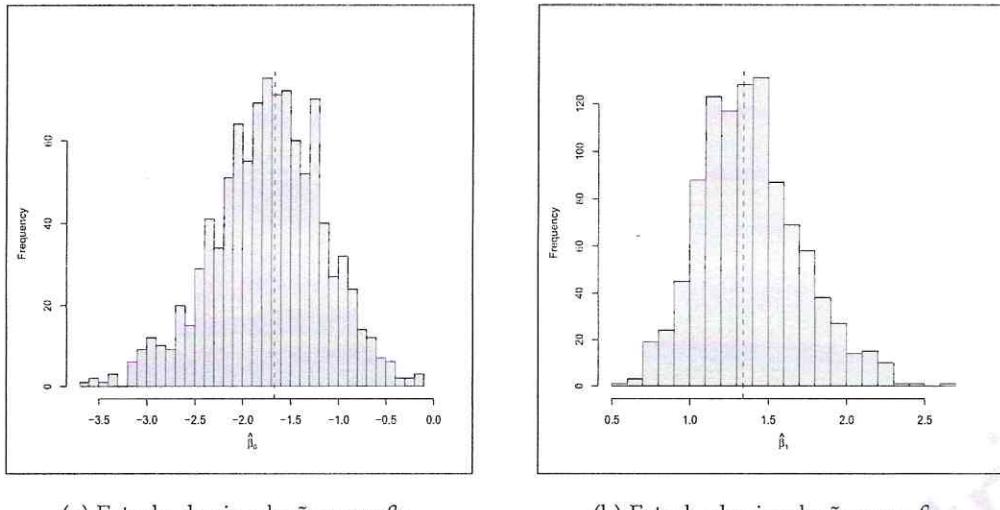
(a) Estudo de simulação para $\hat{\beta}_0$.(b) Estudo de simulação para $\hat{\beta}_1$.

Figura 4.8: Estudo de simulação dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ para tamanho de amostra $n = 100$. A linha tracejada vertical indica os valores utilizados para a simulação: $\beta_0 = -1.67$ e $\beta_1 = +1.34$.

4.7.2 ANÁLISE DA CURVA ROC

A análise de curva ROC (*Receiver Operating Characteristics*) (BRADLEY, 1997), embora tenha origem na área de processamento de sinais, é uma importante ferramenta para a análise da resposta do modelo de regressão logístico.

Inicialmente, define-se a **sensibilidade** observada do modelo logístico como a proporção de acertos das observações referentes à resposta $Y = 1$. A **especificidade** observada do modelo logístico é definida como a proporção de acerto das observações referentes à resposta $Y = 0$. Definindo um limiar de corte τ , a sensibilidade é definida na equação 4.35 e a especificidade é definida na equação 4.36.

$$s(\tau) = \frac{\sum_i \mathbf{1}(\hat{P}(Y=1|x_i) \geq \tau) \times y_i}{n} \quad (4.35)$$

$$e(\tau) = \frac{\sum_i \mathbf{1}(\hat{P}(Y=1|x_i) < \tau) \times (1 - y_i)}{n} \quad (4.36)$$

onde $\mathbf{1}(\cdot)$ é a função indicadora. Como mostram as equações 4.35 e 4.36, a sensibilidade e a especificidade são funções do limiar de corte τ .

Utilizando a base de dados descrita previamente, para $\tau = 0,5$, do total de 20 observações para as quais $y_i = 1$, 12 observações são classificadas corretamente. Portanto, a sensibilidade é de $12/20 = 0,60$ (60%). Dentre as 19 observações para as quais $y_i = 0$, 14 observações são classificadas corretamente. Portanto a especificidade é de $14/19 = 0,7368$ (73,68%). Este fato é ilustrado na Figura 4.9 (a). Caso o valor do limiar de corte τ for reduzido para $\tau = 0,25$ a sensibilidade aumentará para $s(0,25) = 1$ (100%) e a especificidade será reduzida para $e(0,25) = 0,0526$ (5,26%), como mostra a Figura 4.9 (b). Por outro lado, caso o valor do limiar de corte seja aumentado para $\tau = 0,80$ a sensibilidade será reduzida para $s(0,80) = 0,25$ (25%) e a especificidade será aumentada para $e(0,80) = 0,9473$ (94,73%), como mostra a Figura 4.9 (c). Portanto, uma redução do limiar de corte gera o aumento da sensibilidade e a redução da especificidade e o aumento do limiar de corte provoca a redução da sensibilidade e o aumento da especificidade. O comportamento da sensibilidade e da especificidade para diferentes valores de τ é mostrado na Figura 4.9 (d).

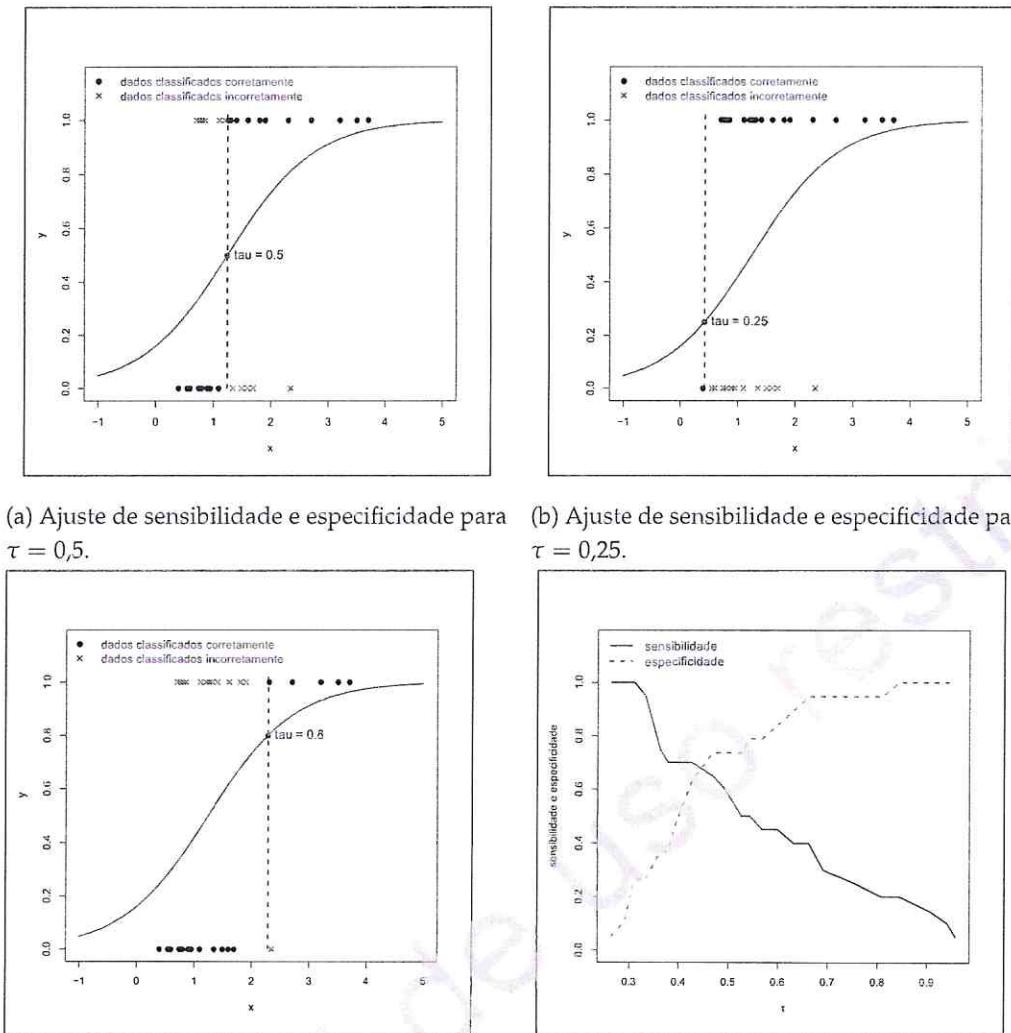


Figura 4.9: Comportamento da sensibilidade e especificidade para diferentes valores de τ .

A curva ROC representa o desenho da *sensibilidade* em função de $1 - \text{especificidade}$ para todos os possíveis valores do limiar de corte: $0 \leq \tau \leq 1$. O resultado é mostrado na Figura 4.10. A *sensibilidade* é também definida como a taxa de verdadeiros positivos (*True Positive Rate* - TPR) e $1 - \text{especificidade}$ é definida como a taxa de falsos positivos (*False Positive Rate* - FPR). Como a sensibilidade e a especificidade representam taxas de acerto, seus valores estão contidos no intervalo $[0 - 1]$ e, como consequência, a curva ROC está definida no quadrado unitário. Neste caso, é possível definir a área abaixo da curva ROC (*area under curve* - AUC) como uma medida do ajuste do modelo logístico cujo comportamento é semelhante ao coeficiente de determinação (R^2). Quanto mais próxima for a resposta estimada pelo modelo logístico da resposta observada, maior será a AUC. O limite superior da estatística AUC é a unidade: $AUC \leq 1$. É comum o uso de curvas ROC e da estatística AUC para comparação de diferentes modelos logísticos e de classificação. Em geral, os modelos com os melhores ajustes apresentam maiores valores para a estatística AUC.

De posse das curvas de sensibilidade e especificidade, o usuário pode definir o ponto que corte τ que favoreça a sensibilidade ou a especificidade, dependendo do interesse da aplicação.

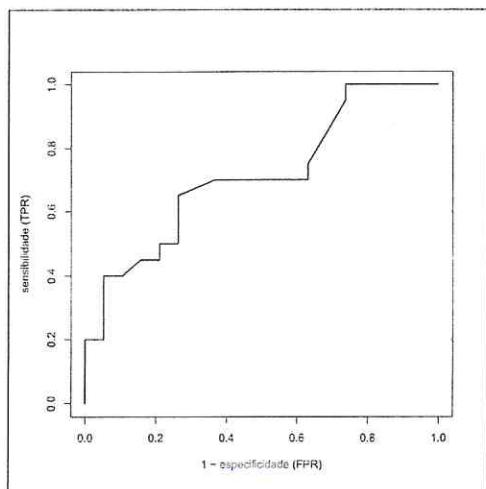


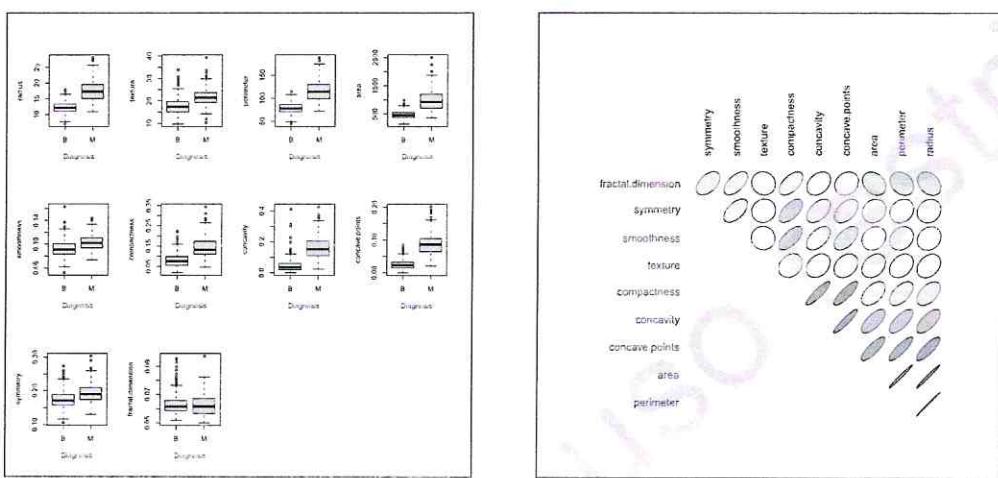
Figura 4.10: Curva ROC.

