Estatística Multivariada
GLM- Modelos Lineares Generalizados
v. 1.0.0

Fábio Rocha da Silva
fabiorochadasilva@decom.cefetmg.br

CEFET-MG
Belo Horizonte “Campus II”

27 de março de 2016

Por que queremos modelar os dados?

- A forma do modelo revela padrões de interação e associação nos dados.
- Através de procedimentos de inferência podemos verificar:
- Quais variáveis explicativas estão relacionadas com a variável resposta;
- Enquanto controlamos outras variáveis relevantes.
- A estimativa dos parâmetros fornece a importância de cada variável no modelo.

No modelo de Regressão Múltipla

Todas as conclusões do modelo estão pautadas em um suposição forte:

o vetor Y tem distribuição normal.

- Muitas vezes essa suposição não será satisfeita.
- Se a resposta é uma variável categórica, isso não será verdade.
- Se Y for quantitativa discreta a suposição também será violada.
- Precisamos, então, usar os Modelos Lineares Generalizados.

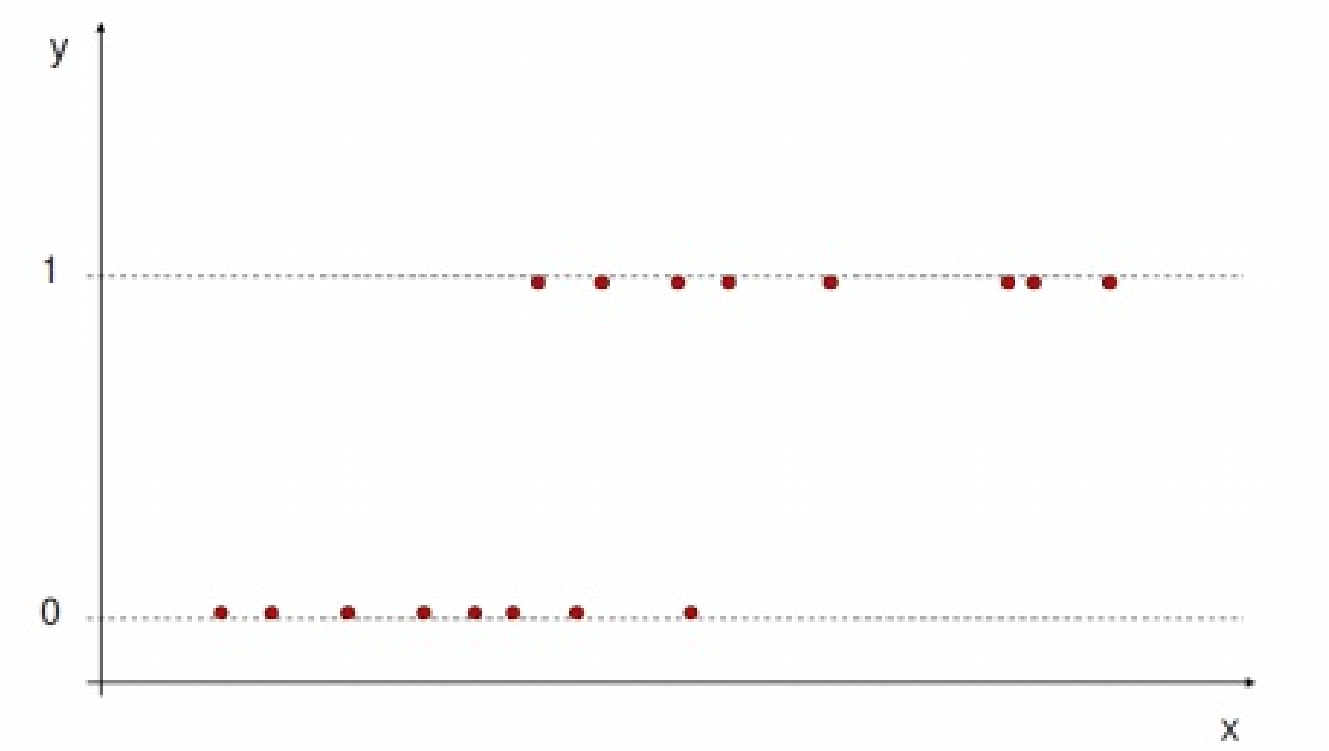
Considere as seguintes situações:

- Um paciente é submetido a um cirurgia.
  - Deseja-se prever a chance do paciente sobreviver.
  - Estima-se essa chance com base em dados clínicos pré-operatórios.
- Estamos analisando casos de dengue em municípios.
  - Queremos tentar prever o número de casos.
  - Podemos usar informações sócio-econômicas do município.

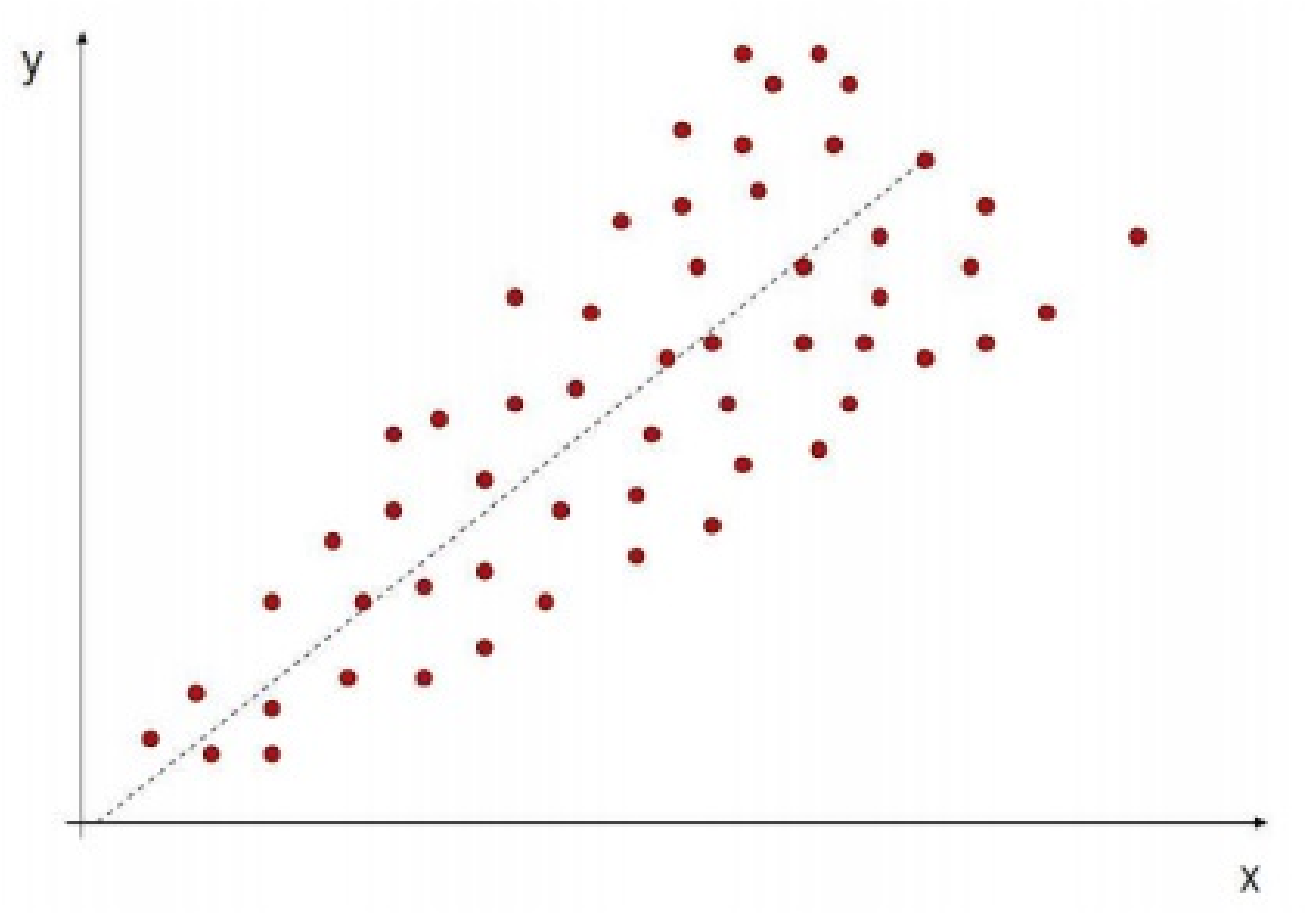
Objetivo dos **GLM** :

- Fornecer ferramentas para a análise de dados que não apresentam uma distribuição Normal:
  - Bernoulli/Binomial;
  - Poisson;
  - Binomial Negativa;
  - Gama.

- Veremos modelos para tratar com dados binários.



- Modelos para lidar com dados heterocedásticos.



- Nelder e Wedderburn (1972), propuseram os Modelos Lineares Generalizados (MLGs), que são uma extensão dos modelos normais

- A ideia básica consiste em abrir o leque de opções para a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições, bem como dar maior flexibilidade para a relação funcional entre a média da variável resposta (μ) e o preditor linear η.

- A ligação entre a média e o preditor linear não é necessariamente a identidade, podendo assumir qualquer forma monótona não-linear.
- Nelder e Wedderburn propuseram também um processo iterativo para a estimação dos parâmetros e introduziram o conceito de desvio que tem sido largamente utilizado na avaliação da qualidade do ajuste dos MLGs, bem como no desenvolvimento de resíduos e medidas de diagnóstico.

- No modelo de Regressão Linear temos que

$$Y_i \sim N(\mu, \sigma^2)$$

  - $E(Y_i) = \mu_i = x_i^T \beta$
  - Y_i são independentes;
  - x_i^T representa a i-ésima linha da matriz X, correspondente ao i-ésimo indivíduo.

Podemos estar interessados em situações mais genéricas.

- A variável resposta tem uma distribuição diferente da normal.
- A relação entre o valor esperado da variável resposta e as explicativas pode ter uma relação diferente de

$$E(Y_i) = \mu_i = x_i^T \beta$$

podemos ter

$$E(Y_i) = \mu_i = g(x_i^T \beta)$$

onde $g(\cdot)$ é uma função genérica.

- A variável Y_i não pode ter **QUALQUER** distribuição.
- Precisamos garantir certas propriedades para:
  - estimar os parâmetros,
  - fazer testes de hipóteses,
  - tirar conclusões sobre o modelo.
- Uma classe de distribuições garante essas propriedades.
- Essa classe é conhecida como **família exponencial**
- O que é a família exponencial?

Família:

Conjunto de distribuições com características similares.

- O que a família exponencial?
  - É uma família de distribuições cuja função densidade pode ser escrita na seguinte forma

$$f(y;\theta,\phi) = exp\left\{a(\phi)^{-1}\left[y\theta - b(\theta) + c(y;\phi)\right]\right\}$$

    onde $b(\cdot)$, $a(\cdot)$, e $c(\cdot)$ são funções não negativas.
  - Um resultado importante

$$\begin{aligned} \mu &= E(Y) = b'(\theta) \\ \sigma^2 &= Var(Y) = a(\phi)b''(\theta) \end{aligned}$$

Exemplo:

- Seja Y uma variável tal que

$$Y \sim Poisson(\theta)$$

.

- Vamos verificar que essa distribuição pertence à família exponencial

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

Exemplo:

- A distribuição Normal pertence à Família Exponencial? Sim.
- Vejamos porque isso é verdade.
- Seja

$$Y \sim N(\mu, \sigma^2)$$

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

- Observe que no exemplo anterior consideramos $\theta = \mu$.
- Tratamos σ^2 como uma constante conhecida
- Ele é chamado parâmetro de ruído (*nuisance parameter*).
- Na prática, precisamos estimá-lo.
- Porém o nosso interesse está em μ e não em σ^2.

Parâmetro Canônico

- Veremos mais a frente que um tipo específico de parâmetro será de grande importância.
- Ele é chamado parâmetro canônico.
- Considere a função densidade escrita na forma

$$f(y;\theta,\phi) = exp\left\{a(\phi)^{-1}\left[y\theta - b(\theta) + c(y;\phi)\right]\right\}.$$

- (θ) é chamado *parâmetro canônico* da distribuição.