O código para a estimativa da curva ROC e da estatística AUC é mostrado a seguir.

```
> require(pROC)
> curva.roc <- roc(dados$y, predict(modelo, type="response"))
> plot(curva.roc)
> curva.roc$auc
  Área under the curve: 0.7118
```

4.7.3 ESTUDO DE CASO PARA O MODELO LOGÍSTICO

O *Machine Learning Repository* (LICHMAN, 2013) (<<http://archive.ics.uci.edu/ml/>>) apresenta uma base de dados, inicialmente apresentada por Street, Wolberg e Mangasarian (1993) e citada posteriormente em inúmeros trabalhos. Denominada de *Breast Cancer Wisconsin (Diagnostic) Data Set*, a base de dados contém 569 observações de características citológicas de células de câncer de mama que foram posteriormente classificadas como *benignas* ou *malignas*. Valores médios para 10 potenciais variáveis preditoras estão disponíveis na base de dados. A Figura 4.11 apresenta boxplots das variáveis preditoras com relação à variável resposta e a matriz de correlação linear entre as variáveis preditoras. A Figura 4.11 (a) indica variáveis preditoras que apresentam alta discriminação para as classes *B* (Benigno) e *M* (Maligno). Por outro lado, é evidente a presença de pares de variáveis preditoras com forte correlação linear, conforme Figura 4.11 (b).



(a) Boxplot das variáveis preditoras com relação à variável resposta.

(b) Gráfico de correlação das variáveis preditoras.

Figura 4.11: Análise exploratória da base de dados *Breast Cancer Winsconsin (Diagnostic)*

A referência para a variável resposta ($Y = 1$) foi definida como a característica maligna ($Y = M$). O modelo logístico foi inicialmente ajustado considerando todas as 10 variáveis preditoras. Em seguida, as variáveis que apresentavam um valor-P maior que 0.10 foram sucessivamente removidas do modelo, com excessão do intercepto que foi mantido no modelo. O resultado do ajuste do modelo final é apresentado a seguir.

```

> modelo <- glm(Diagnosis ~ radius + texture + area + smoothness
+                  + concave.points, family=binomial, data=dados)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(modelo)

Call:
glm(formula = Diagnosis ~ radius + texture + area + smoothness +
    concave.points, family = binomial, data = dados)

Deviance Residuals:
    Min      1Q   Median      3Q      Max 
-1.94769 -0.16164 -0.04965  0.00378  2.77371 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.64306   7.96274 -0.583  0.55983    
radius       -2.88770   1.17018 -2.468  0.01360 *  
texture        0.37317   0.06258  5.963 2.47e-09 *** 
area          0.04400   0.01422  3.094  0.00197 ** 
smoothness    63.18020  23.87993  2.646  0.00815 ** 
concave.points 81.11419  15.91206  5.098 3.44e-07 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 751.44  on 568  degrees of freedom
Residual deviance: 150.68  on 563  degrees of freedom
AIC: 162.68

```

Variáveis preditoras que apresentam coeficientes positivos contribuem para o aumento da chance da classificação maligna, ao passo que variáveis preditoras que apresentam coeficiente negativos contribuem para a redução da chance de classificação maligna das células. As estimativas das razões de chance (OR) com os respectivos intervalos de confiança são apresentado a seguir.

```

> exp(coef(modelo))
            (Intercept)      radius      texture      area      smoothness
concave.points
9.628197e-03 5.570424e-02 1.452335e+00 1.044981e+00 2.7456710e+27 1.688278e
+35
> exp(confint(modelo))
Waiting for profiling to be done...
              2.5 %    97.5 %
(Intercept) 4.293367e-10 2.151366e+04
radius       5.664391e-03 5.795693e-01
texture      1.294167e+00 1.657015e+00
area         1.016397e+00 1.075133e+00
smoothness   6.692051e+06 1.292984e+48
concave.points 1.710837e+22 4.223623e+49

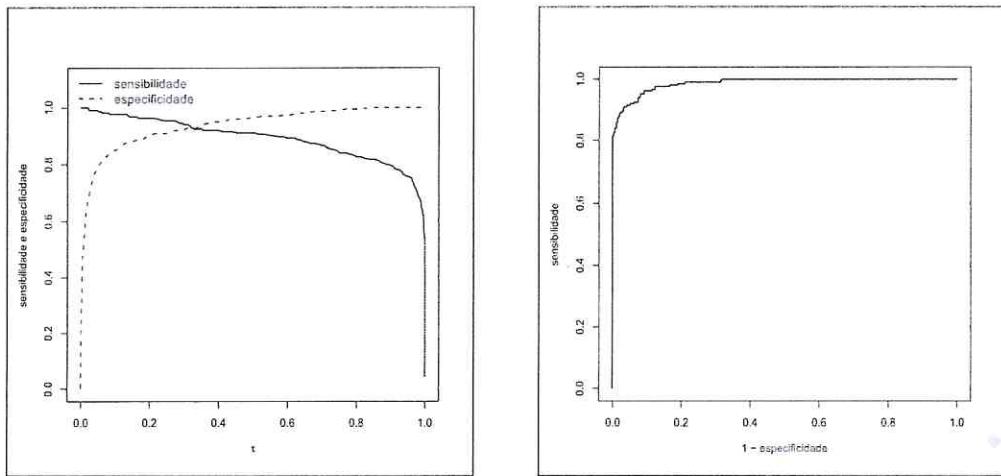
```

As curvas de sensibilidade e especificidade são mostradas na Figura 4.12 (a), ao passo que a curva ROC é mostrada na Figura 4.12 (b). A área sobre a curva ROC (AUC) é de:

```

> require(pROC)
> curva.roc <- roc(dados$Diagnosis, predict(modelo, type="response"))
> curva.roc$auc
Area under the curve: 0.9874

```



(a) Curvas de sensibilidade e especificidade.

(b) Curva ROC.

Figura 4.12: Ajustes de sensibilidade e especificidade do modelo de regressão logístico (*Breast Cancer Wisconsin (Diagnostic)*).

Os resultados mostram que o modelo logístico apresenta grande capacidade de discriminação entre as classes de tumores malignos e benignos.

4.8 O MODELO DE REGRESSÃO BINOMIAL

Uma variável aleatória Binomial pode ser representada pela soma de n variáveis aleatórias de Bernoulli com média p , $Y = \sum_{i=1}^n Y_i^*$, onde $Y_i^* \sim Bernoulli(p)$. Então, o modelo de regressão Binomial pode ser definido na forma:

$$Y_i \sim Binomial(n_i, p_i) \quad (4.37)$$

onde p_i pode ser expresso por variáveis preditoras utilizando a função logística:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad (4.38)$$

Portanto, a interpretação dos estimadores da equação de regressão do modelo Binomial é idêntica à interpretação do modelo logístico (Bernoulli). A distribuição de probabilidade do modelo Binomial é:

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} \quad (4.39)$$

A partir da equação 4.39, a equação log-Verossimilhança pode ser definida por:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^N \log \binom{n_i}{y_i} + y_i \log p_i + (n_i - y_i) \log (1 - p_i) \\ &= \sum_{i=1}^N \log \binom{n_i}{y_i} + y_i \log \left(\frac{p_i}{1 - p_i} \right) + n_i \log (1 - p_i) \end{aligned} \quad (4.40)$$

onde $\log(p_i/(1 - p_i)) = X\beta$. Lembrando que, para o modelo Binomial, $E[Y_i] = n_i p_i$ e $Var[Y_i] = n_i p_i (1 - p_i)$.

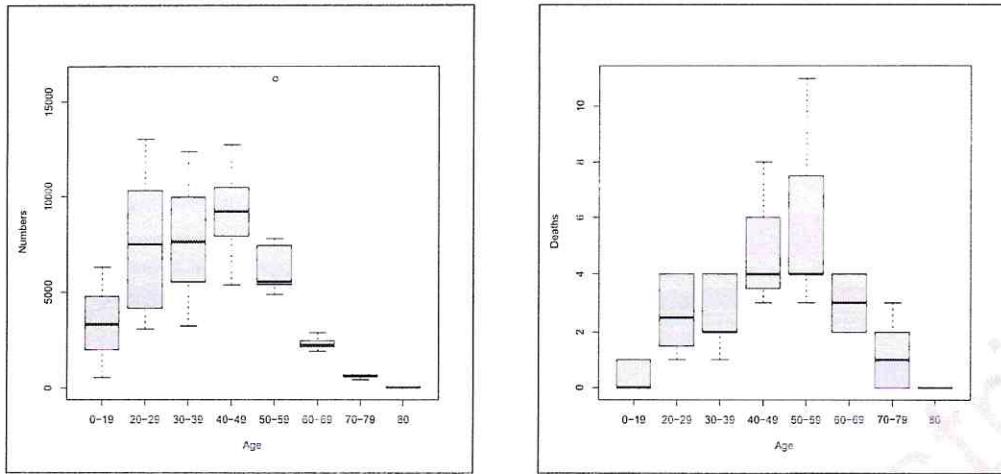
4.8.1 ESTUDO DE CASO: AVIATION DEATHS

A base de dados denominada *Aviation Deaths* é composta por 64 observações, referentes a mortes de pilotos de carreira em diferentes faixas etárias no período de 1992 a 1999 na Austrália. A base de dados contém as variáveis: ano de interesse (Year), número de pilotos registrados (Numbers), número de mortes (Deaths) e faixa etária dos pilotos (Age). A variável resposta é o número de mortes e a população de interesse é o número de pilotos registrados. As variáveis faixa etária e ano são as variáveis preditoras. As oito observações iniciais da base de dados são apresentadas a seguir.

```
> dados <- read.csv("aviationdeaths.dat", sep="")
> head(dados, n=8)
  Year Numbers Deaths  Age
1 1992      546      1 0-19
2 1992     3141      4 20-29
3 1992     5875      4 30-39
4 1992    12731      8 40-49
5 1992    16230      3 50-59
6 1992     2175      3 60-69
7 1992      418      0 70-79
8 1992      24       0   80
```

A Figura 4.13 mostra o comportamento do número de mortes e do número de pilotos para os diferentes anos de interesse. Vale destacar que a análise somente da variável número de mortes em relação às variáveis preditoras é inconsistente neste contexto, uma vez que um número maior de mortes pode estar associado a uma maior população de interesse, e não necessariamente ao efeito das variáveis preditoras. Portanto, nos problemas onde existem populações, que é o caso

da regressão Binomial, a análise exploratória deve ser feita utilizando como variável resposta as taxas brutas, $taxa_i = y_i/n_i$, como mostra a Figura 4.14.