O modelo normal linear é definido por

1. Y independentes com $Y_i \sim N(\mu_i, \sigma^2)$
2. $\mu_i = \eta_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip}$

O MLGs são definidos por

1. Y_i são independentes com $Y_i \sim FE(\mu_i, \phi), \quad i = 1, \ldots n,$
2. $\mu_i = g^{-1}(\eta_i),$

em que $E(Y_i) = \mu_i$, $Var(Y_i) = \phi^{-1} V(\mu_i)$ função de variância, $g(\cdot)$ função de ligação, ϕ^{-1} parâmetro de dispersão e η_i é o preditor linear.

O MLGs são definidos por

O modelo linear generalizado pode ser dividido em três partes:

1. Componente Aleatório;
2. Parte Sistemática;
3. Função de Ligação.

1-Componente aleatório:

- É representado por um conjunto de variáveis aleatórias independentes $Y_1, \ldots, Y_n$.
- Todas provenientes de uma mesma distribuição que faz parte da família exponencial de distribuições;
- cada uma das variáveis com médias $\mu_1, \ldots, \mu_n$

$$E(Y_i) = \mu_i, \quad i = 1, \ldots, n$$

2-Componente sistemático:

- As variáveis explicativas entram na forma de uma soma linear de seus efeitos

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_r = \mathbf{x}_i^T\beta \text{ ou } \eta = \mathbf{X}\beta$$

- sendo
  - $\mathbf{X} = (x_1, x_2, \ldots, x_n)^T$ a matriz do modelo,
  - $\beta = (\beta_1, \ldots \beta_p)^T$ o vetor de parâmetros e
  - $\eta = (\eta_1, \ldots, \eta_n)^T$ o preditor linear.
- *Se um parâmetro tem valor conhecido, o termo correspondente na estrutura linear é chamado* **offset**

3-Função de ligação:

- É uma função que relaciona o componente aleatório ao componente sistemático, ou seja, relaciona a média ao preditor linear, isto é,

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

- sendo $g(\cdot)$ uma função monótona e diferenciável.
- As funções de ligação usuais são:
  - potência $\eta = \mu^\lambda$, λ real.
  - logística $\eta = \log\left[\mu/(m-\mu)\right]$
  - probito $\eta = \Phi^{-1}\left(\mu/m\right)$ sendo $\Phi(\cdot)$ a função de distribuição acumulada (f.d.a.) da distribuição normal padrão
  - complemento loglog $\eta = \log\left[-\log\left(1-\mu/m\right)\right]$, em que m é o número de ensaios independentes.

- Se a função de ligação é escolhida de tal forma que $g(\mu_i) = \theta_i = \eta_i$ o preditor linear modela diretamente o parâmetro canônico θ_i
- Esta ligação é chamada de função de ligação canônica.
- Os modelos correspondentes são denominados canônicos.
- Isso resulta, frequentemente, em uma escala adequada para a modelagem com interpretação prática para os parâmetros de regressão

Tabela: Funções de ligação canônicas

Distribuição	Função de ligação canônica
Normal	Identidade: $\eta = \mu$
Poisson	Logarítmica: $\eta = \log(\mu)$
Binomial	Logística: $\eta = \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{\mu}{m-\mu}\right)$
Gama	Recíproca: $\eta = \frac{1}{\mu}$
Normal Inversa	Recíproca do quadrado: $\eta = \frac{1}{\mu^2}$

Alguns modelos na família exponencial

Modelo logístico-linear

1. $Y_i \sim \text{Bernoulli}(\mu_i), \quad i = 1, \ldots, n, \quad 0 < \mu_i < 1$
2. $\mu_i = \frac{\eta_i}{1+\exp\{\eta_i\}}$

Modelo recíproco gama

1. $Y_i \sim \text{Gama}(\mu_i, \phi) \quad i = 1, \ldots, n$
2. $\mu_i = \eta_i^{-1}$

Modelo log-linear de Poisson

1. $Y_i \sim \text{Poisson}(\mu_i)$ $i = 1, \ldots, n$
2. $\mu_i = \exp\{\eta_i\}$

em que $Y_i = 0, 1, 2, \ldots$, $\mu_i > 0$ e $Var(Y_i) = \mu_i$

Modelo log-linear Binomial Negativa

1. $Y_i \sim \text{Binomial negativa}(\mu_i, \phi)$ $i = 1, \ldots, n$
2. $\mu_i = \exp\{\eta_i\}$

em que $Y_i = 0, 1, 2, \ldots$, $\mu_i > 0$ e $Var(Y_i) = \mu_i + \frac{\mu_i^2}{\phi}$

- Consideremos n variáveis aleatórias independentes $y_1, \ldots, y_n$, cada uma com função densidade (ou de probabilidade) na família exponencial da forma

$$f(y; \theta_i, \phi) = \exp\left\{\phi\left[y\theta_i - b(\theta_i)\right] + c(y, \phi)\right\}, i = 1, \ldots, n \tag{1}$$

onde $b(\cdot)$ e $c(\cdot)$ são funções conhecidas. Para o modelo em 1 valem as seguintes relações:

$$E(y_i) = \mu_i = b'(\theta_i), \quad Var(y_i) = \phi^{-1} V_i, \quad i = 1, \ldots, n$$

sendo ϕ^{-1} o parâmetro de dispersão e $V = \frac{d\mu}{d\theta}$ a função de variância (caracteriza a distribuição).

- Os MLGs são definidos por 1 e pela componente sistemática

$$(g\mu_i) = \eta_i, \quad i = 1, \ldots, n \tag{2}$$

onde $g(\cdot)$ é uma função monótona e diferenciável, denominada função de ligação,

  - $\eta = \mathbf{X}\beta$,
  - $\beta = (\beta_1, \beta_2, \ldots \beta_p)^T \quad p \leq n$
  - $\mathbf{X} = (x_1, x_2, \ldots, x_p)$
  - $x_j = \left(x_{1j}, x_{2j}, \ldots x_{nj}\right)^T$

Consideremos um MLG definido por 1 e 2,O logaritmo da função de verossimilhança como função de β pode ser expresso na forma

$$L(y;\beta) = \sum_{i=1}^{n} \phi \left[y_i \theta_i - b(\theta_i) + c(y_i) \right] + a(y_i, \phi).$$

- $\theta_i = \eta_i = \sum_{j=1}^{n} x_{ij}\beta_j, \quad i = 1, \ldots n.$

$$L(y\beta) = \sum_{i=1}^{n} \phi \left[y_i \sum_{j=1}^{p} x_{ij}\beta_j - b\left(\sum_{j=1}^{p} x_{ij}\beta_j \right) \right] + \phi \sum_{i=1}^{n} c(y_i) + \sum_{i=1}^{n} a(y_i, \phi)$$

- Para os modelos normal, Poisson, binomial, gama e normal inversa as ligações canônicas são dadas por

$$\eta = \mu, \;\; \eta = \log(\mu), \;\; \eta = \log\left(\frac{\mu}{1-\mu}\right), \;\; \eta = \mu^{-1} \;\; \text{e} \;\; \eta = \mu^{-2}.$$

  garantem a concavidade de $L(y;\beta)$, isto é, garantem a unicidade da estimativa de máxima verossimilhança de β, quando essa existe e, consequentemente, muitos resultados assintóticos são obtidos mais facilmente.
- Para ligações não-canônicas Wedderburn (1976) discute condições à existência da concavidade $L(y;\beta)$.

- Os MLGs são ajustados no R através da função glm, onde devemos especificar formula (a definição do modelo) e family (a distribuição assumida pela variável resposta com a função de ligação a ser usada). Por exemplo,

$$ajuste = glm(y\ 1 + x,\ family = gaussian).$$

- Se a função de ligação usada for diferente do *'default'*, basta especificar a função de ligação desejada através do comando *link*. Por exemplo,

$$ajuste=glm(y\ 1+x,family=gaussian(link='log')).$$

- O comando *summary(ajuste)* dá um resumo do resultado do ajuste.

- Sem perda de generalidade, suponha que o logaritmo da função de verossimilhança seja agora definido por

$$L(y;\mu) = \sum_{i=1}^{n} L(y_i;\mu_i)$$

em que $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = x_i^T\beta$. Para o modelo saturado ($p = n$) a função $L(y;\mu)$ é estimada por

$$L(y;y) = \sum_{i=1}^{n} L(y_i;y_i)$$

Ou seja, a estimativa de máxima verossimilhança de μ_i fica nesse caso dada por $\mu_i^0 = y_i$

- Quando $p < n$, denotamos a estimativa de $L(y; \mu)$ por $L(y; \hat{\mu})$. Aqui, a estimativa de máxima verossimilhança de μ_i será dada por $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ em que $\hat{\eta}_i = x_i^T \hat{\beta}$.
- A qualidade do ajuste de um MLG é avaliada através da função desvio dada por

$$D^*(y, \hat{\mu}) = \phi D(y, \hat{\mu}) = 2\{L(y; y) - L(y, \hat{\mu})\}$$

onde

$$D(y, \hat{\mu}) = 2 \sum_{i=1}^{n} \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) + (b(\tilde{\theta}_i) - b(\hat{\theta}_i)) \right],$$

denotamos por $\hat{\theta}_i = \theta_i(\hat{\mu}_i)$ e $\tilde{\theta}_i = \theta_i(\tilde{\mu}_i)$, respectivamente, as estimativas de máxima verossimilhança de θ_i para os modelos com p parâmetros ($p < n$) e saturado ($p = n$).

Apresentaremos a seguir a função desvio dos casos especiais citados anteriormente.

- Normal

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

- Binomial

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{n_i\hat{\mu}}\right) + (n_i - y_i)\log\left(\frac{1-(y_i/n_i)}{1-\hat{\mu}}\right)\right]$$

- Gama:

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left[-\log\left(\frac{y_i}{\hat{\mu}_i} + \frac{(y_i-\hat{\mu}_i)}{\hat{mu}_i}\right)\right].$$

- Poisson

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i-\hat{\mu}_i)\right]$$

- Normal Inversa

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

- Um valor pequeno para a função desvio indica que, para um menor número de parâmetros, obtém-se um ajuste tão bom quanto o ajuste com o modelo saturado.
- Embora seja usual comparar os valores observados da função desvio com os percentis da distribuição qui-quadrado com $n - p$ graus de liberdade, em geral, $D(y, \hat{mu})$ não segue assintoticamente uma distribuiçãoqui-quadrado com $n - p$ graud de liberdade.

Suponha para o vetor de parâmetros β a partição $\beta = (\beta_1^T, \beta_2^T)^T$, em que β_1 é um vetor q-dimensional enquanto β_2 tem dimensão p- q.

- Portanto podemos estar interessados em testar as hipóteses $H_0 : \beta_1 = \beta_1^{(0)}$ contra $H_1 : \beta_1 \neq \beta_1^{(0)}$.
- As funções desvio correspondentes aos modelos sob H_0 e H_1 são dadas por $D\left(y; \mu^{(0)}\right)$ e $D(y; \hat{\mu})$, respectivamente, onde $\mu^{(0)}$ é a estimativa de máxima verossimilhança de μ sob H_0.

Análise do desvio

A análise de desvio *(ANODEV)* é uma generalização da análise de variância para os MLGs.

- Podemos definir a seguinte estatística

$$F = \frac{\left\{D(y;\mu^{(0)}) - D(y;\hat{\mu})\right\}/q}{D(y;\hat{\mu})/(n-p)},$$

  cuja distribuição nula assintótica é $F_{q,(n-p)}$.
- Não depende de ϕ e é invariante sob reparametrização
- Pode ser obtida diretamente de funções desvio, é muito conveniente para uso prático.

Através do comando anova, o R fornece uma tabela *ANODEV* para os ajustes colocados como objetos (ajustes de um MLG). Por exemplo, suponha que os objetos *ajuste*, *ajuste1* correspondam aos ajustes de um MLG com um, dois fatores, respectivamente. Então, o comando

anova(ajuste, ajuste1,test = "Chi")

fornece uma tabela comparando os três fatores.

Contribuições dos MLGs

Algumas contribuições importantes dos MLGs.