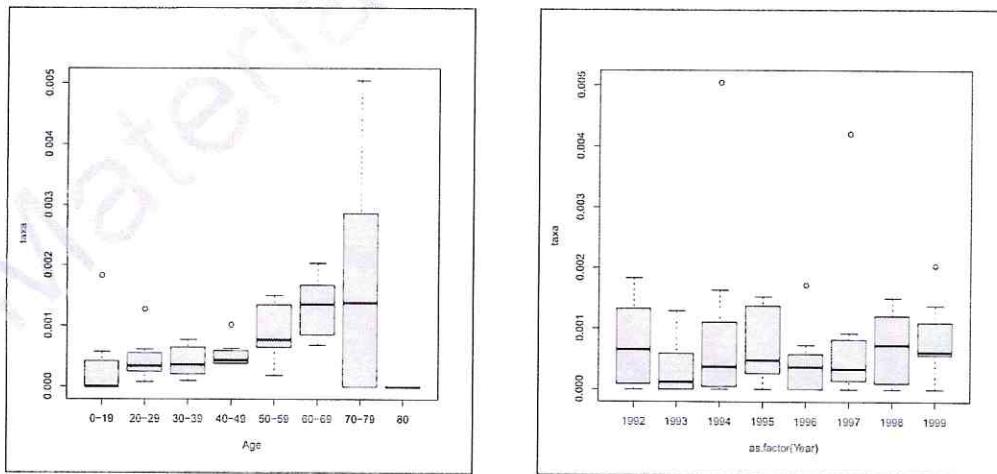
(a) Número de pilotos registrados para as diferentes faixas etárias.

(b) Número de mortes registradas para as diferentes faixas etárias.

Figura 4.13: Análise do número de mortes e das populações de interesse para as diferentes faixas etárias.

A Figura 4.14 mostra as taxas de morte para as variáveis faixa etária e ano de interesse. A Figura 4.14 (a) mostra um aumento das taxas brutas com o aumento da faixa etária, com exceção da última faixa etária (acima de 80 anos), que apresenta taxas brutas iguais a zero. Provavelmente, este fato se deve ao pequeno número de pilotos com idade igual ou acima de 80 anos que ainda trabalham. Com relação aos anos de interesse, a Figura 4.14 (b) evidencia um pequeno aumento das medianas das taxas com o aumento progressivo dos anos, com excessão do primeiro ano, 1992, que apresentou grande variabilidade e um valor alto para a mediana.

```
dados$taxa <- with(dados, Deaths/Numbers)
```



(a) Taxas brutas para as diferentes faixas etárias.

(b) Taxas brutas para os diferentes anos.

Figura 4.14: Análise das taxas brutas para as variáveis faixa etária e ano de interesse.

O ajuste do modelo de regressão Binomial é apresentado a seguir.

```
> modelo <- glm(cbind(Deaths, Numbers-Deaths) ~ as.factor(Year) + Age,
+                 family=binomial, data=dados)
> summary(modelo)

Call:
glm(formula = cbind(Deaths, Numbers - Deaths) ~ as.factor(Year) +
    Age, family = binomial, data = dados)

Deviance Residuals:
    Min      1Q   Median      3Q     Max 
-2.90895 -0.59837 -0.00015  0.45364  2.23942 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -9.19430   0.62059 -14.815 < 0.0000000000000002 *** 
as.factor(Year)1993 -0.26700   0.33225 -0.804   0.421624    
as.factor(Year)1994 -0.04044   0.31061 -0.130   0.896420    
as.factor(Year)1995  0.34756   0.28952  1.200   0.229954    
as.factor(Year)1996  0.04585   0.31520  0.145   0.884346    
as.factor(Year)1997  0.01970   0.32941  0.060   0.952302    
as.factor(Year)1998  0.46492   0.29605  1.570   0.116319    
as.factor(Year)1999  0.40325   0.32299  1.248   0.211852    
Age20-29          1.15539   0.61731  1.872   0.061254 .  
Age30-39          1.12504   0.61749  1.822   0.068461 .  
Age40-49          1.51847   0.60089  2.527   0.011503 *  
Age50-59          1.91395   0.59935  3.193   0.001406 ** 
Age60-69          2.37890   0.61364  3.877   0.000106 *** 
Age70-79          2.70317   0.66824  4.045   0.0000523 *** 
Age80             -13.14911  2439.11054 -0.005   0.995699   

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.139  on 63  degrees of freedom
Residual deviance: 57.231  on 49  degrees of freedom
AIC: 226.03

Number of Fisher Scoring iterations: 18
```

A inclusão da população de referência no código R consiste em especificar o número de mortos e o número de não-mortos na forma de uma variável resposta com duas colunas:

```
modelo <- glm(cbind(Deaths, Numbers-Deaths) ~ as.factor(Year) + Age,
               family=binomial, data=dados)
```

Em um primeiro instante, os resultados mostram que a variável ano (Year) não é estatisticamente significativa, pois nenhum dos valores-P é menor que 0.05 (5%). Uma alternativa é explorar a variável ano como uma variável numérica e ajustar novamente o modelo. Este fato se justifica a partir da análise exploratória, que indica uma leve tendência de aumento na mediana com o passar dos anos. Para isso a variável ano foi recodificada, como mostrado a seguir.

```
> dados$Year <- dados$Year - min(dados$Year)
> modelo <- glm(cbind(Deaths, Numbers-Deaths) ~ Year + Age,
+                 family=binomial, data=dados)
> summary(modelo)

Call:
glm(formula = cbind(Deaths, Numbers - Deaths) ~ Year + Age, family = binomial,
     data = dados)

Deviance Residuals:
    Min      1Q   Median      3Q     Max 
-2.6374 -0.6115 -0.1585  0.5389  2.4499 
```

```

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
(Intercept) -9.35001   0.59024 -15.858 < 0.000000000000002 ***
Year          0.07506   0.03535   2.124     0.033705 *
Age20-29     1.15003   0.61727   1.879     0.060203 .
Age30-39     1.13757   0.61729   1.843     0.065352 .
Age40-49     1.55596   0.59976   2.594     0.009478 **
Age50-59     1.98111   0.59642   3.322     0.000895 ***
Age60-69     2.42090   0.61262   3.952     0.0000776 ***
Age70-79     2.74172   0.66736   4.108     0.0000399 ***
Age80        -12.13173  1496.25660 -0.008     0.993531
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.139 on 63 degrees of freedom
Residual deviance: 61.557 on 55 degrees of freedom
AIC: 220.36

Number of Fisher Scoring iterations: 17

```

Os novos resultados indicam que a variável ano é estatisticamente significativa (valor-P < 0.05), bem como a variável faixa etária. A interpretação do modelo é realizada sobre a exponencial dos valores estimados, mostrados a seguir.

```

> round(as.matrix(exp(coef(modelo))), digits=7)
      [,1]
(Intercept) 0.0000861
Year         1.0779472
Age20-29    3.1900386
Age30-39    3.1191658
Age40-49    4.7396321
Age50-59    7.2507950
Age60-69    11.2560120
Age70-79    15.5137078
Age80       0.0000054

```

Considerando a faixa etária "< 20 anos" como a referência (intercepto), os resultados indicam que para cada aumento de um ano, a razão de chances da morte em relação a não morte aumenta em 7,79%. Indivíduos na faixa etária "20 – 29 anos" apresentam um razão de chance 3,19 vezes maior que a referência. Há um aumento do efeito da faixa etária na razão de chance para indivíduos com idade avançada, com excessão de indivíduos na faixa etária " ≥ 80 " anos, cujo valor-P é não significativo. Este resultado indica que esses indivíduos se comportam de maneira semelhante aos indivíduos de referência.

Finalmente, o valor-P do modelo e alguns valores da resposta estimada do número médio de mortes são apresentados a seguir. Os resultados demonstram coerência estatística entre o modelo escolhido e a distribuição empírica dos dados (valor-P > 0.05).

```

> 1 - pchisq(modelo$deviance, modelo$df.residual)
[1] 0.2530336
> dados$ajustado <- dados$Numbers * predict(modelo, type="response")
> head(dados, n=10)
  Year Numbers Deaths   Age      ajustado
1   0      546     1 0-19  0.04700625717320
2   0      3141     4 20-29  0.86247203900861
3   0      5875     4 30-39  1.57735751335840
4   0     12731     8 40-49  5.19314420861700
5   0     16230     3 50-59  10.12589741394087
6   0      2175     3 60-69  2.10583113201209
7   0      418      0 70-79  0.55758694754679
8   0      24       0 80      0.00000001112933
9   1      6278     0 0-19  0.58261129480452
10  1     13026     1 20-29  3.85546122626838

```

4.8.2 ESTUDO DE CASO: POTENCY

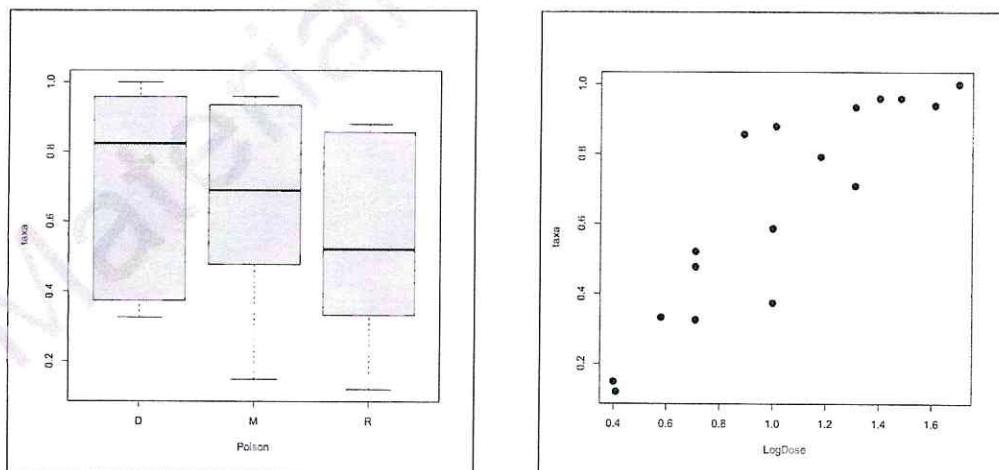
A base de dados denominada *Potency* (FINNEY; TATTERSFIELD, 1952) é composta por 17 observações, referentes ao número de insetos mortos (Kill) quando na aplicação de três possíveis tipos de veneno (Poison) em diferentes concentrações (LogDose). O número total de insetos (Number) submetidos à aplicação das diferentes substâncias também está disponível. A base de dados é apresentada a seguir.

	Kill	Number	Poison	LogDose
1	44	50	R	1.01
2	42	49	R	0.89
3	24	46	R	0.71
4	16	48	R	0.58
5	6	50	R	0.41
6	48	48	D	1.70
7	47	50	D	1.61
8	47	49	D	1.48
9	34	48	D	1.31
10	18	48	D	1.00
11	16	49	D	0.71
12	48	50	M	1.40
13	43	46	M	1.31
14	38	48	M	1.18
15	27	46	M	1.00
16	22	46	M	0.71
17	7	47	M	0.40

A partir da base de dados foram calculadas as taxas de morte.

```
dados$taxa <- with(dados, Kill/Number)
```

A Figura 4.15 mostra o comportamento das taxas de mortes para as diferentes substâncias utilizadas e a gráfico de dispersão entre as taxas e as diferentes concentrações utilizadas. Os resultados indicam a substância D como sendo a substância que apresenta as maiores taxas de morte, seguida pelas substâncias M e R. A Figura 4.15 (b) evidencia que o aumento da concentração provoca o aumento da taxa de mortes. É importante destacar o fato de que existem taxas próximas de 100%, ou seja, experimentos nos quais o número de insetos mortos é próximo ou igual ao número de insetos submetidos às diferentes substâncias (população de interesse).



(a) Taxas observadas para as diferentes substâncias aplicadas.

(b) Gráfico de dispersão das taxas brutas em função da concentração (LogDose).

Figura 4.15: Análise gráfica da base de dados *Potency*.

O resultado do ajuste do modelo e o valor-P do modelo é apresentado a seguir.

```
> modelo <- glm(cbind(Kill, Number-Kill) ~ Poison + LogDose,
+                 data=dados, family=binomial )
> summary(modelo)

Call:
glm(formula = cbind(Kill, Number - Kill) ~ Poison + LogDose,
     family = binomial, data = dados)

Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-1.7725 -1.0948  0.5153  1.4039  2.2419 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.8688    0.4201 -11.590 < 2e-16 ***
PoisonM      0.9123    0.2449   3.725 0.000196 *** 
PoisonR      1.6034    0.2656   6.036 1.58e-09 *** 
LogDose       4.8277    0.3395  14.222 < 2e-16 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

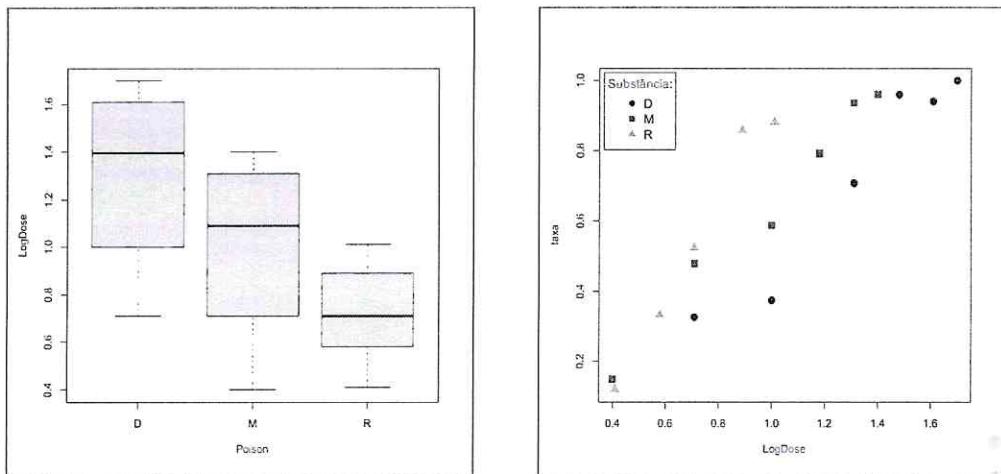
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 350.069  on 16  degrees of freedom
Residual deviance: 31.971  on 13  degrees of freedom
AIC: 98.955

Number of Fisher Scoring iterations: 4

> with(modelo, 1-pchisq(deviance, df=df.residual))
[1] 0.00242618
```

Os resultados indicam uma falta de consistência entre o modelo utilizado (Binomial) e a distribuição empírica dos dados, uma vez que o valor-P do modelo é menor que 0.05 (5%). Outro fato intrigante é a indicação da substância R como sendo a substância que apresenta o maior efeito na mortalidade dos insetos ($e^{1.6034} = 4.9699$). Este fato pode ser explicado a partir da análise das concentrações utilizadas para as diferentes substâncias, como mostra a Figura 4.16. No experimento analisado, as concentrações utilizadas para a substância D foram as maiores e, por consequência, as maiores taxas de mortes estão associadas a esta substância. Por outro lado, o resultado do modelo não indica a substância R como sendo a mais potente. Os resultados indicam que é possível utilizar a substância R em concentrações menores para obter um efeito semelhante ao observado para a substância D. Segundo a Figura 4.16 (b) mesmo utilizado em baixas concentrações, a substância R apresenta alguns poucos resultados com elevadas taxas de mortalidade.



(a) Variável concentração (LogDose) para os diferentes tipos de substâncias.

(b) Taxas de morte em função da concentração, para os diferentes tipos de substâncias.

Figura 4.16: Análise dos dados estratificada para os diferentes tipos de substâncias.

Vale destacar, que existe a possibilidade de interação entre as substâncias e a concentração, ou seja, dependendo da substância o efeito da concentração seria diferenciado. Para testar esta hipótese foi realizado o ajuste do modelo considerando a interação entre as variáveis tipos de veneno e concentração. Os resultados são apresentados a seguir.

```
> modelo <- glm(cbind(Kill, Number-Kill) ~ Poison * LogDose,
+                 data=dados, family=binomial )
> summary(modelo)

Call:
glm(formula = cbind(Kill, Number - Kill) ~ Poison * LogDose,
     family = binomial, data = dados)

Deviance Residuals:
    Min      1Q   Median      3Q      Max 
-1.77331 -0.67995  0.03622  1.00671  2.02531 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.4509    0.6104 -7.291 3.07e-13 *** 
PoisonM      1.0440    0.7795  1.339  0.1805    
PoisonR     -0.3878    0.8828 -0.439  0.6605    
LogDose       4.4602    0.5155  8.652 < 2e-16 *** 
PoisonM:LogDose -0.2244    0.7140 -0.314  0.7533    
PoisonR:LogDose  2.6079    1.0221  2.551  0.0107 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 350.069 on 16 degrees of freedom 
Residual deviance: 22.723 on 11 degrees of freedom 
AIC: 93.707

Number of Fisher Scoring iterations: 4

> with(modelo, 1-pchisq(deviance, df=df.residual))
[1] 0.01933496
```

Os resultados para o modelo de interação indicam uma interação entre a substância R e a dose utilizada. O valor-P do modelo apresentou um aumento em relação ao modelo sem a

interação. Considerando um nível de significância de 0.05 (5%), o resultado ainda aponta uma inconsistência entre o modelo Binomial e a distribuição empírica dos dados. A interpretação do modelo com interação pode ser facilitada considerando os efeitos das concentração para as diferentes substâncias. Caso a substância D seja utilizada, o efeito da dose pode ser descrito pela equação:

$$\text{Substância D: } \log\left(\frac{p_i}{1-p_i}\right) = -4.4509 + 4.4602 \times \text{LogDose}$$

Caso a substância M seja utilizada, o efeito da dose pode ser descrito pela equação:

$$\text{Substância M: } \log\left(\frac{p_i}{1-p_i}\right) = (-4.4509 + 1.0440) + (4.4602 - 0.2244) \times \text{LogDose}$$

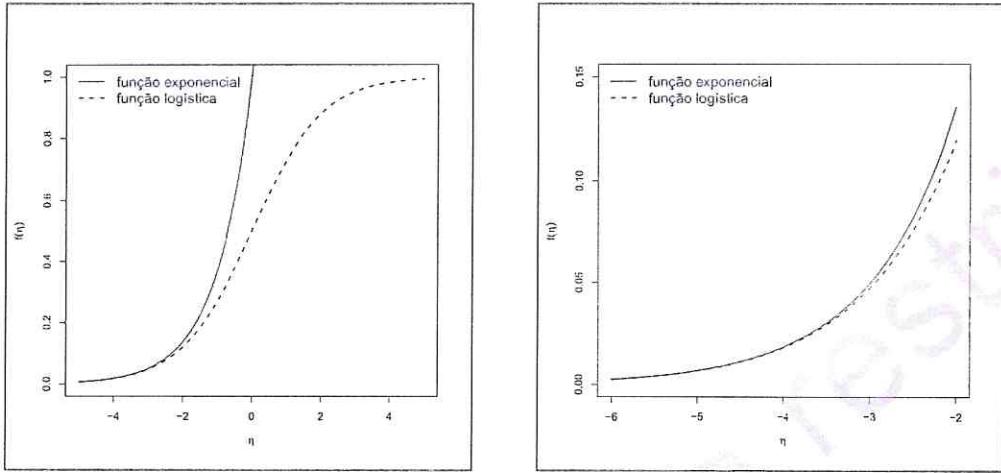
Caso a substância R seja utilizada, o efeito da dose pode ser descrito pela equação:

$$\text{Substância R: } \log\left(\frac{p_i}{1-p_i}\right) = (-4.4509 - 0.3878) + (4.4602 + 2.6079) \times \text{LogDose}$$

As equações mostram que, utilizando a substância R, o efeito da concentração é de $4.4602 + 2.6079 = 7.0681$, ou seja, para cada aumento de uma unidade na concentração da substância D a razão de chance de morte é multiplicada por $e^{7.0681} = 1173.92$. Portanto, a substância R apresenta o maior efeito nas taxas de mortalidade dos insetos para pequenas concentrações. É interessante destacar novamente a Figura 4.16 que indica que, no experimento, a substância R foi utilizada nas menores concentrações, se comparadas às concentrações utilizadas para as substâncias D e M.

4.9 O MODELO DE REGRESSÃO DE POISSON COMO APROXIMAÇÃO DO MODELO BINOMIAL

Em casos particulares, o modelo de regressão Binomial pode ser aproximado pelo modelo de regressão de Poisson. Na prática, isso se deve ao fato da função logística se comportar de forma semelhante à função exponencial quando ambas as funções são utilizadas para estimar respostas próximas de zero. Esta característica é ilustrada na Figura 4.17.



(a) Função exponencial versus função logística. (b) Função exponencial versus função logística para valores próximos de zero.

Figura 4.17: Função exponencial versus função logística.

Considerando o preditor linear: $\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$, caso seja possível verificar que a função logística pode ser aproximada pela função exponencial, é possível assumir que o número médio de casos estimado pelo modelo Binomial pode ser reescrito como:

$$\begin{aligned} E[Y] &= n_i \times p_i \\ &= n_i \times \frac{e^\eta}{1 + e^\eta} \\ &\approx n_i \times e^\eta \end{aligned} \tag{4.41}$$

Ou seja, o comportamento estatístico do número esperado de casos passa a ser interpretado segundo o modelo de regressão de Poisson: supondo um modelo com uma única variável preditora (x_1), para cada aumento de uma unidade na variável preditora a média da variável resposta seria multiplicada por e^{β_1} :

$$\begin{aligned} E[Y|x_1 + 1] &\approx n_i \times e^{\beta_0 + \beta_1(x_1 + 1)} \\ &\approx n_i \times e^{\beta_0 + \beta_1 x_1} \times e^{\beta_1} \end{aligned} \tag{4.42}$$

No caso do uso da aproximação do modelo Binomial pelo modelo de Poisson, a população (n_i) pode ser incorporada ao preditor linear (η) formando uma variável preditora à qual não está associado nenhum parâmetro, denominada de termo de *offset*.

$$\begin{aligned} E[Y] &\approx n_i \times e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k} \\ &= e^{\log(n_i) + \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k} \end{aligned} \tag{4.43}$$

onde $offset = \log(n_i)$.

4.9.1 ESTUDO DE CASO: AVIATION DEATHS

Na prática, a aproximação do modelo de regressão Binomial pelo modelo de regressão de Poisson resulta em valores estimados muito próximos, incluse deviance.

```
> modelo <- glm(Deaths ~ offset(log(Numbers)) + Year + Age,
+                 family=poisson, data=dados)
> summary(modelo)

Call:
glm(formula = Deaths ~ offset(log(Numbers)) + Year + Age, family = poisson,
     data = dados)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-2.6368 -0.6112 -0.1584  0.5386  2.4491 

Coefficients:
            Estimate Std. Error z value     Pr(>|z|)    
(Intercept) -9.35932   0.59019 -15.859 < 0.000000000000002 *** 
Year          0.07500   0.03533   2.123     0.033772 *  
Age20-29     1.15978   0.61722   1.879     0.060240 .  
Age30-39     1.13732   0.61724   1.843     0.065391 .  
Age40-49     1.55553   0.59971   2.594     0.009492 ** 
Age50-59     1.98040   0.59637   3.321     0.000898 *** 
Age60-69     2.41972   0.61254   3.950     0.0000781 *** 
Age70-79     2.74003   0.66718   4.107     0.0000401 *** 
Age80        -12.69976  1987.86347 -0.006     0.994903  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 118.063  on 63  degrees of freedom
Residual deviance: 61.515  on 55  degrees of freedom
AIC: 220.36

Number of Fisher Scoring iterations: 16

> with(modelo, 1-pchisq(deviance, df.residual))
[1] 0.2542289
```

Se comparados com os resultados apresentados na sessão 4.8.1, os resultados apresentados utilizando o modelo de regressão de Poisson são muito próximos dos resultados utilizando o modelo de regressão Binomial. Vale destacar que o modelo de regressão de Poisson, neste caso, é uma aproximação. A principal vantagem desta aproximação é a interpretação dos coeficientes segundo a função exponencial, como apresentado para o modelo de regressão de Poisson.