1. Ligação entre a média e o preditor linear: $\mu_i = g^{-1}(\eta_i)$
   Para cada distribuição da família exponencial novos modelos podem ser gerados variando-se a função de ligação $g(\cdot)$.
2. Função desvio: $D(y; \hat{\mu}) = 2\{L(y, y) - L(\hat{\mu}, y)\}$.
   É uma distância entre as log-verossimilhanças do modelo saturado e do modelo postulado. Para alguns modelos a distribuição do desvio é uma qui-quadrado facilitando avaliar a qualidade do ajuste.

3 Resíduo componente do desvio: $t_{D_i} = \pm\sqrt{d^2(y_i, \hat{\mu}_i)}$
Esse resíduo é muito utilizado para detectar pontos aberrantes e para avaliar a adequação da distribuição utilizada para a resposta.
4. Função de variância: $V(\mu)$
Caracteriza a distribuição da família exponencial. Ou seja, para cada $V(\mu)$ existe apenas uma distribuição na família exponencial e vice-versa. Além disso, quando $\phi \to \infty$ tem-se que $\sqrt{\phi}(Y - \mu) \xrightarrow{d} N(0, V(\mu))$

5-Processo iterativo na forma de mínimos quadrados

A estimativa de máxima verossimilhança $\hat{beta}$ pode ser obtida através do processo iterativo de mínimos quadrados ponderados

$$\beta^{(m+1)} = (X^T W^{(m)} X)^{-1} X^T W^{(m)} z^{(m)},$$

com matriz modelo X, matriz de pesos W e variável dependende modificada z. Esse processo iterativo é inicializado nos próprios valores observados e em geral converge em um número finito de passos.

EXEMPLOS LIVRO GILBERTO A. PAULA

Exemplos Regressão Dados Binários

Dados Binários Agrupados 1

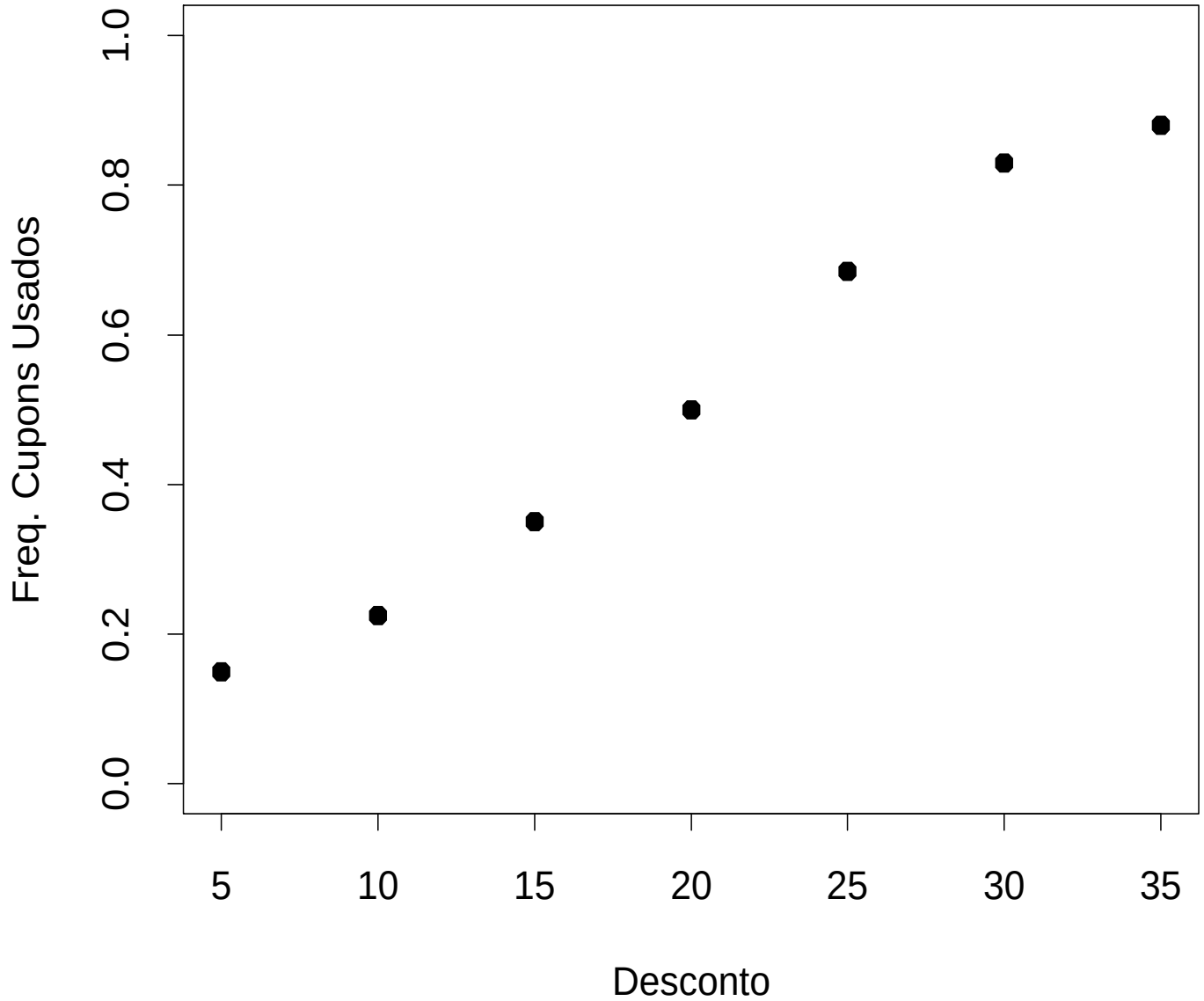
Vamos considerar inicialmente os dados sobre o uso de cupons com descontos, enviados para clientes de uma rede de supermercados (Neter et al., 1996). Cupons com descontos de 5, 10, 15, 20, 25, 30 e 35 reais são enviados a clientes da rede de supermercados escolhidos aleatoriamente e deseja-se estimar a probabilidade de um cupom ser utilizado num prazo de 2 semanas após o envio pelo correio. Inicialmente vamos observar o gráfico da proporção de cupons usados.

Tabela de Cupons Usados

Desconto	Cupons Enviados	Cupons Usados
5	200	30
10	200	45
15	200	70
20	200	100
25	200	137
30	200	166
35	200	176

Na Figura 1 tem-se o comportamento da proporção de cupons usados no período de duas semanas.

Figura 1. Proporção de Cupons Usados.



Nota-se pela Figura 1 que a probabilidade do cupom ser usado aumenta com o desconto do cupom. O modelo para explicar a probabilidade $\mu(x)$ de um cupom com desconto x ser usado pode ser expresso na forma:

- $Y(x) \sim \mathrm{B}(n(x), \mu(x))$
- $\mu(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$,

Nota-se pela Figura 1 que a probabilidade do cupom ser usado aumenta com o desconto do cupom. O modelo para explicar a probabilidade $\mu(x)$ de um cupom com desconto x ser usado pode ser expresso na forma:

- $Y(x) \sim \mathrm{B}(n(x), \mu(x))$
- $\mu(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$,

em que $Y(x)$ denota o número de cupons usados e $n(x)$ o número de cupons enviados com desconto x.

Estimativas dos Parâmetros

Efeito	Estimativa	E/E. Padrão
Constante	-2,535	-16,11
Desconto	0,132	17,91

Portanto, nota-se que há um crescimento significativo da probabilidade do cupom ser usado em duas semanas com o aumento do desconto.

O desvio do modelo é dado por $\mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2,16$ (para 5 graus de liberdade), obtendo-se o nível descritivo P=0,83 que indica que o modelo está bem ajustado.

Modelo ajustado

$$\hat{\mu}(x) = \frac{e^{-2{,}535+0{,}132x}}{1+e^{-2{,}535+0{,}132x}},$$

em que $\hat{\mu}(x)$ é a probabilidade ajustada do cupom com desconto x ser usado no período de duas semanas.

Modelo ajustado

$$\hat{\mu}(x) = \frac{e^{-2{,}535+0{,}132x}}{1+e^{-2{,}535+0{,}132x}},$$

em que $\hat{\mu}(x)$ é a probabilidade ajustada do cupom com desconto x ser usado no período de duas semanas.

A chance do cupom com desconto x ser usado no período de duas semanas fica dado por

$$\frac{\mu(x)}{1-\mu(x)} = \exp\{\alpha + \beta x\}.$$

Chance ajustada

$$\frac{\hat{\mu}(x)}{1-\hat{\mu}(x)} = \exp\{-2,535 + 0,132x\},$$

ou seja, a chance aumenta com o valor do desconto.

Chance ajustada

$$\frac{\hat{\mu}(x)}{1-\hat{\mu}(x)} = \exp\{-2,535 + 0,132x\},$$

ou seja, a chance aumenta com o valor do desconto. A razão de chances entre um cupom com desconto $(x+1)$ e um cupom com desconto x é definida por:

$$\psi(x) = \frac{\frac{\mu(x+1)}{1-\mu(x+1)}}{\frac{\mu(x)}{1-\mu(x)}}.$$

Razão de chances ajustada:

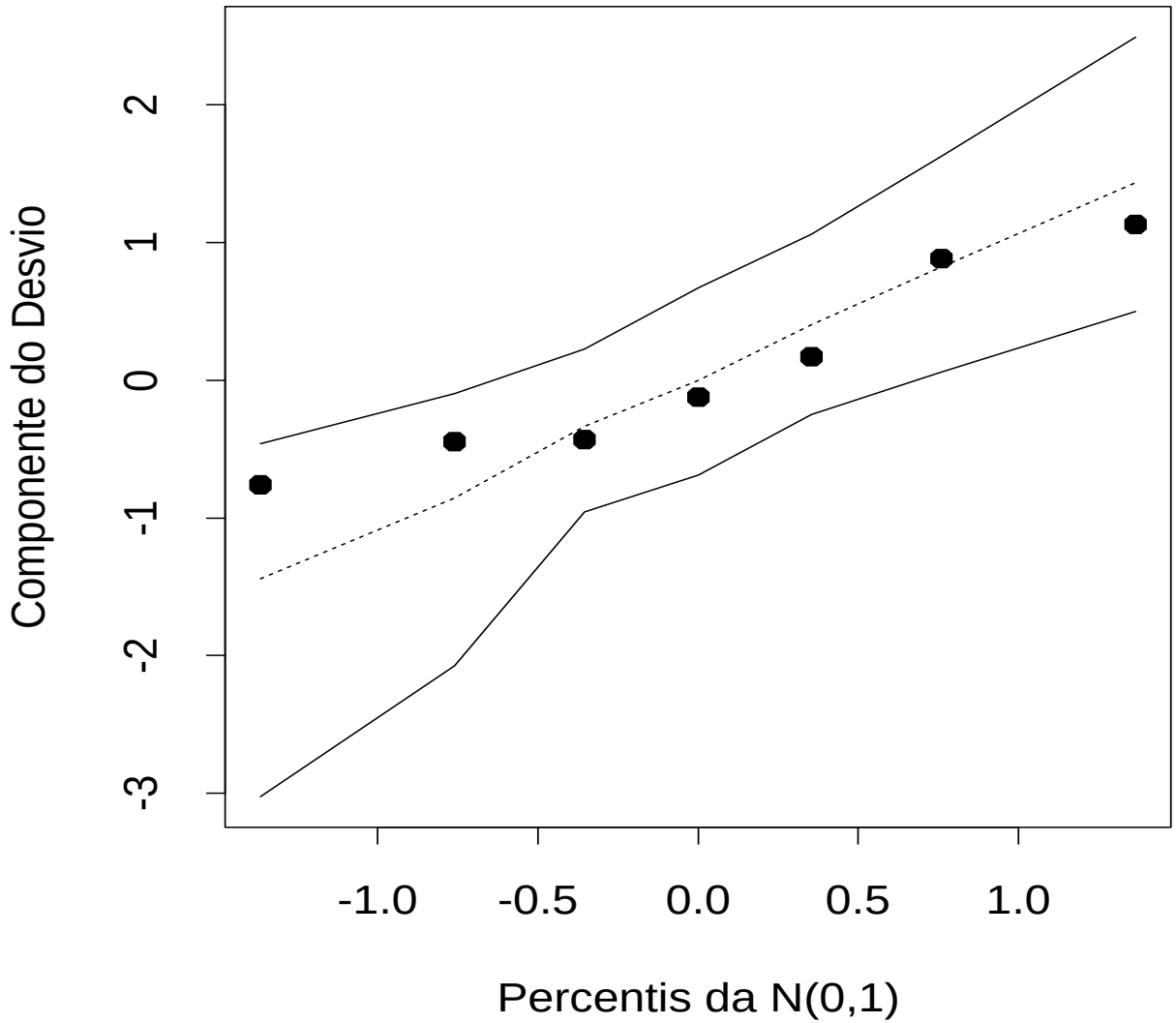
$$\begin{aligned}
\hat{\psi}(x) &= \frac{\exp\{-2,535 + 0,132(x+1)\}}{\exp\{-2,535 + 0,132x\}} \\
&= \exp(0,132) \\
&= 1,14.
\end{aligned}$$

Razão de chances ajustada:

$$\begin{aligned}\hat{\psi}(x) &= \frac{\exp\{-2,535 + 0,132(x+1)\}}{\exp\{-2,535 + 0,132x\}} \\ &= \exp(0,132) \\ &= 1,14.\end{aligned}$$

Interpretação: aumentando em 1 unidade o desconto a chance do cupom ser usado aumenta em aproximadamente 14%.