4.9.2 ESTUDO DE CASO: POTENCY

Os resultados para a base de dados *Potency* utilizando a aproximação do modelo de regressão de Poisson são apresentados a seguir.

```
> modelo <- glm(Kill ~ offset(log(Number)) + Poison * LogDose,
+                 data=dados, family=poisson )
> summary(modelo)

Call:
glm(formula = Kill ~ offset(log(Number)) + Poison * LogDose,
     family = poisson, data = dados)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.3082 -0.5904  0.1314  0.3504  1.1158 
```

```

Coefficients:
            Estimate Std. Error z value    Pr(>|z|)
(Intercept) -2.035890  0.344822 -5.904 0.00000000354 ***
PoissonM    -0.007741  0.465181 -0.017   0.98672
PoissonR    -0.730991  0.524648 -1.393   0.16353
LogDose     1.244544  0.235919  5.275 0.00000013253 ***
PoissonM:LogDose 0.246048  0.353004  0.697   0.48579
PoissonR:LogDose 1.513752  0.517280  2.926   0.00343 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 127.7119 on 16 degrees of freedom
Residual deviance: 9.9556 on 11 degrees of freedom
AIC: 108.96

Number of Fisher Scoring iterations: 4

> with(modelo, 1-pchisq(deviance, df=df.residual))
[1] 0.5343858

```

Neste caso, diferente dos resultados apresentados na seção 4.8.2, há uma diferença expressiva entre os resultados utilizando o modelo de regressão Binomial e o modelo de regressão de Poisson. Mesmo na presença de um *valor-P* de 0.53438, indicando que a distribuição empírica dos dados se comporta segundo modelo de regressão de Poisson, o resultado está equivocado. Neste caso, uma vez que os valores estimados dos coeficientes são diferentes, a aproximação pelo modelo de regressão de Poisson não deve ser aplicada.

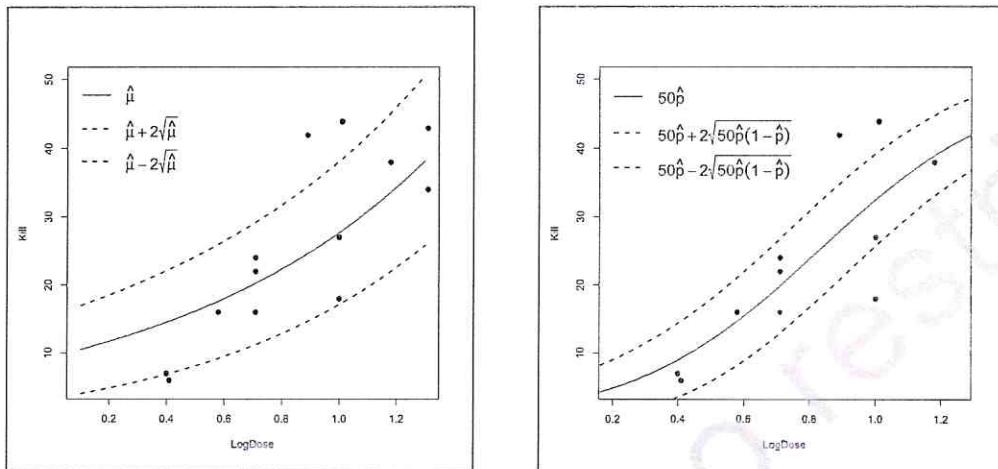
Em geral, a aproximação do modelo de regressão Binomial pelo modelo de regressão de Poisson ocorre na presença de eventos raros em grandes populações. É o caso, por exemplo, da base de dados *Aviation Deaths* na qual o número de mortes de pilotos é muito pequena se comparada à população de pilotos. O mesmo não ocorre para a base de dados *Potency* na qual, para alguns registros, o número de insetos mortos é muito próximo ao número de insetos (população) submetida ao efeito do veneno. No primeiro caso, os valores das taxas brutas são muito reduzidos o que permite a aproximação da função logística pela função exponencial, conforme ilustrado na Figura 4.17. No segundo caso, a aproximação da função logística pela função exponencial não é válida.

4.9.3 COMENTÁRIOS FINAIS SOBRE BINOMIAL VERSUS POISSON PARA A BASE DE DADOS Potency

Com o objetivo de esclarecer a *falta de ajuste* do modelo Binomial frente ao modelo de Poisson para a base de dados *Potency*, algumas simplificações serão feitas na base de dados. A título de ilustração, inicialmente será considerada que a população de insetos é igual a 50. Também será considerada somente a variável preditora concentração (LogDose). O problema, portanto, pode ser visualizado em um gráfico de dispersão. O objetivo é detalhar o fato do modelo Binomial, *teoricamente correto*, ter obtido um resultado inferior ao modelo errado (Poisson). A diferença, neste caso, está associada tanto à equação da média mas principalmente à estrutura de variância dos modelos.

Para o modelo de Poisson, assumindo uma variável preditora x e uma população N , a equação da média é definida por $\mu = \exp(\log(N) + \beta_0 + \beta_1 \cdot x)$ e a equação de variância é $Var(Y) = \mu$. Para o modelo Binomial, a equação da média é definida por $\mu = N \cdot p$ e a equação de variância é $Var(Y) = N \cdot p(1 - p)$, onde $p = \frac{\exp(\beta_0 + \beta_1 \cdot x)}{1 + \exp(\beta_0 + \beta_1 \cdot x)}$. Para fins de ilustração, vamos utilizar as estruturas de média e variância para *desenhar* intervalos de predição para as observações. Vamos considerar uma amplitude de dois desvios em relação à média, $\mu \pm 2 \cdot \sigma$. Utilizando os resultados do ajuste dos modelos Binomial e Poisson, assumindo as simplificações descritas anteriormente, a Figura 4.18 apresenta os resultados obtidos. A partir dos resultados, desejamos distinguir qual

o modelo apresenta a melhor *cobertura* dos dados, ou seja, qual o intervalo preditivo que contém a maior proporção de pontos em seu interior. O resultado mostra que o modelo de Poisson apresenta o intervalo preditivo mais largo. Portanto, o modelo de Poisson apresenta a melhor cobertura dos dados e a melhor estatística de ajuste. Entretanto, a média do modelo de Poisson aumenta exponencialmente com a concentração do veneno e, proporcionalmente, a sua variância também está aumentando. Como consequência, o modelo de Poisson pode estimar um número de insetos mortos maior que a população de interesse, fornecendo respostas equivocadas para o problema.



(a) Resultado do ajuste do modelo de Poisson. (b) Resultado do ajuste do modelo Binomial.

Figura 4.18: Comparação entre os ajustes do modelo de Poisson e do modelo Binomial para a base de dados Potency, supondo N=50.

Portanto, o problema (Potency) se beneficiou de um modelo com estruturas de média e variância *superdimensionadas*. Por definição, o problema ainda está inserido no contexto de uma distribuição Binomial. Entretanto, para a distribuição Binomial, a estrutura de variância é limitada. No contexto da família exponencial, é possível ajustar a estrutura de variância. Lembrando que para a família exponencial a variância da variável resposta é definida por $Var(Y) = a(\phi)b''(\theta)$, ou ainda $Var(Y) = \phi b''(\theta)$. Para a distribuição Binomial *tradicional*, o parâmetro ϕ é igual a 1 ($\phi = 1$). Mas, o mesmo pode ser também estimado, gerando um novo modelo denominado *quasi-Binomial*.

O ajuste do modelo *quasi-Binomial* é mostrado a seguir.

```
> modelo <- glm(cbind(Kill, Number-Kill) ~ LogDose + Poison, family=quasibinomial, data=dados)
> summary(modelo)

Call:
glm(formula = cbind(Kill, Number - Kill) ~ LogDose + Poison,
family = quasibinomial, data = dados)

Deviance Residuals:
Min      1Q  Median      3Q     Max 
-1.7725 -1.0948  0.5153  1.4039  2.2419 

Coefficients:
Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.8688    0.6426 -7.576 4.04e-06 ***
LogDose      4.8277    0.5193  9.297 4.16e-07 ***
PoisonM     0.9123    0.3747  2.435  0.03005 *  
PoisonR     1.6034    0.4063  3.946  0.00167 ** 
---

```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
(Dispersion parameter for quasibinomial family taken to be 2.340001)
```

Como pode ser verificado no ajuste do novo modelo, o parâmetro de dipersão foi estimado como $\phi = 2.340001$. Portanto, neste modelo, a estrutura de variância é praticamente duas vezes maior que o modelo Binomial tradicional, $Var(Y) = \phi \cdot N \cdot p(1 - p)$.

Materiais de uso restritivo

4.10 O MODELO DE REGRESSÃO GAMA

Uma variável aleatória contínua positiva Y ($Y > 0$) com distribuição Gama possui a seguinte densidade de probabilidade:

$$f_Y(y) = \frac{y^{-1}}{\Gamma(\nu)} \left(\frac{y\nu}{\mu} \right)^{\nu} e^{-y\nu/\mu} \quad (4.44)$$

onde μ é o parâmetro de média, $E(Y) = \mu$, ν é o parâmetro de dispersão, $Var(Y) = \frac{\mu^2}{\nu}$, e $\Gamma(\cdot)$ é a função gama, apresentada na equação 4.45.

$$\Gamma(u) = \int_0^{\infty} x^{u-1} e^{-x} dx \quad (4.45)$$

Diferente das distribuições apresentadas, a relação entre o parâmetro de média e o preditor linear $x_i^T \beta$, utilizando a ligação canônica, é na forma:

$$\mu_i = -\frac{1}{x_i^T \beta} \quad (4.46)$$

Portanto, neste caso em particular, é preferível definir um modelo de regressão na forma:

$$\mu_i = e^{x_i^T \beta} \quad (4.47)$$

garantindo que a média do modelo será sempre positiva e que a interpretação dos parâmetros do modelo é semelhante à interpretação do modelo de regressão de Poisson.

A principal característica de um modelo de regressão Gama é a sua estrutura de variância, $Var(Y|x_i^T) \propto \mu_{(x_i^T)}^2$. Ou seja, a variância do modelo é proporcional ao quadrado da média. As propriedades do modelo Gama são semelhantes às propriedades do modelo de regressão log-linear. Razão pela qual há, em geral, preferência para o ajuste do modelo log-linear.

O modelo log-linear pode ser definido na forma:

$$\log Y_i = x_i^T \beta + \epsilon_i \quad (4.48)$$

onde $\epsilon_i \sim Normal(0, \sigma^2)$. A partir da equação 4.48, as propriedades de esperança e variância podem ser avaliadas.

$$\begin{aligned} Y_i &= \exp(x_i^T \beta + \epsilon_i) \\ &= \exp(x_i^T \beta) \times e^{\epsilon_i} \\ &= \mu_i \cdot e^{\epsilon_i} \end{aligned} \quad (4.49)$$

Então, $E(Y_i)$ é definida por:

$$\begin{aligned} E(Y_i) &= E(\mu_i \cdot e^{\epsilon_i}) \\ &= \mu_i \cdot E(e^{\epsilon_i}) \\ &= \mu_i \cdot e^{\sigma^2/2} \end{aligned} \quad (4.50)$$

e a variância, $Var(Y_i)$, é definida por:

$$\begin{aligned} Var(Y_i) &= Var(\mu_i \cdot e^{\epsilon_i}) \\ &= \mu_i^2 \cdot Var(e^{\epsilon_i}) \\ &= \mu_i^2 \cdot (e^{\sigma^2} - 1) \cdot e^{\sigma^2} \end{aligned} \quad (4.51)$$

onde e^{ϵ_i} é uma variável aleatória log-normal (CASELLA; BERGER, 2002).

É importante destacar que, embora existam semelhanças entre o modelo log-Normal e o modelo Gama, existem importantes diferenças que podem produzir resultados distintos. As diferenças são apresentadas na tabela 4.2.

Tabela 4.2: Diferenças entre o modelo Gama e o modelo Log-Linear

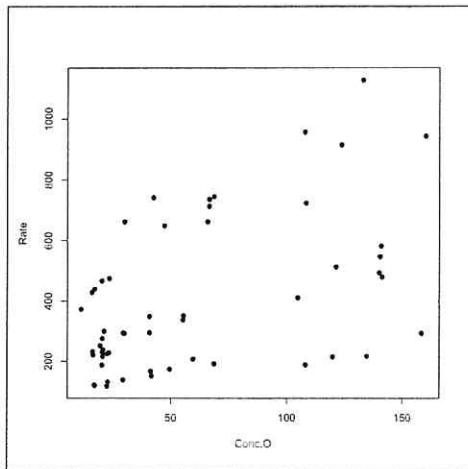
Modelo de Regressão	Esperança	Variância
Gama	$E[Y] = \mu$	$Var[Y] = \phi \cdot \mu^2$
Log-Linear	$E[Y] = \mu \cdot e^{\sigma^2/2}$	$Var[Y] = \mu^2 \cdot (e^{\sigma^2} - 1) \cdot e^{\sigma^2}$

onde $\mu = \mathbf{x}^T \boldsymbol{\beta}$. Como pode ser visto na tabela 4.2, a diferença entre o modelo Gama e o modelo Log-Normal é a maneira como o parâmetro de dispersão (ϕ para o modelo Gama e σ^2 para o modelo log-linear) está associado às variâncias e médias da variável resposta Y .

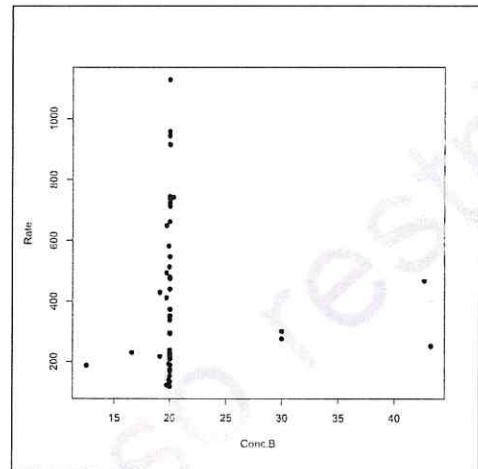
4.10.1 ESTUDO DE CASO: TAXA DE OXIDAÇÃO DE BENZENO

A base de dados Taxa de Oxidação de Benzeno (*oxidation rate of benzene*) contém 54 observações referentes à taxa inicial de oxidação de benzeno sobre um catalisador de óxido de vanádio e disponibiliza quatro potenciais variáveis preditoras: concentração de oxigênio (Conc.O), concentração de Benzeno (Conc.B), Temperatura do processo (Temp) e o número observado de moles de oxigênio consumido por mole de benzeno (O.per.B).

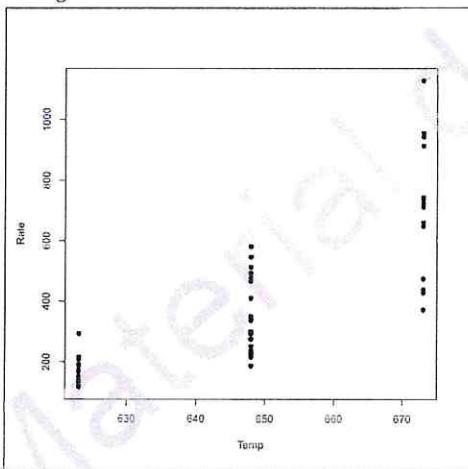
A figura 4.19 apresenta os gráficos de dispersão entre a variável resposta e as variáveis preditoras. É possível perceber nas figuras 4.19 (a), 4.19 (c) e 4.19 (d) o aumento da dispersão dos dados com o aumento das variáveis preditoras, indicando um possível modelo heterocedástico.



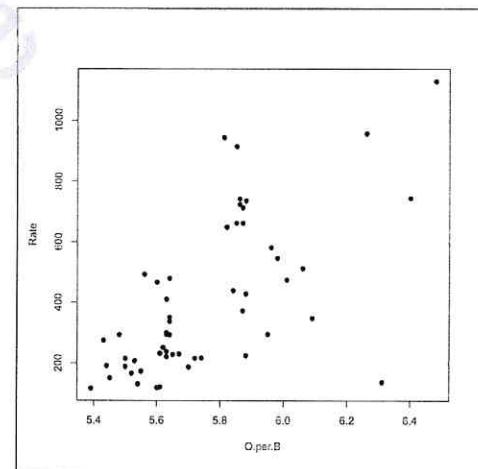
(a) Taxa de oxidação em função da concentração de oxigênio.



(b) Taxa de oxidação em função da concentração de benzeno.



(c) Taxa de oxidação em função da concentração da temperatura.



(d) Taxa de oxidação em função da razão oxigênio/benzeno.

Figura 4.19: Gráficos de dispersão da variável resposta em função das variáveis preditoras.

Embora exista evidência para o ajuste de um modelo heterocedástico, será analisado inicialmente o ajuste de um modelo linear padrão (homocedástico), mostrado a seguir.

```
> modelo <- lm(Rate ~ Conc.O + Conc.B + Temp + O.per.B, data=dados)
> summary(modelo)
```

```

Call:
lm(formula = Rate ~ Conc.O + Conc.B + Temp + O.per.B, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max 
-145.38  -60.35  -21.46   78.45  204.88 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6798.2122   414.1358 -16.415 < 2e-16 ***
Conc.O        2.4688    0.2606   9.472 1.18e-12 ***
Conc.B        4.5621    2.4733   1.845  0.07116 .  
Temp         9.1860    0.7819  11.748 7.34e-16 ***
O.per.B      173.0007   61.5987   2.809  0.00713 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 85.67 on 49 degrees of freedom
Multiple R-squared:  0.8922,    Adjusted R-squared:  0.8834 
F-statistic: 101.4 on 4 and 49 DF,  p-value: < 2.2e-16

```

O modelo linear ajustado apresenta um coeficiente de determinação ajustado $R^2_{adj} = 88.34\%$, indicando que o modelo explica uma parcela considerável da variabilidade da resposta. A análise dos resíduos do modelo é apresentada da figura 4.20. Comparando o resultado obtido com a figura 2.3, é possível identificar uma possível não linearidade no modelo ajustado, bem como baixa evidência de heterocedasticidade nos resíduos do modelo linear. Neste caso, será explorado o ajuste do modelo Log-linear.

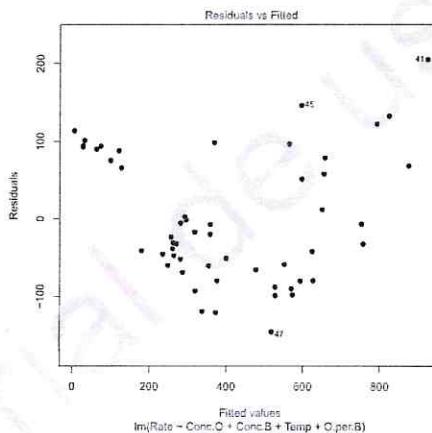


Figura 4.20: Análise dos resíduos do modelo de regressão linear.

O ajuste do modelo Log-linear é apresentado a seguir.

```

> modelo <- lm(log(Rate) ~ Conc.O + Conc.B + Temp + O.per.B, data=dados)
> summary(modelo)

Call:
lm(formula = log(Rate) ~ Conc.O + Conc.B + Temp + O.per.B, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.34737 -0.08976 -0.00199  0.09276  0.27813 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.271e+01  6.475e-01 -19.621 < 2e-16 ***
Conc.O       6.011e-03  4.075e-04  14.750 < 2e-16 ***

```

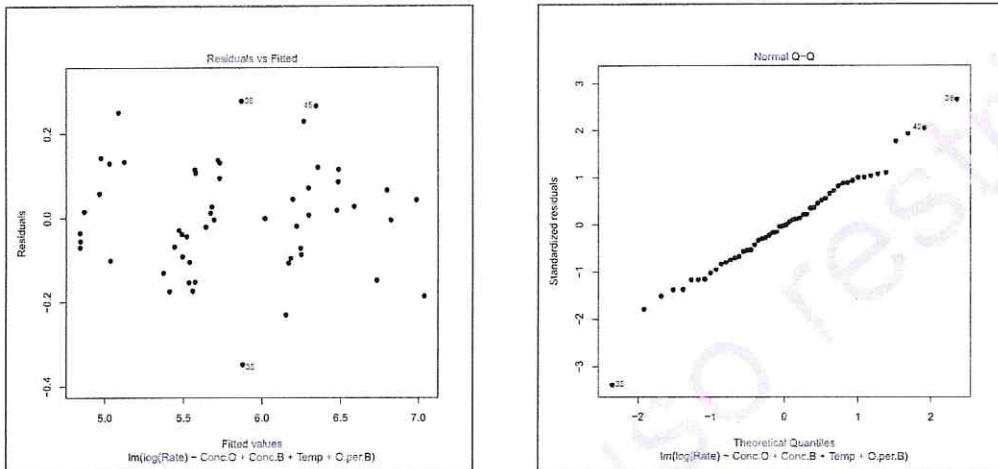
```

Conc.B      1.572e-02  3.867e-03   4.064  0.000174 ***
Temp       2.598e-02  1.223e-03  21.249 < 2e-16 ***
O.per.B     1.690e-01  9.631e-02   1.755  0.085575 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1339 on 49 degrees of freedom
Multiple R-squared:  0.9553,    Adjusted R-squared:  0.9516
F-statistic: 261.8 on 4 and 49 DF,  p-value: < 2.2e-16

```

A figura 4.21 mostra a análise dos resíduos do modelo Log-linear ajustado. Os resultados mostram que os resíduos apresentam um comportamento homocedástico (ver figura 4.21 (a)) e se aproximam de uma distribuição normal (ver figura 4.21 (b)).



(a) Análise gráfica dos resíduos do modelo Log-linear.

(b) Análise de normalidade dos resíduos.

Figura 4.21: Análise dos resíduos do ajuste do modelo Log-linear.

Finalmente, o ajuste do modelo Gama é apresentado a seguir:

```

> modelo <- glm(Rate ~ Conc.O + Conc.B + Temp + O.per.B, data=dados,
+                  family=Gamma(link=log))
> summary(modelo)