Figura 2. Envelope Exemplo Cupons.



Dados Binários Agrupados 2

Vamos considerar como outra ilustração o conjunto de dados apresentado em Innes et al. (1969) referente a um estudo para avaliar o possível efeito cancerígeno do fungicida Avadex. No estudo 403 camundongos são observados. Desses, 65 receberam o fungicida e foram acompanhados durante 85 semanas, verificando-se o desenvolvimento ou não de tumor. Os demais animais não receberam o fungicida e também foram acompanhados pelo mesmo período. Os dados são resumidos a seguir.

Distribuição dos Camundongos

	Macho		Fêmea	
Tumor	Tratado	Controle	Tratado	Controle
Sim	6	8	5	13
Não	26	158	28	159
Total	32	166	33	172

Os dados estão descritos no arquivo **camundongos.dat** na seguinte ordem: sexo (1: macho, 0: fêmea), tratamento (1: tratado, 0: controle), ratos que desenvolveram o tumor e total de ratos expostos.

Seja $\pi(x_1, x_2)$ a probabilidade de desenvolvimento de tumor dados x_1 (x_1=1 macho, x_1=0 fêmea) e x_2 (x_2=1 tratado, x_2=0 controle) e vamos denotar por $Y(x_1, x_2)$ o número de camundongos na condição (x_1, x_2) com desenvolvimento de tumor no período. Vamos assumir que $Y(x_1, x_2)$ segue uma binomial com parte sistemática dada por

$$\log\left\{\frac{\pi(x_1, x_2)}{1 - \pi(x_1, x_2)}\right\} = \alpha + \gamma x_1 + \beta x_2 + \delta x_1 x_2,$$

em que δ denota a interação entre os dois fatores.

Para testar a hipótese de ausência de interação entre os fatores sexo e grupo ($H_0 : \delta = 0$) comparamos o desvio do modelo sem interação $D(\mathbf{y}; \hat{\boldsymbol{\mu}}^0) = 0,832$ com os percentis da distribuição qui-quadrado com 1 grau de liberdade. O nível descritivo obtido é dado por P$= 0,362$, indicando pela não rejeição da hipótese nula (homogeneidade das razões de chances). Ou seja, a razão de chances de desenvolvimento de tumor (entre tratado e controle) é a mesma nos grupos macho e fêmea.

Ajustamos então o modelo logístico sem interação

$$\log\left\{\frac{\pi(x_1,x_2)}{1-\pi(x_1,x_2)}\right\} = \alpha + \gamma x_1 + \beta x_2,$$

em que γ e β denotam, respectivamente, os efeitos de sexo e grupo. As estimativas são dadas abaixo:

Efeito	Estimativa	E/E.Padrão
Constante	-2,602	-9,32
Sexo	-0,241	-0,64
Grupo	1,125	2,81

Portanto, tem efeito de grupo mas não tem efeito de sexo.

Note que $\hat{\psi} = e^{\hat{\beta}}$ é a razão de chances estimada entre tratado e controle (que é a mesma para macho e fêmea).

Um intervalo assintótico de confiança para ψ com coeficiente $(1 - \alpha)$, terá os limites

$$(\hat{\psi}_I, \hat{\psi}_S) = \exp\{\hat{\beta} \pm z_{(1-\alpha/2)} \sqrt{\mathrm{Var}(\hat{\beta})}\}.$$

Logo, para o exemplo acima e assumindo um intervalo de 95%, esses limites ficam dados por $[1,403; 6,759]$.

Dados Binários Não Agrupados

Como exemplo neste tópico vamos considerar os dados sobre a preferência de automóveis (1: americano, 0: japonês) de uma amostra aleatória de 263 consumidores (Foster, Stine e Waterman, 1998, pp. 338-339). A probabilidade de preferência por carro americano será relacionada com as seguintes variáveis explicativas do comprador(a): (i) idade (em anos), (ii) sexo (0: masculino; 1: feminino) e (iii) estado civil (0:casado, 1:solteiro). Os dados estão descritos no arquivo **prefauto.dat** na ordem acima.

Dados Binários Não Agrupados

Como exemplo neste tópico vamos considerar os dados sobre a preferência de automóveis (1: americano, 0: japonês) de uma amostra aleatória de 263 consumidores (Foster, Stine e Waterman, 1998, pp. 338-339). A probabilidade de preferência por carro americano será relacionada com as seguintes variáveis explicativas do comprador(a): (i) idade (em anos), (ii) sexo (0: masculino; 1: feminino) e (iii) estado civil (0:casado, 1:solteiro). Os dados estão descritos no arquivo **prefauto.dat** na ordem acima. A seguir tem-se algumas análises descritivas.

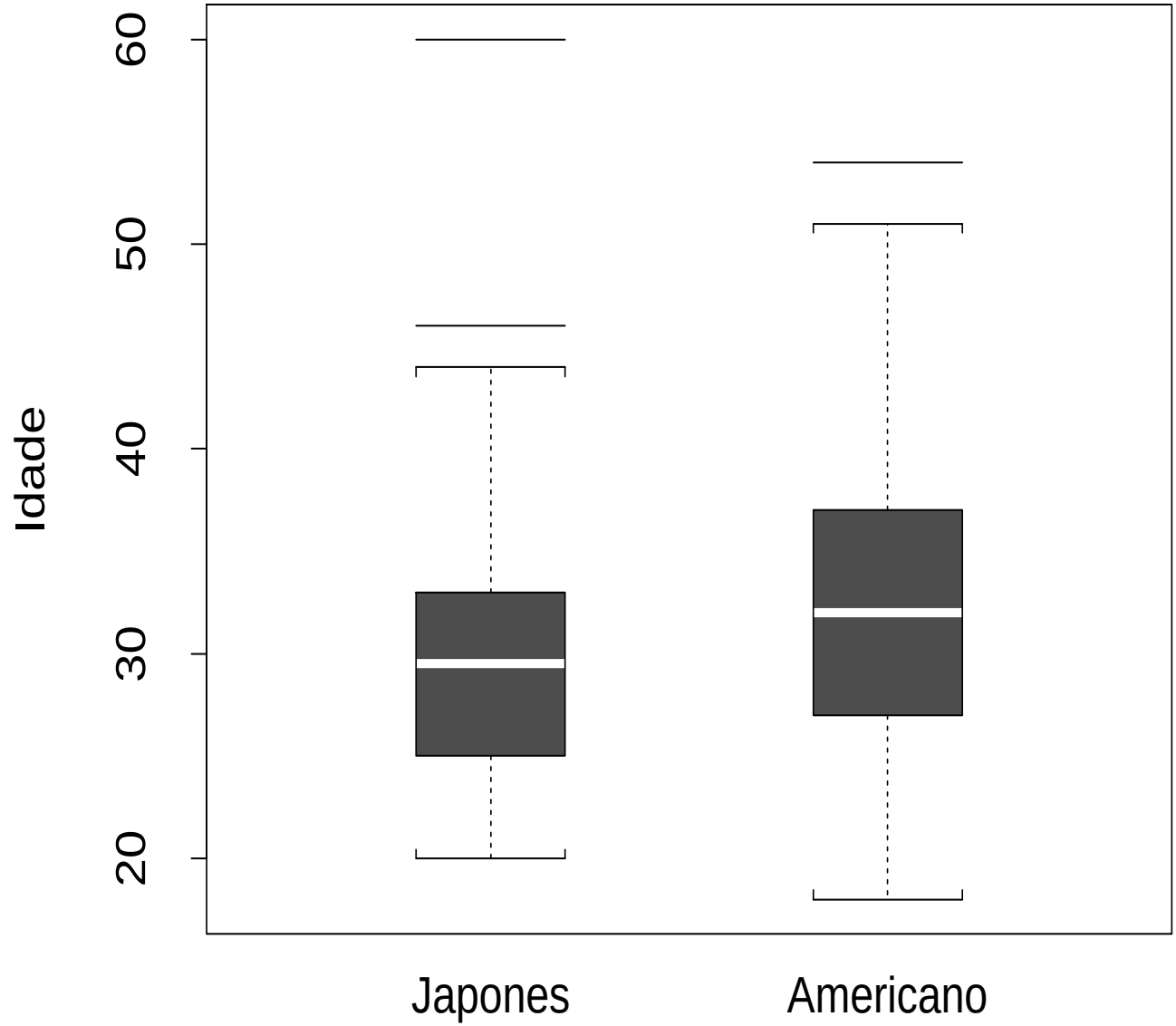
Preferência por Sexo e E. Civil

	Masculino	Feminino
Americano	61 (42,4%)	54 (45,4 %)
Japonês	83 (57,6%)	65 (54,6 %)
Total	144	119

	Casado	Solteiro
Americano	83 (48,8%)	32 (34,4 %)
Japonês	87 (51,2%)	65 (65,6 %)
Total	170	93

Ambos os sexos preferem mais carro joponês. Dentre os casados há pequena vantagem por carro japonês. Essa preferência é bem mais acentuada entre os solteiros.

Figura 3. Idade segundo preferência.



Vamos supor que cada resposta seja Bernoulli com

$$\log\left\{\frac{\mu_i}{1-\mu_i}\right\} = \beta_1 + \beta_2 \times \text{Idade}_i + \beta_3 \times \text{Sexo}_i + \beta_4 \times \text{Ecivil}_i,$$

em que μ_i denota a probabilidade do i-ésimo comprador preferir carro americano. As estimativas são dadas abaixo:

Efeito	Estimativa	E/E.Padrão
Constante	-1,653	-2,33
Idade	0,050	2,31
Sexo	-0,094	-0,37
E.Civil	-0,518	-1,90

Nota-se que a variável sexo é não significativa. As novas estimativas sem essa variável são dadas por:

Efeito	Estimativa	E/E.Padrão
Constante	-1,600	-2,31
Idade	0,049	2,30
E.Civil	-0,526	-1,94

Para testar a inclusão da interação Idade*E.Civil aplicamos o teste da razão de verossimilhanças cujo resultado foi RV=0,81 (1 g.l.). O valor-P foi de P=0,368, portanto não incluímos a interação no modelo.

O modelo ajustado é dado por:

$$\log\left\{\frac{\hat{\mu}}{1-\hat{\mu}}\right\} = -1,600 + 0,049 \times \text{Idade} - 0,526 \times \text{E.Civil}.$$

O modelo ajustado é dado por:

$$\log\left\{\frac{\hat{\mu}}{1-\hat{\mu}}\right\} = -1,600 + 0,049 \times \text{Idade} - 0,526 \times \text{E.Civil.}$$

Portanto, a preferência por automóvel americano aumenta com a idade do comprador. Com relação ao estado civil nota-se que os casados preferem mais carro americano do que os solteiros. Essa razão de chances (entre casados e solteiros) por carro americano pode ser estimada por $\hat{\psi} = \exp(0,526) = 1,69$.

Figura 4. Diagnóstico Exemplo Preferência.

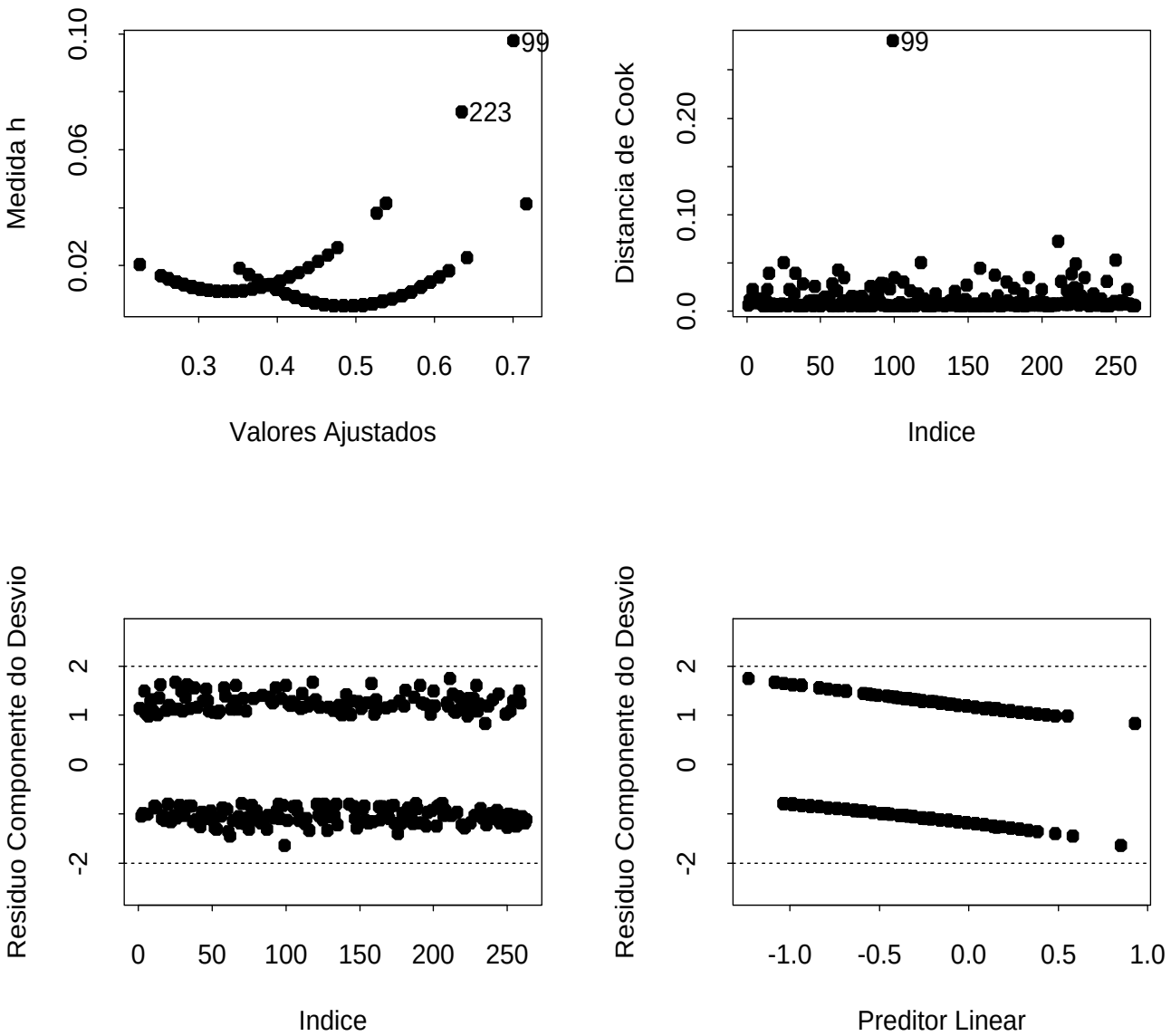
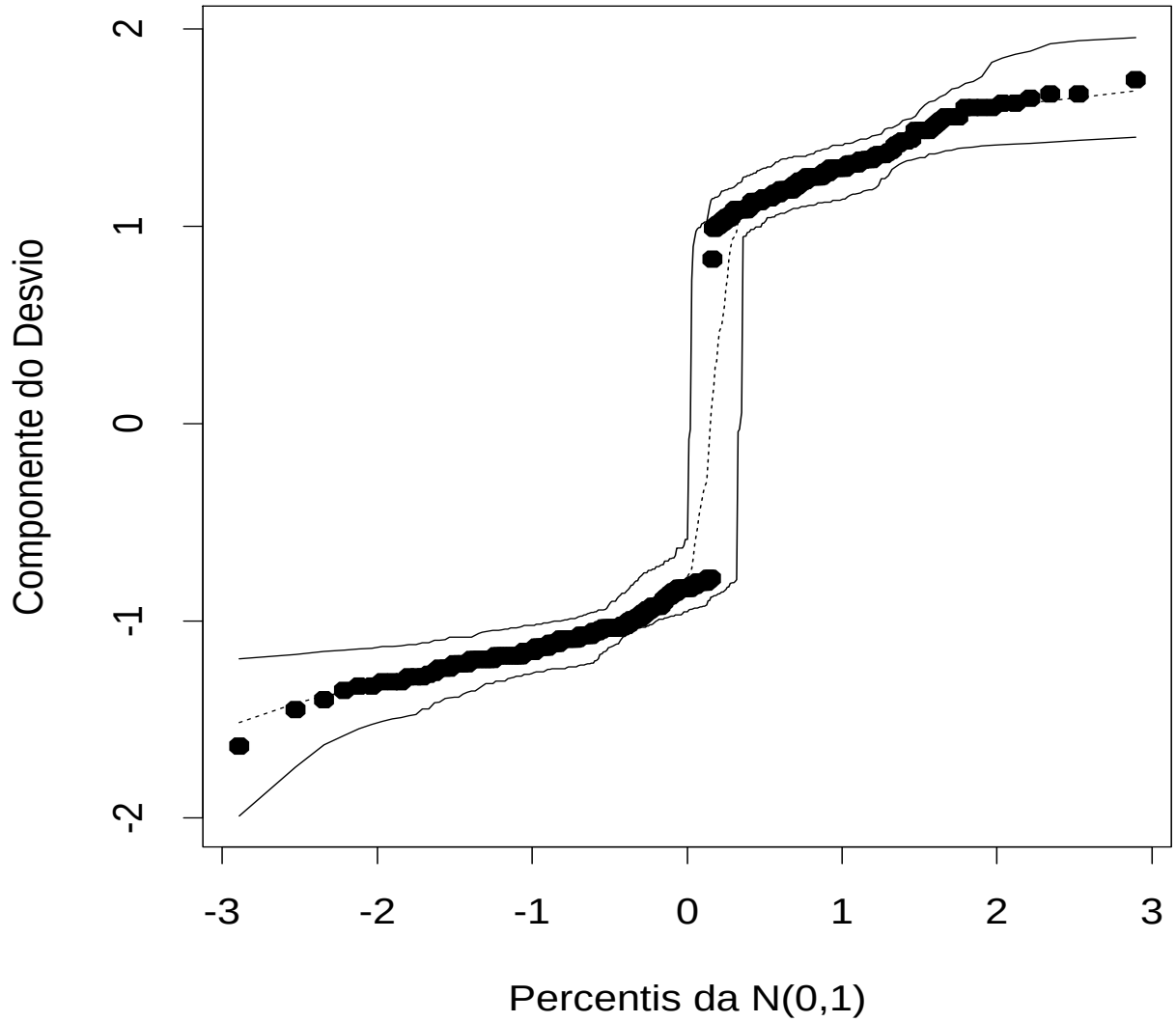


Figura 5. Envelope Exemplo Preferência.



Eliminação Influentes

Apresentamos abaixo as estimativas e variações eliminando-se as observações #99 e #223.

Efeito	Estimativa	z-valor	Variação
Constante	-1,942	-2,65	-17,5%
Idade	0,060	2,65	18,3%
E.Civil	-0,474	-1,72	9,9%

Efeito	Estimativa	z-valor	Variação
Constante	-1,463	-2,07	8,7%
Idade	0,045	2,05	-8,9%
E.Civil	-0,550	-2,02	-4,8%

Conclusões

Neste exemplo em que ajustamos a probabilidade de um comprador preferir carro de marca americana em relação a marca japonesa, notamos que a idade do comprador e o estado civil são variáveis importantes. Com essas duas variáveis o modelo logístico se ajusta bem aos dados. Os dois pontos influentes, referentes a dois compradores com perfil atípico, embora mudem de forma desproporcinal as estimativas não mudam a inferência. Não há indícios de que a distribuição das respostas não seja Bernoulli.

Referências

- Foster, D. P.; Stine, R. A. e Waterman, R. P. (1998). *Business Analysis using Regression*. New York: Springer.
- Innes, J. R. M.; Ulland, B. M.; Valerio, M. G.; Petrucelli, L.; Fishbein, L.; Hart, E. R.; Pallota, A. J.; Bates, R. R.; Falk, H. L.; Gart, J. J.; Klein, M.; Mitchell, I. e Peters, J. (1969). Biossay of pesticides and industrial chemicals for tumorigenicity in mice: A preliminary note. *Journal of the National Cancer Institute* **42**, 1101-1114.

- Neter, J.; Kutner, M. H.; Nachtsheim, C. J. e Wasserman, W.(1996). *Applied Linear Regression Models*, 3rd Edition. Irwin, Illinois,

Regressão Dados de Contagem

Exposição de bactérias

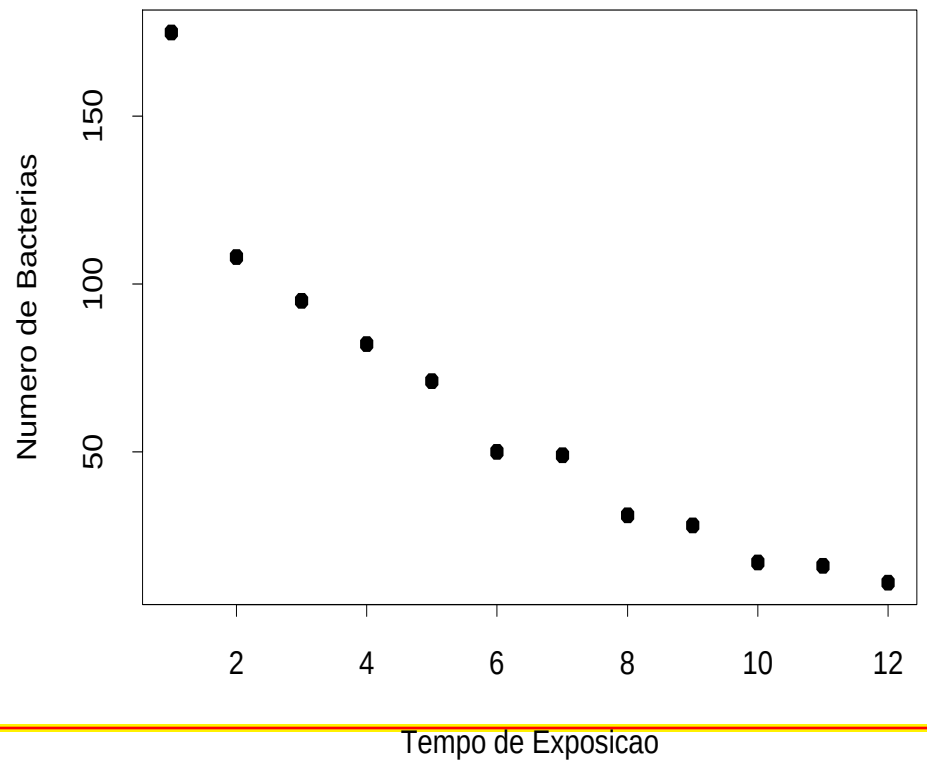
Vamos considerar um exemplo em que modelos de regressão normal linear são comparados com um modelo log-linear de Poisson para ajustar dados de contagem.

- resposta: número de bactérias sobreviventes em amostras de um produto elimentício exposto a uma temperatura de 300^{o}F.
- variável explicativa: tempo de exposição do produto (em minutos).

(Montgomery, Peck e Vining, 2001) (Paula, 2010).