Call:
glm(formula = Rate ~ Conc.O + Conc.B + Temp + O.per.B, family = Gamma(link = log),
     data = dados)

Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-0.36203 -0.09668 -0.00819  0.08651  0.26855 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.276e+01  6.455e-01 -19.765 < 2e-16 ***
Conc.O       5.984e-03  4.062e-04  14.730 < 2e-16 ***
Conc.B       1.699e-02  3.855e-03   4.407 5.7e-05 ***
Temp        2.605e-02  1.219e-03  21.379 < 2e-16 ***
O.per.B      1.667e-01  9.601e-02   1.736  0.0888 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.01782698)

Null deviance: 19.60466 on 53 degrees of freedom
Residual deviance: 0.87634 on 49 degrees of freedom
AIC: 571.16

```

Number of Fisher Scoring iterations: 4

É importante destacar as semelhanças entre os resultados apresentados pelo modelo Log-linear e os resultados apresentados para o modelo Gama. Como mencionado, ambos os modelos definem uma estrutura de variância proporcional ao quadrado da média. Na prática, nem sempre o resultado para o ajuste do modelo Gama é semelhante ao resultado do ajuste do modelo Log-linear. É aconselhável a comparação dos dois modelos. Para o modelo Gama ajustado, o valor-P do modelo é:

```
> with(modelo, 1-pchisq(deviance, df.residual))
[1] 1
```

Material de apoio restrito

4.11 O MODELO DE REGRESSÃO MULTINOMIAL

O modelo de regressão multinomial é aplicado em problemas nos quais a variável resposta é categórica. Mas, diferente do modelo logístico, a variável resposta possui mais de 2 (duas) categorias, como será ilustrado no exemplo a seguir. A base de dados *shopping.csv* possui como variável resposta o registro da compra de três possíveis marcas (brand), $Y \in \{A, B, C\}$. Para prever a compra estão disponíveis as variáveis preditoras gênero (female) e idade (age). Caso um indivíduo não realize uma compra, esta informação poderia ser acrescentada na base de dados como, por exemplo, uma nova marca $Y = D$.

```
> mydata <- read.csv("shopping.csv")
> head(mydata)
  brand female age
1     A      0  24
2     A      0  26
3     A      0  26
4     A      1  27
5     A      1  27
6     C      1  27
```

A princípio, seria possível realizar uma adaptação do problema para o modelo logístico. Para isso, a variável resposta poderia ser decomposta em três variáveis respostas, conforme apresentado na tabela a seguir:

Y	Y^A	Y^B	Y^C
A	1	0	0
B	0	1	0
C	0	0	1

onde, caso $Y = A$ então $Y^A = 1$, $Y^B = 0$ e $Y^C = 0$ e assim por diante. Esta abordagem é muito semelhante à codificação *duminy* de variáveis preditoras, conforme apresentado na tabela a seguir.

x	x^A	x^B	x^C
A	1	0	0
B	0	1	0
C	0	0	1

Na sequência, seria possível ajustar três modelos de regressão logística, um para cada nova variável resposta Y^A , Y^B e Y^C , obtendo então, três diferentes estimativas para $P(Y^A = 1)$, $P(Y^B = 1)$ e $P(Y^C = 1)$:

$$P(Y^A = 1) = \pi_A = \frac{e^{\beta_0^A + \beta_1^A x_1 + \beta_2^A x_2}}{1 + e^{\beta_0^A + \beta_1^A x_1 + \beta_2^A x_2}} \quad (4.52)$$

$$P(Y^B = 1) = \pi_B = \frac{e^{\beta_0^B + \beta_1^B x_1 + \beta_2^B x_2}}{1 + e^{\beta_0^B + \beta_1^B x_1 + \beta_2^B x_2}} \quad (4.53)$$

$$P(Y^C = 1) = \pi_C = \frac{e^{\beta_0^C + \beta_1^C x_1 + \beta_2^C x_2}}{1 + e^{\beta_0^C + \beta_1^C x_1 + \beta_2^C x_2}} \quad (4.54)$$

A solução apresenta nas equações 4.52, 4.53 e 4.54 seria uma solução razoável para aqueles que só conhecem o modelo de regressão logístico. Entretanto, o mesmo apresenta limitações. A primeira limitação é o fato de que não é garantida que a soma das probabilidades estimadas por cada modelo é a unidade, ou seja, $\pi_A + \pi_B + \pi_C \neq 1$.

Utilizando alguns princípios já apresentados para o modelo logístico, podemos definir o modelo multinomial para o exemplo apresentado. Inicialmente, é necessário definir a restrição de que a soma das probabilidades estimadas pelo modelo deve ser igual a 1:

$$\pi_A + \pi_B + \pi_C = 1 \quad (4.55)$$

Pela equação 4.55 é possível verificar que é necessário definir equações de regressão para apenas duas das categorias, uma vez que a probabilidade da categoria de referência pode ser calculada a partir das demais. Por exemplo, suponha que π_A é definida como a probabilidade de referência, então $\pi_A = 1 - (\pi_B + \pi_C)$. Segundo, podemos redefinir razões de chance para as demais probabilidades em relação à probabilidade de referência:

$$\frac{P(Y^B = 1)}{P(Y^A = 1)} = \frac{\pi_B}{\pi_A} = e^{\beta_0^B + \beta_1^B x_1 + \beta_2^B x_2} \quad (4.56)$$

$$\frac{P(Y^C = 1)}{P(Y^A = 1)} = \frac{\pi_C}{\pi_A} = e^{\beta_0^C + \beta_1^C x_1 + \beta_2^C x_2} \quad (4.57)$$

Então, considerando as equações 4.55, 4.56 e 4.57 é possível definir um sistema de equações com relação às quantidades π_A , π_B e π_C :

$$\begin{cases} \pi_A + \pi_B + \pi_C = 1 \\ \frac{\pi_B}{\pi_A} = e^{\beta_0^B + \beta_1^B x_1 + \beta_2^B x_2} \\ \frac{\pi_C}{\pi_A} = e^{\beta_0^C + \beta_1^C x_1 + \beta_2^C x_2} \end{cases} \quad (4.58)$$

cuja solução é:

$$\begin{cases} \pi_A = \frac{1}{1 + e^{\eta_B} + e^{\eta_C}} \\ \pi_B = \frac{e^{\eta_B}}{1 + e^{\eta_B} + e^{\eta_C}} \\ \pi_C = \frac{e^{\eta_C}}{1 + e^{\eta_B} + e^{\eta_C}} \end{cases} \quad (4.59)$$

onde $\eta_B = \beta_0^B + \beta_1^B x_1 + \beta_2^B x_2$ e $\eta_C = \beta_0^C + \beta_1^C x_1 + \beta_2^C x_2$ são os preditores lineares (η_B e η_C). Neste caso, não há a necessidade de definir um preditor linear para a classe A pois, como mencionado, $\pi_A = 1 - (\pi_B + \pi_C)$. De forma sucinta, uma variável resposta que apresenta k categorias necessita de $k - 1$ preditores lineares. Os cálculos realizados podem ser facilmente aplicados a variáveis respostas com mais de três dimensões.

4.11.1 INTERPRETAÇÃO DOS PARÂMETROS DO MODELO MULTINOMIAL

É importante destacar que o modelo logístico, como apresentado na seção 4.7, é um caso particular do modelo multinomial. No modelo logístico, existem somente duas classes A (ou $Y = 0$) e B (ou $Y = 1$). Então, tornando a classe A ($Y = 0$) como a referência, a razão de chance do modelo logístico é:

$$\frac{P(Y = B)}{P(Y = A)} = \frac{P(Y = 1)}{P(Y = 0)} = \frac{P(Y = 1)}{1 - P(Y = 1)} = e^{\beta_0^B + \beta_1^B x_1 + \beta_2^B x_2} \quad (4.60)$$

dessa forma, a interpretação dos coeficientes do modelo consiste em aumentar ou reduzir a razão de chance de ocorrência da classe B em relação à classe de referência A , dependendo do sinal positivo ou negativo do coeficiente. Semelhante ao modelo logístico, a interpretação do valor da exponencial do coeficiente é a mais adequada.

Utilizando o exemplo apresentado no início desta seção, o modelo multinomial pode ser ajustado utilizando o código a seguir.

```

> library(mlogit)
> mldata <- mlogit.data(mydata, choice="brand", shape="wide")
> mlogit.model <- mlogit(brand ~ 1|female + age, data = mldata, reflevel="A")
> summary(mlogit.model)

Call:
mlogit(formula = brand ~ 1 | female + age, data = mldata, reflevel = "A",
method = "nr", print.level = 0)

Frequencies of alternatives:
A      B      C
0.28163 0.41769 0.30068

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 0.00158
successive function values within tolerance limits

Coefficients :
Estimate Std. Error t-value Pr(>|t|)
B:(intercept) -11.774478 1.774612 -6.6350 3.246e-11 ***
C:(intercept) -22.721201 2.058028 -11.0403 < 2.2e-16 ***
B:female       0.523813 0.194247  2.6966  0.007004 **
C:female       0.465939 0.226090  2.0609  0.039316 *
B:age          0.368201 0.055003  6.6942 2.169e-11 ***
C:age          0.685902 0.062627 10.9523 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -702.97
McFadden R^2:  0.11676
Likelihood ratio test : chisq = 185.85 (p.value = < 2.22e-16)

```

Com base nos resultados apresentados e utilizando a classe, ou melhor, a marca *A* como a marca de referência é possível concluir que: (a) a razão de chance de um indivíduo do sexo feminino de adquirir a marca *B* em relação à marca *A* é de $e^{0.523813} = 1.688453$, ou seja, um indivíduo do sexo feminino é mais propenso a adquirir um produto da marca *B* em relação à marca *A*; (b) a razão de chance de um indivíduo do sexo feminino de adquirir a marca *C* em relação à marca *A* é de $e^{0.465939} = 1.59351$, ou seja, um indivíduo do sexo feminino é mais propenso a adquirir um produto da marca *C* em relação à marca *A*. Uma análise dos resultados encontrados nos itens (a) e (b), considerando que a referência nos dois casos é a mesma, permite concluir que indivíduos do sexo feminino têm preferências de compra para a marca *B*, seguida pela marca *C*, seguida pela marca *A*. Na prática, um vendedor poderia trazer para os clientes, ou apresentar no mostruário, com maior evidência, os itens *B* e *C*. Ou colocar a marca *A* em promoção!

Com relação à faixa etária (*age*), os resultados mostram que: (a) para cada aumento de uma unidade na idade do indivíduo ocorre um aumento na razão de chance de $e^{0.368201} = 1.445132$ (44,5%) de um indivíduo adquirir a marca *B* em relação à marca *A*; (b) para cada aumento de uma unidade na idade do indivíduo ocorre uma aumento na razão de chance de $e^{0.685902} = 1.985562$ (98,6%) de um indivíduo adquirir a marca *C* em relação à marca *A*. Avaliando os resultados encontrados nos itens (a) e (b), considerando que a referência nos dois casos é a mesma, é possível concluir que indivíduos de idade mais avançada têm preferência de compra pela marca *C*, seguida da marca *B*, seguida pela marca *A*.

O modelo ajustado pode ser utilizado para estimar as probabilidades de compra das três marcas *A*, *B* e *C*, para um indivíduo do sexo masculino (*female* = 0) com idade de 35 anos (*age* = 35).