Descrição dados bactérias

Bactérias	175	108	95	82	71	50
Exposição	1	2	3	4	5	6
Bactérias	49	31	28	17	16	11
Exposição	7	8	9	10	11	12



Ajuste modelos nomais

Com base na aproximação da Poisson pela normal vamos propor inicialmente os seguintes modelos:

$$\sqrt{y_i} = \alpha + \beta \text{tempo}_i + \epsilon_i$$

e

$$\sqrt{y_i} = \alpha + \beta \text{tempo}_i + \gamma \text{tempo}_i^2 + \epsilon_i,$$

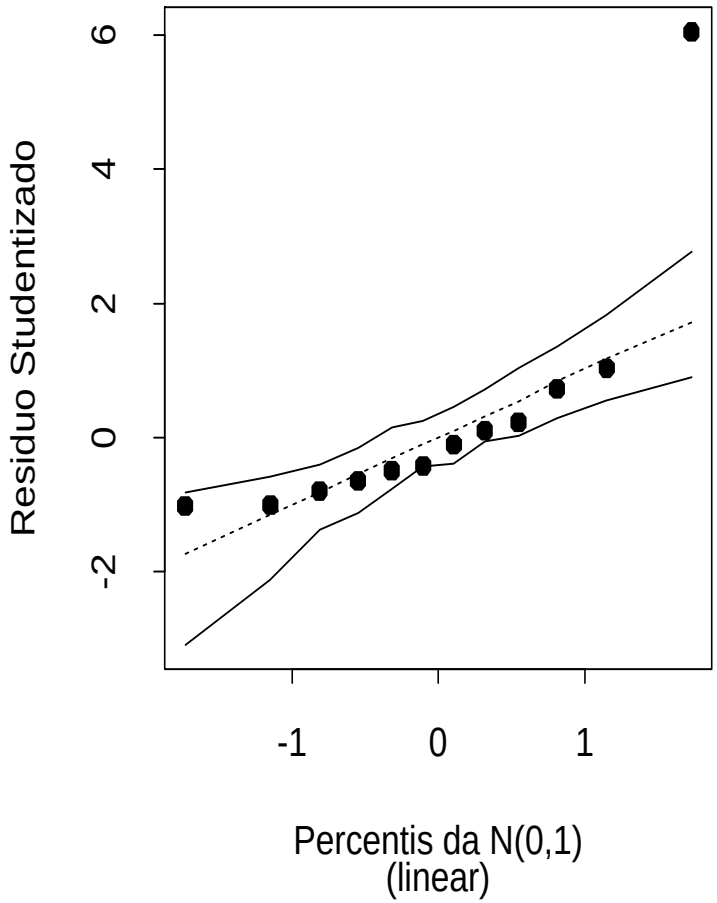
em que $\epsilon_i \sim \text{N}(0, \sigma^2)$ são erros mutuamente independentes.

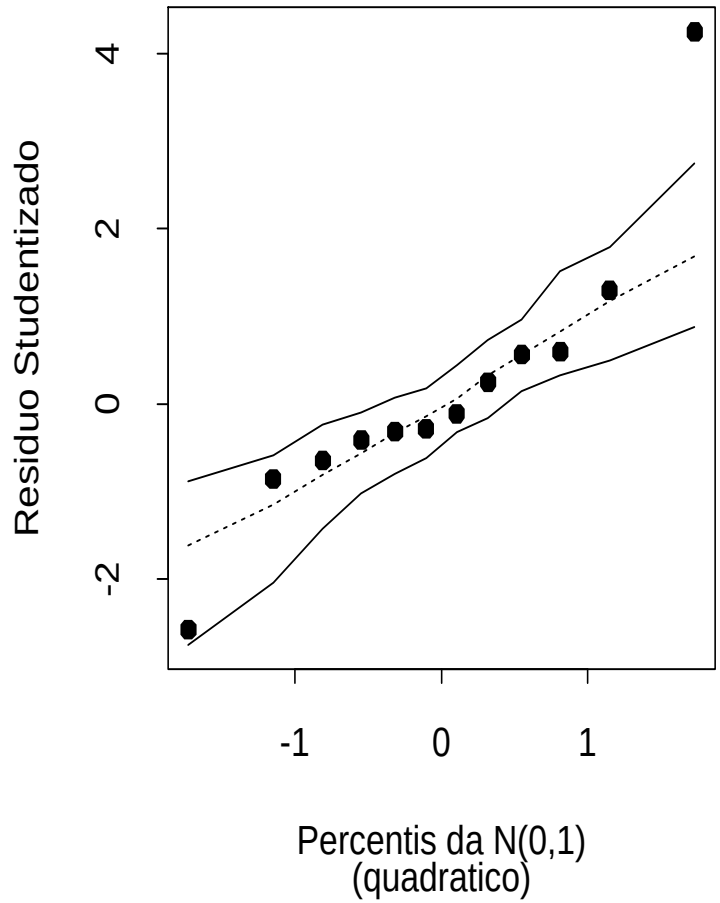
A seguir apresentamos as estimativas dos parâmetros dos modelos ajustados e gráficos de resíduos.

Estimativas dos parâmetros dos modelos normal linear e normal quadrático para explicar $\sqrt{\mathrm{Y}_i}$ em fução do tempo de exposição.

Parâmetro	Linear-$\sqrt{\mathrm{Y}}$	Quadrático-$\sqrt{\mathrm{Y}}$
α	12,57(0,38)	13,64(0,51)
β	-0,82(0,05)	-1,27(0,18)
γ		0,04(0,01)
σ	0,62	0,98
R^2	96,1%	97,8%

Resíduos modelos normais





Ajuste modelo de Poisson

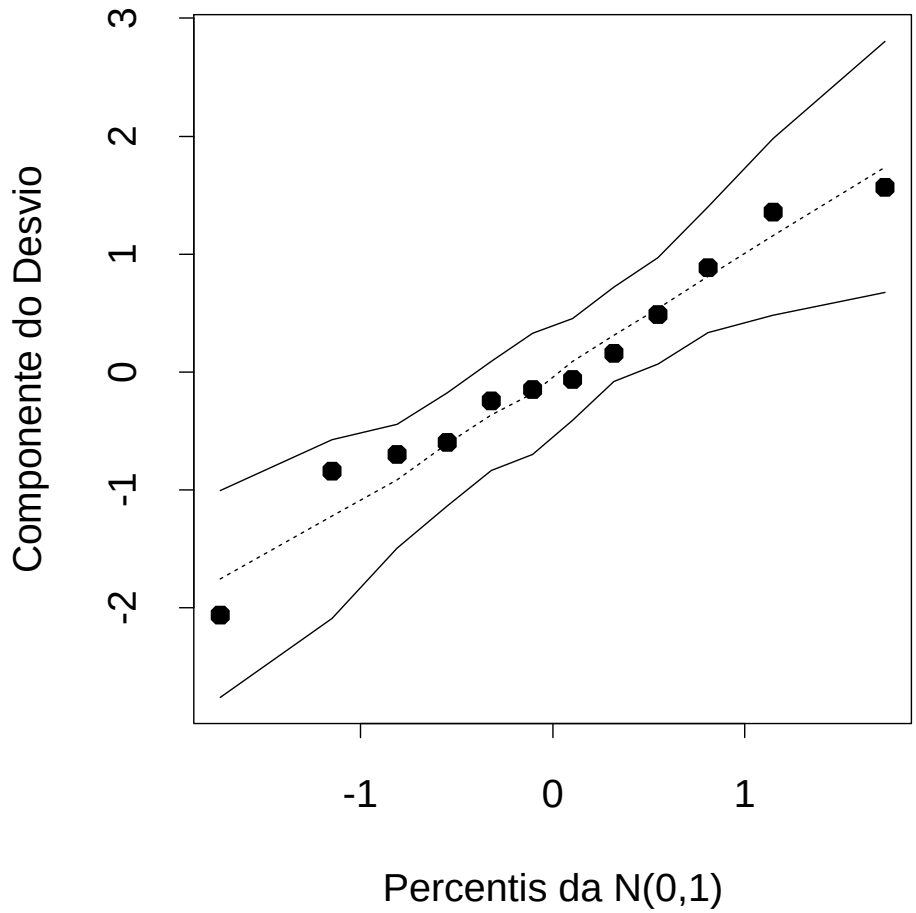
Vamos supor agora o seguinte modelo log-linear de Poisson

$$\log\mu_i = \alpha + \beta \text{tempo}_i,$$

em que $Y_i \sim \text{P}(\mu_i)$. As estimativas desse modelo são apresentadas na tabela abaixo.

Parâmetro	Estimativa	E/E.Padrão
α	5,30	88,34
β	-0,23	-23,00
Desvio		8,42 (10 g.l.)

Resíduos modelo de Poisson



Interpretação modelo de Poisson

O modelo ajustado fica então dado por

$$\hat{\mu}(x) = e^{5{,}30 - 0{,}23x},$$

em que x denota o tempo de exposição.

Interpretação modelo de Poisson

O modelo ajustado fica então dado por

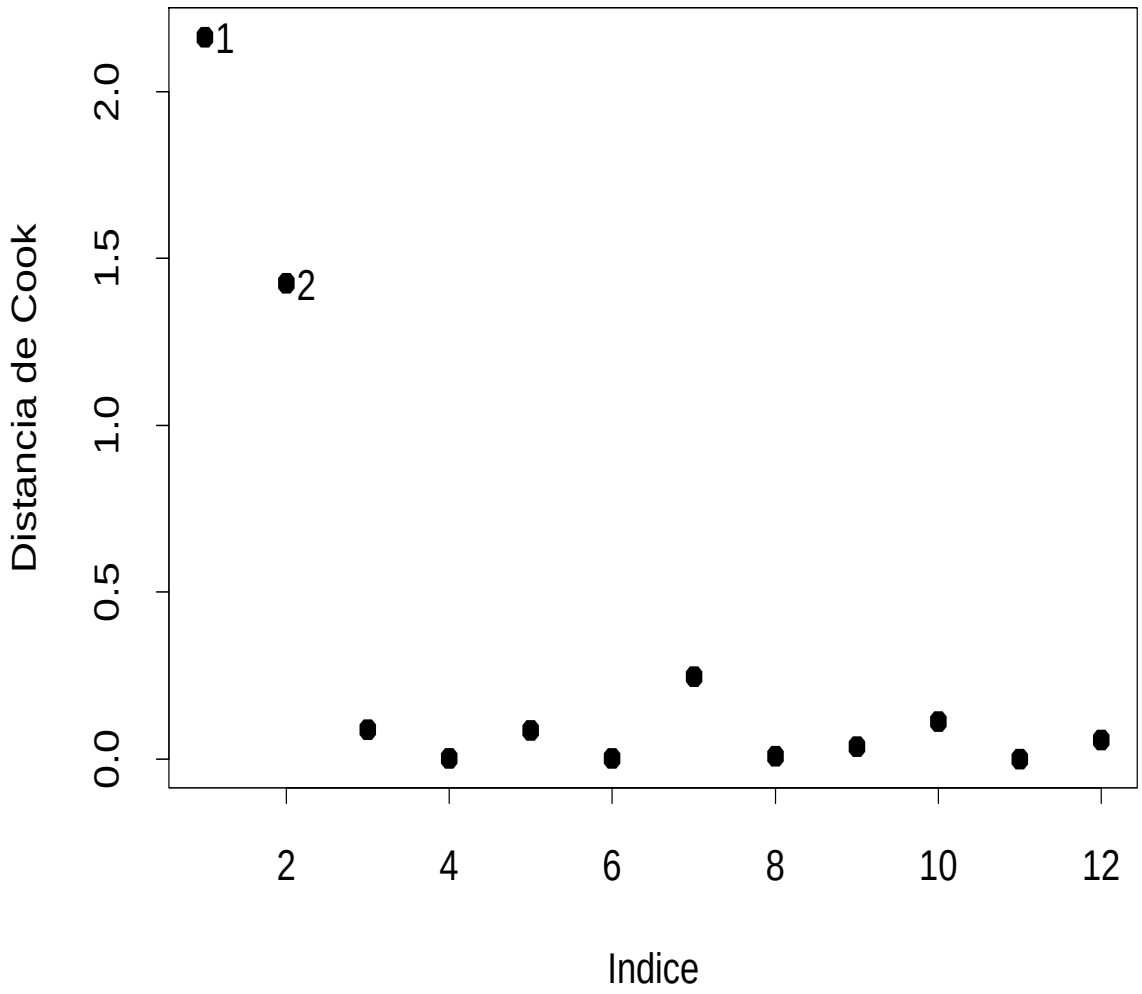
$$\hat{\mu}(x) = e^{5{,}30 - 0{,}23x},$$

em que x denota o tempo de exposição. Logo, se diminuirmos de uma unidade o tempo de exposição a variação no valor esperado fica dada por

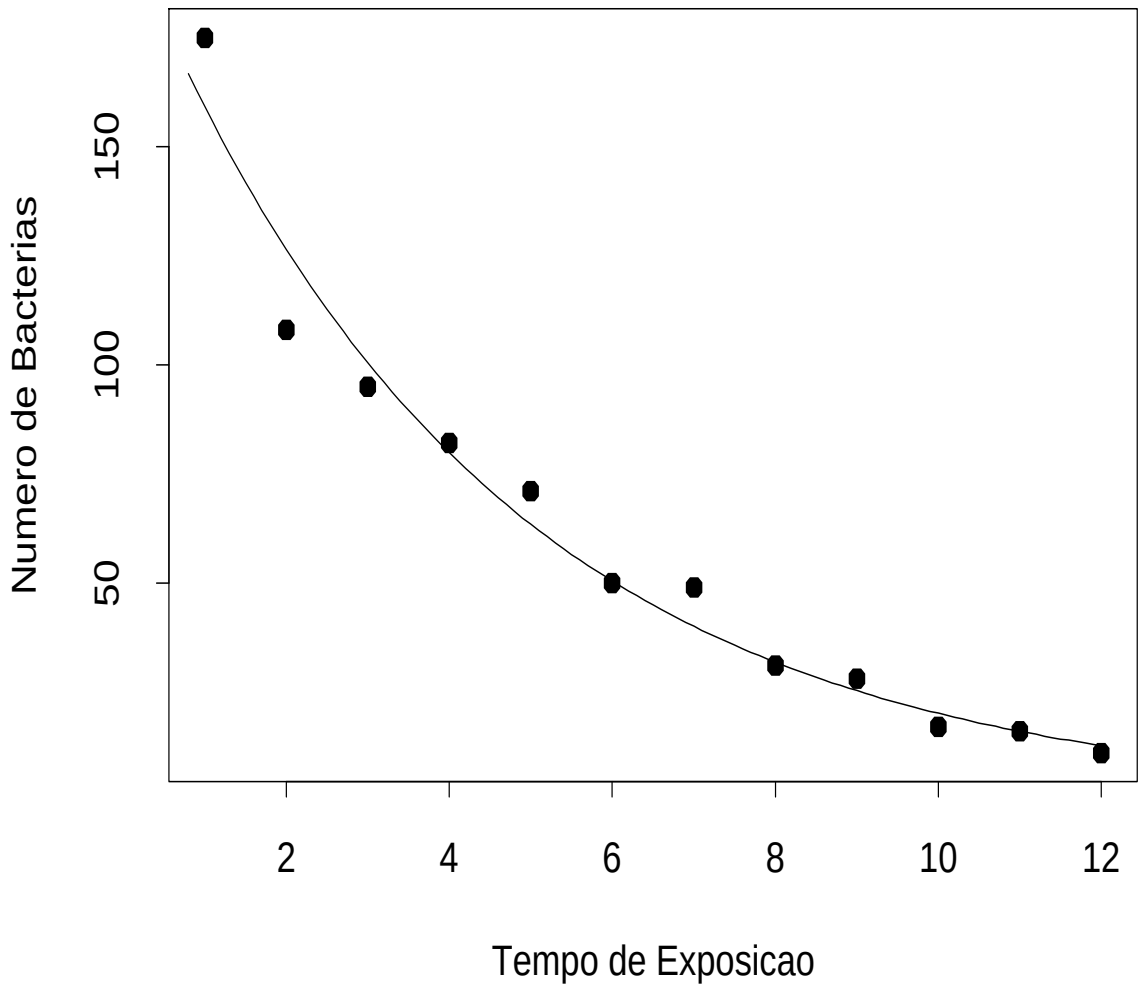
$$\frac{\hat{\mu}(x-1)}{\hat{\mu}(x)} = e^{0{,}23} = 1{,}259.$$

Ou seja, o número esperado de sobreviventes diminui aproximadamente 25,9%.

Diagnóstico modelo de Poisson



Curva ajustada modelo de Poisson



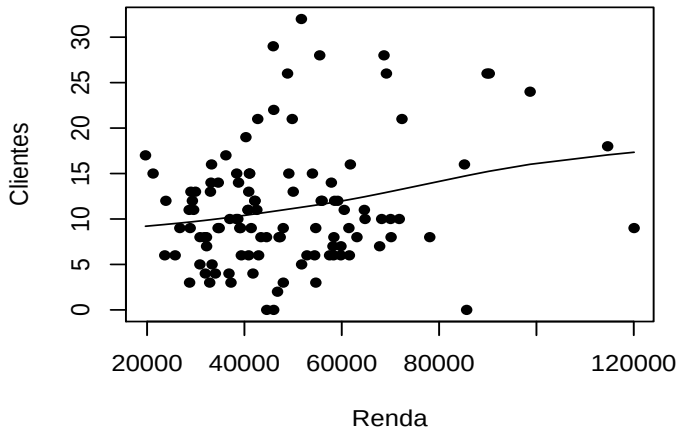
Conclusões

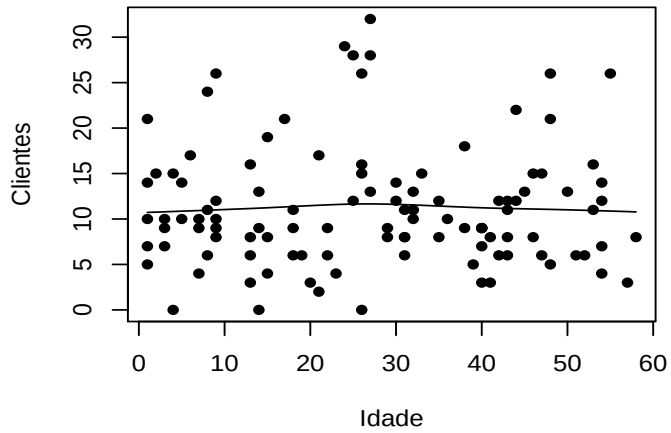
Nota-se neste exemplo a superioridade do modelo log-linear de Poisson quando comparado aos dois modelos aproximados pela distribuição normal. Essa vantagem se reflete não somente pela qualidade do ajuste mas também na interpretação dos parâmetros uma vez que a escala da variável resposta foi preservada. A retirada da observação #1, que se destaca nos gráficos de dispersão e influência , muda um pouco as estimativas dos parâmetros, contudo não muda a inferência.

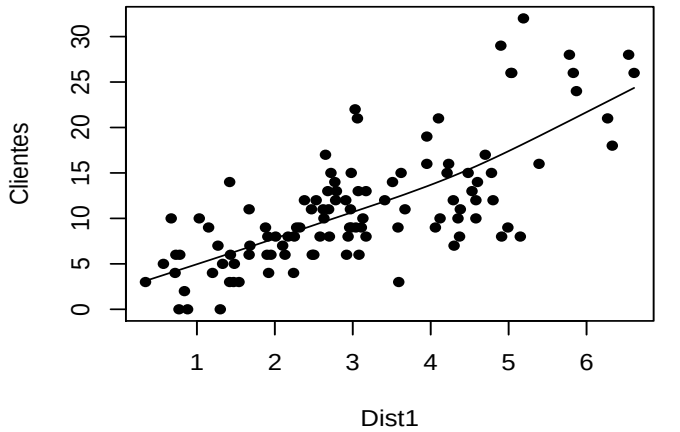
Perfil de clientes

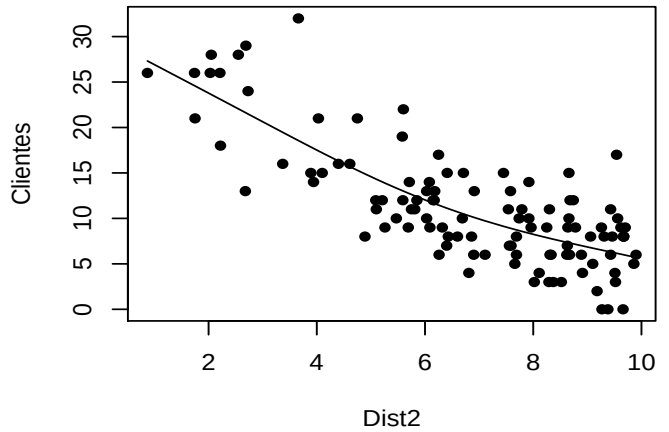
Considere os dados apresentados em Neter et al. (1996) sobre o perfil dos clientes de uma determinada loja oriundos de 110 áreas de uma cidade. O objetivo do estudo é relacionar o número de clientes em cada área (Nclientes) com as seguintes variáveis explicativas em cada área: número de domicílios (em mil) (Domic), renda média anual (em mil USD) (Renda), idade média dos domicílios (em anos) (Idade), distância ao concorrente mais próximo (em milhas) (Dist1) e distância à loja (em milhas) (Dist2).

Diagramas de dispersão









Estimativas modelo de Poisson

Supor o MLG:

(1) $\text{Nclientes}_i \sim \text{P}(\mu_i)$,

(2) $\log\mu_i = \alpha + \beta_1\text{Domic}_i + \beta_2\text{Renda}_i + \beta_3\text{Idade}_i + \beta_4\text{Dist1}_i + \beta_5\text{Dist2}_i$ $(i = 1, \ldots, 110)$.

Efeito	Parâmetro	Estimativa	E/E.Padrão
Constante	α	2,942	14,21
Domicílio	β_1	0,606	4,27
Renda	β_2	-0,012	-5,54
Idade	β_3	-0,004	-2,09
Dist1	β_4	0,168	6,54
Dist2	β_5	-0,129	-7,95

Interpretações

Nota-se que as estimativas são altamente significativas. O desvio do modelo foi de $\mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 114,98$ (104 g.l.) que equivale a P$= 0,35$ indicando ajuste adequado. Notamos pela tabela que o número esperado de clientes na loja cresce com o aumento do número de domicílios na área e da distância ao concorrente mais próximo, porém diminui com o aumento da renda média e da idade média dos domicílios bem como da distância da área à loja.

Interpretações

Nota-se que as estimativas são altamente significativas. O desvio do modelo foi de $\mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 114,98$ (104 g.l.) que equivale a P$= 0,35$ indicando ajuste adequado. Notamos pela tabela que o número esperado de clientes na loja cresce com o aumento do número de domicílios na área e da distância ao concorrente mais próximo, porém diminui com o aumento da renda média e da idade média dos domicílios bem como da distância da área à loja. Isso sugere que deve ser uma loja de conveniência.

Interpretações

Podemos notar pelas estimativas que se aumentarmos, por exemplo, em 1 mil USD a renda média dos domicílios de uma determinada área esperamos aumento relativo no número de clientes que irão à loja de $\exp(-0,012) = 0,988$. Ou seja, decrescimento de 1,2%.

Interpretações

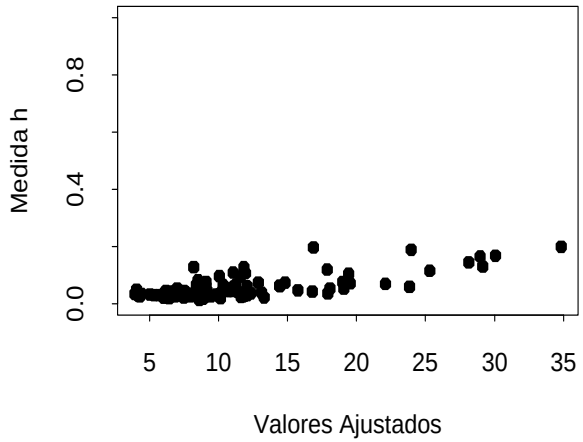
Podemos notar pelas estimativas que se aumentarmos, por exemplo, em 1 mil USD a renda média dos domicílios de uma determinada área esperamos aumento relativo no número de clientes que irão à loja de $\exp(-0,012) = 0,988$.