```
> saída ← predict(mlogit.modelo, newdata=data.frame(female=rep(0, 3), age=rep(35, 3)))
> saída
A          B          C
0.1305800 0.3972399 0.4721801
```

4.11.2 O ALGORITMO DE ESTIMAÇÃO DO MODELO MULTINOMIAL

Considerando o modelo multinomial com três classes ($Y = A, B, C$), a distribuição de probabilidade de Y pode ser escrita como:

$$P(Y_i^A = y_i^A, Y_i^B = y_i^B, Y_i^C = y_i^C) = (\pi_i^A)^{y_i^A} \times (\pi_i^B)^{y_i^B} \times (\pi_i^C)^{y_i^C} \quad (4.61)$$

Uma vez definida a classe de referência A , temos que $y_i^A = 1 - (y_i^B + y_i^C)$ e $\pi_A = 1 - (\pi_B + \pi_C)$. Utilizando a equação 4.61 e as propriedades descritas, podemos definir a função log-verossimilhança na forma:

$$\begin{aligned} \text{logLik} &= \sum_{n_A} y_i^A \cdot \log(\pi_i^A) + \sum_{n_B} y_i^B \cdot \log(\pi_i^B) + \sum_{n_C} y_i^C \cdot \log(\pi_i^C) \\ &= \sum_i (1 - y_i^B - y_i^C) \log(1 - \pi_i^B - \pi_i^C) + y_i^B \log(\pi_i^B) + y_i^C \log(\pi_i^C) \\ &= \sum_i y_i^B \log\left(\frac{\pi_i^B}{\pi_i^A}\right) + y_i^C \log\left(\frac{\pi_i^C}{\pi_i^A}\right) + \log(1 - \pi_i^B - \pi_i^C) \\ &= \sum_i y_i^B \eta_i^B + y_i^C \eta_i^C + \log(1 - \pi_i^B - \pi_i^C) \end{aligned} \quad (4.62)$$

onde $\log\left(\frac{\pi_i^B}{\pi_i^A}\right) = \eta_i^B = \mathbf{x}_i \beta^B$ e $\log\left(\frac{\pi_i^C}{\pi_i^A}\right) = \eta_i^C = \mathbf{x}_i \beta^C$, como pode ser verificado nas equações 4.56 e 4.57. Também é possível mostrar que:

$$\begin{aligned} \log(1 - \pi_i^B - \pi_i^C) &= \log\left(1 - \frac{e^{\eta_i^B} + e^{\eta_i^C}}{1 + e^{\eta_i^B} + e^{\eta_i^C}}\right) \\ &= \log\left(\frac{1}{1 + e^{\eta_i^B} + e^{\eta_i^C}}\right) \\ &= -\log\left(1 + e^{\eta_i^B} + e^{\eta_i^C}\right) \end{aligned} \quad (4.63)$$

o que permite reescrever a equação de log-verossimilhança (equação 4.62) na forma:

$$\text{logLik} = \sum_i y_i^B \eta_i^B + y_i^C \eta_i^C - \log\left(1 + e^{\eta_i^B} + e^{\eta_i^C}\right) \quad (4.64)$$

A equação 4.64 é muito semelhante à equação de verossimilhança de um modelo logístico. O algoritmo de estimação da família exponencial, supondo ligação canônica, é apresentado na equação 4.5 e replicado a seguir.

$$\beta^{(k+1)} = \beta^{(k)} + [\mathbf{K}^{(k)}]^{-1} \mathbf{U}(\beta^{(k)})$$

No caso da multinomial, vamos assumir o vetor de parâmetros como $\beta^T = [\beta^B, \beta^C]$. Então, a Função Escore $\mathbf{U}(\beta)$ pode ser escrita como:

$$\mathbf{U}^T = \left[\frac{\partial \text{logLik}}{\partial \beta^B}, \frac{\partial \text{logLik}}{\partial \beta^C} \right] \quad (4.65)$$

onde:

$$\begin{aligned}\frac{\partial \log L}{\partial \beta^B} &= \sum_i y_i^B x_i - \frac{e^{\eta_i^B}}{1 + e^{\eta_i^B} + e^{\eta_i^C}} \cdot x_i \\ &= \sum_i y_i^B x_i - \pi_i^B x_i \\ &= \sum_i (y_i^B - \pi_i^B) x_i \\ &= X^T (Y^B - \pi^B)\end{aligned}\quad (4.66)$$

De forma semelhante, o cálculo da derivada da log-verossimilhança com relação ao vetor β^C é definido como:

$$\frac{\partial \log L}{\partial \beta^C} = X^T (Y^C - \pi^C)$$

A matriz K , no exemplo, pode ser definida por:

$$K = \begin{bmatrix} \frac{\partial^2 \log L}{\partial \beta^B \partial \beta^B} & \frac{\partial^2 \log L}{\partial \beta^B \partial \beta^C} \\ \frac{\partial^2 \log L}{\partial \beta^C \partial \beta^B} & \frac{\partial^2 \log L}{\partial \beta^C \partial \beta^C} \end{bmatrix} \quad (4.67)$$

onde:

$$\begin{aligned}\frac{\partial}{\partial \beta^C} \left(\frac{\partial \log L}{\partial \beta^B} \right) &= \frac{\partial}{\partial \beta^C} \left(\sum_i y_i^B - \pi_i^B x_i \right) \\ &= - \sum_i x_i^T \cdot \left(\frac{\partial \pi_i^B}{\partial \beta^C} \right) \\ &= - \sum_i x_i^T \cdot \left(\frac{\partial \pi_i^B}{\partial \eta_i^C} \right) \cdot \left(\frac{\partial \eta_i^C}{\partial \beta^C} \right)\end{aligned}\quad (4.68)$$

A segunda parte da equação 4.68 pode ser solucionada como $\frac{\partial \eta_i^C}{\partial \beta^C} = x_i$, já a primeira parte pode ser solucionada na forma:

$$\begin{aligned}\frac{\partial \pi_i^B}{\partial \eta_i^C} &= \frac{\partial}{\partial \eta_i^C} \left[e^{\eta_i^B} \cdot (1 + e^{\eta_i^B} + e^{\eta_i^C})^{-1} \right] \\ &= e^{\eta_i^B} \cdot (-1) \cdot (1 + e^{\eta_i^B} + e^{\eta_i^C})^{-2} \cdot e^{\eta_i^C} \\ &= - \frac{e^{\eta_i^B}}{1 + e^{\eta_i^B} + e^{\eta_i^C}} \cdot \frac{e^{\eta_i^C}}{1 + e^{\eta_i^B} + e^{\eta_i^C}} \\ &= -\pi_i^B \cdot \pi_i^C\end{aligned}\quad (4.69)$$

A partir do resultados obtido na equação 4.69, é possível concluir que a derivada de segunda ordem $\frac{\partial^2 \log L}{\partial \beta^C \partial \beta^B}$ pode ser escrita na forma:

$$\frac{\partial^2 \log L}{\partial \beta^C \partial \beta^B} = X^T W^{BC} X \quad (4.70)$$

onde W^{BC} é uma matriz diagonal cujos elementos são definidos por $[W^{BC}]_{ii} = -\pi_i^B \cdot \pi_i^C$. Aplicando este mesmo princípio, é possível mostrar que $\frac{\partial^2 \log L}{\partial \beta^B \partial \beta^C} = X^T W^{CB} X$, onde W^{CB} também é uma matriz diagonal na qual $[W^{CB}]_{ii} = -\pi_i^C \cdot \pi_i^B$. Conclui-se que $W^{CB} = W^{BC}$.

No caso dos elementos da diagonal da matriz \mathbf{K} , no exemplo proposto, é possível calcular as derivadas de segunda ordem como:

$$\begin{aligned}
 \frac{\partial^2 \log L}{\partial \beta^B \partial \beta^B} &= \frac{\partial}{\partial \beta^B} \sum_i (y_i^B - \pi_i^B) \mathbf{x}_i \\
 &= -\mathbf{x}_i^T \cdot \frac{\partial \pi_i^B}{\partial \beta^B} \\
 &= -\mathbf{x}_i^T \cdot \frac{\partial}{\partial \beta^B} \left(\frac{e^{\eta_i^B}}{1 + e^{\eta_i^B} + e^{\eta_i^C}} \right) \\
 &= -\mathbf{x}_i^T [\pi_i^B (1 - \pi_i^B)] \mathbf{x}_i \\
 &= -\mathbf{X}^T \mathbf{W}^{BB} \mathbf{X}
 \end{aligned} \tag{4.71}$$

onde \mathbf{W}^{BB} é uma matriz diagonal e $[\mathbf{W}^{BB}]_{ii} = \pi_i^B (1 - \pi_i^B)$. Aplicando o mesmo princípio, é possível mostrar que $\frac{\partial^2 \log L}{\partial \beta^C \partial \beta^C} = -\mathbf{X}^T \mathbf{W}^{CC} \mathbf{X}$, onde $[\mathbf{W}^{CC}]_{ii} = \pi_i^C (1 - \pi_i^C)$. É interessante observar as semelhanças das equações encontradas com a variância de uma bernoulli, $\pi (1 - \pi)$. Esta estrutura também é utilizada na matriz \mathbf{K} , no respectivo algoritmo de estimativa do modelo logístico.

Como principal conclusão, é possível reescrever a matriz \mathbf{K} na forma:

$$\mathbf{K} = \begin{bmatrix} -\mathbf{X}^T \mathbf{W}^{BB} \mathbf{X} & -\mathbf{X}^T \mathbf{W}^{BC} \mathbf{X} \\ -\mathbf{X}^T \mathbf{W}^{CB} \mathbf{X} & -\mathbf{X}^T \mathbf{W}^{CC} \mathbf{X} \end{bmatrix} \tag{4.72}$$

Os resultados obtidos podem ser facilmente estendidos para modelos multinomiais com mais de três classes. É importante destacar que foi apresentada uma versão simplificada do modelo multinomial onde ocorre uma única escolha dentre três possíveis classes (A , B ou C). Em uma versão mais ampla, existe a possibilidade de escolher n itens dentre as classes disponíveis. Na média, o número de itens selecionado para cada classe é o número total de itens n multiplicado pela probabilidade da escolha de cada item, ou seja, $n_A = n \cdot \pi_A$, $n_B = n \cdot \pi_B$ e $n_C = n \cdot \pi_C$ onde n_A , n_B e n_C são os números médios de itens selecionados para as classes A , B e C , respectivamente, e $n_A + n_B + n_C = n$.

Existem outras classes de modelos multinomiais como os modelos multinomiais ordinais. Nesta classe de modelos, existe uma hierarquia com relação às classes da variável resposta. É o caso, por exemplo, de uma variável resposta do tipo $Y \in \{\text{gelado}, \text{frio}, \text{morno}, \text{quente}\}$, onde não há uma associação numérica predefinida para as classes mas deseja-se manter a hierarquia: gelado < frio < morno < quente. Esta classe de modelos multinomiais não será objeto de análise neste texto.

4.11.3 EXERCÍCIOS**Exercício 4.1**

Para o modelo de Bernoulli descrito na forma da família exponencial (ver equação), mostre que: $E[Y] = b'(\theta) = p$ e $Var[Y] = b''(\theta) = p(1 - p)$.

Exercício 4.2

Para o modelo Bernoulli, $Y_i \sim Bernoulli(p_i)$, defina a equação da função *Deviance*.

Exercício 4.3

Para o modelo Binomial, descrito na equação 4.37, verifique se o mesmo pertence à família exponencial e analise as suas propriedades.

Exercício 4.4

Para a distribuição Gama, descrita na equação 4.44, verifique que a mesma pertence à família exponencial e identifique as propriedades da ligação canônica.