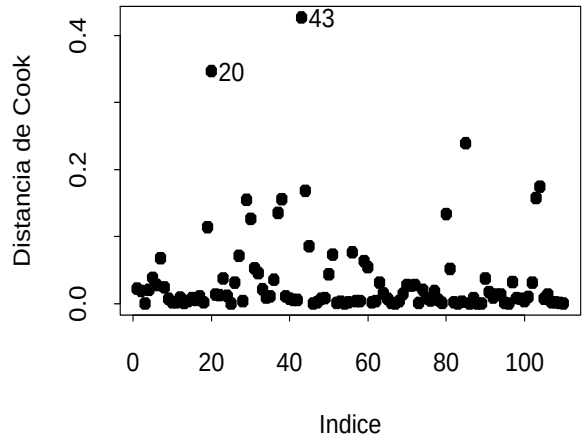
Ou seja, decrescimento de 1,2%.

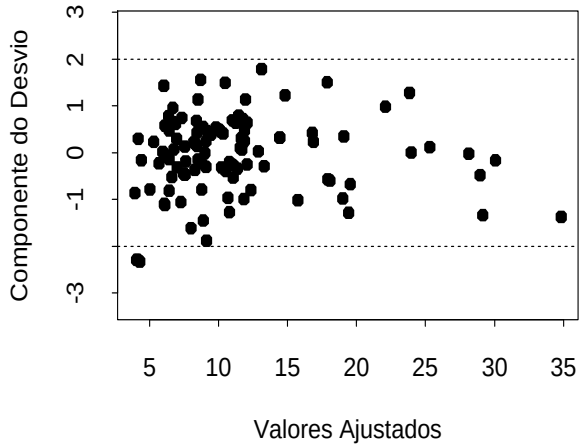
Por outro lado, se a distância ao concorrente mais próximo aumentar em uma milha esperamos aumento relativo no número de clientes que irão à loja de $\exp(0,168) = 1,183$.

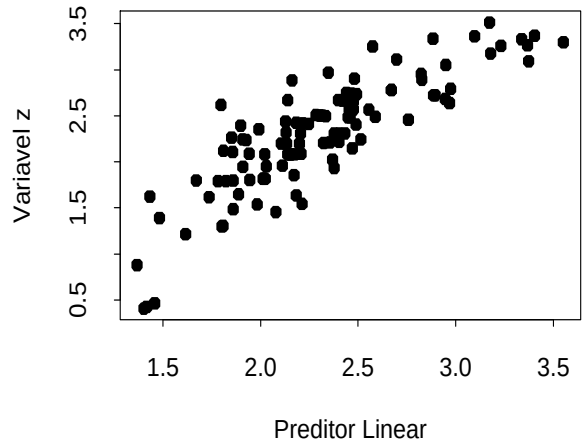
Ou seja, aumento de 18,3%.

Diagnóstico modelo de Poisson

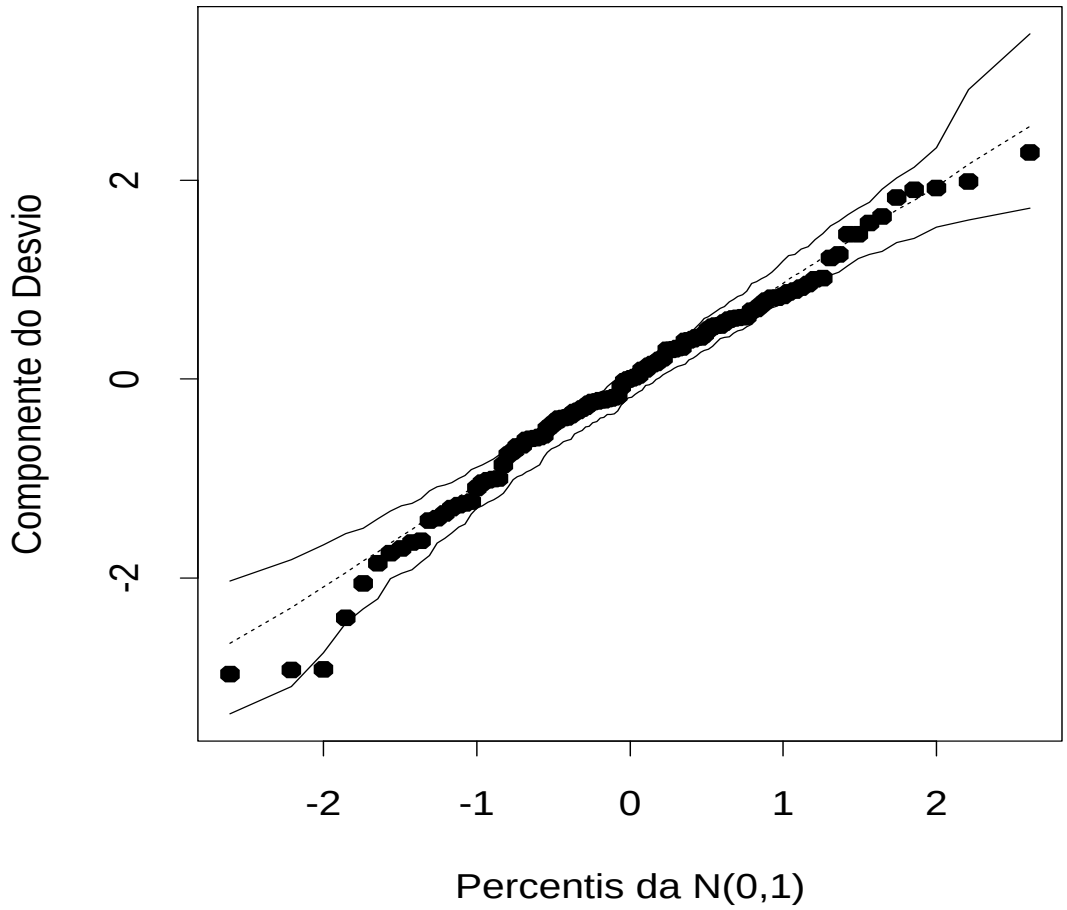








Resíduos modelo de Poisson



Conclusões

Dentre as observações destacadas pelos gráficos de diagnóstico, apenas as áreas #20 e #43 apresentam algumas variações desproporcionais nas estimativas dos parâmetros, porém não houve mudança inferencial.

Nota-se também que não há indícios de que a ligação utilizada seja inapropriada e o gráfico de resíduos não apresenta indicações de afastamentos sérios da suposição de distribuição de Poisson para o número de clientes que frequentam a loja de conveniência.

Estudo Seriado com Ratos

Os dados descritos na tabela a seguir são provenientes de um estudo seriado com 204 ratos em que um tipo de tumor maligno foi inoculado nos animais para avaliar a influência da série (passagem do tumor) na morte (caquexia) do rato (Paula, Barbosa e Ferreira, 1989).

Para cada animal foram observados as seguintes variáveis: grupo de passagem (P0 a P28), desenvolvimento de massa tumoral (sim ou não), ocorrência de caquexia (sim ou não) e tempo de observação (em dias).

Distribuição Ratos Caquéticos

Massa tumoral		Grupo de passagem P0-P6	P7-P18	P19-P28
Sim	Casos	6	13	8
	R-Dias	2597	3105	2786
Não	Casos	12	3	1
	R-Dias	1613	411	232

Portanto, dos 204 ratos acompanhados apenas 43 ficaram caquéticos e os 161 ratos restantes tiveram o tempo de sobrevida censurado.

Modelo de Poisson

Seja O_{ij} o número de ratos caquéticos no nível i de massa tumoral e grupo de passagem j $(i = 1, 2)$ e $(j = 1, 2, 3)$.

Vamos supor $O_{ij} \sim \mathrm{P}(\mu_{ij})$, em que $\mu_{ij} = \lambda_{ij} t_{ij}$ e parte sistemática dada por

$$\log \lambda_{ij} = \alpha + \beta_i + \gamma_j,$$

com as restrições $\beta_1 = 0$ e $\gamma_1 = 0$. Assim, teremos

$$\log \mu_{ij} = \log t_{ij} + \alpha + \beta_i + \gamma_j.$$

Estimativas modelo de Poisson

Efeito	Estimativa	E/E.Padrão
α	5,875	18,83
β_2	0,860	2,51
γ_2	0,334	0,92
γ_3	-0,040	-0.09

O teste da razão de verossimilhanças para testar

$H_0 : \gamma_2 = 0, \gamma_3 = 0$ contra $H_1 : \gamma_2 \neq 0$ ou $\gamma_3 \neq 0$ fica dado por

$\xi_{RV} = 1,15$ para 2 graus de liberdade (P=0,56). Logo, não rejeita-se a hipótese nula, ou seja, não foi detectado efeito de passagem.

Modelo reduzido

Eliminando-se o efeito de passagem teremos o seguinte:

$$\log\lambda_{ij} = \alpha + \beta_i,$$

com a restrição $\beta_1 = 0$. As estimativas (erro padrão) ficam dadas por: $\hat{\alpha} = -5,750(0,192)$ e $\hat{\beta}_2 = 0,802(0,315)$.

Modelo reduzido

Eliminando-se o efeito de passagem teremos o seguinte:

$$\log\lambda_{ij} = \alpha + \beta_i,$$

com a restrição $\beta_1 = 0$. As estimativas (erro padrão) ficam dadas por: $\hat{\alpha} = -5,750(0,192)$ e $\hat{\beta}_2 = 0,802(0,315)$.

Modelo alternativo:

$$\log\mu_{ij} = \delta\log t_{ij} + \alpha + \beta_i + \gamma_j,$$

em que $\hat{\delta} = 1,390(0,439)$. Para testar $H_0 : \delta = 1$ contra $H_1 : \delta \neq 1$ obtém-se $\xi_W = \{(1,390 - 1)/0,439\}^2 = 0,789$, portanto não rejeita-se H_0.

Conclusões

Foi detectado apenas efeito de desenvolvimento de massa tumoral. A função desvio do modelo final foi de $\mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 1,99$ para 4 graus de liberdade indicando um ajuste adequado.

A estimativa para a razão entre as taxas de caquexia (entre não desenvolveu e desenvolveu massa tumoral) é dada por: $\exp(0,802) = 2,23$.

Conclusões

Foi detectado apenas efeito de desenvolvimento de massa tumoral. A função desvio do modelo final foi de $\mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 1,99$ para 4 graus de liberdade indicando um ajuste adequado.

A estimativa para a razão entre as taxas de caquexia (entre não desenvolveu e desenvolveu massa tumoral) é dada por: $\exp(0,802) = 2,23$. Isso quer dizer que os ratos que não desenvolveram massa tumoral sobrevivem menos do que aqueles que desenvolveram massa tumoral!

Regressão Gama

Comparação de Turbinas de Avião

Como ilustração vamos considerar os dados descritos na tabela a seguir em que cinco tipos de turbina de avião são comparados segundo o tempo (em milhões de ciclos) até a perda da velocidade (Lawless 1982, p. 201).

Apresentamos em seguida a distribuição empírica do tempo (ignorando-se o tipo de turbina) os boxplots dos tempos para cada grupo e uma tabela com a média, desvio padrão e coeficiente de variação do tempo de duração do motor para cada tipo de turbina.

Tempo de Duração do Motor

Tipo de turbina				
Tipo I	Tipo II	Tipo III	Tipo IV	Tipo V
3,03	3,19	3,46	5,88	6,43
5,53	4,26	5,22	6,74	9,97
5,60	4,47	5,69	6,90	10,39
9,30	4,53	6,54	6,98	13,55
9,92	4,67	9,16	7,21	14,45
12,51	4,69	9,40	8,14	14,72
12,95	5,78	10,19	8,59	16,81
15,21	6,79	10,71	9,80	18,39
16,04	9,37	12,58	12,28	20,84
16,84	12,75	13,41	25,46	21,51

Figura 1. Densidade Tempo de Duração.

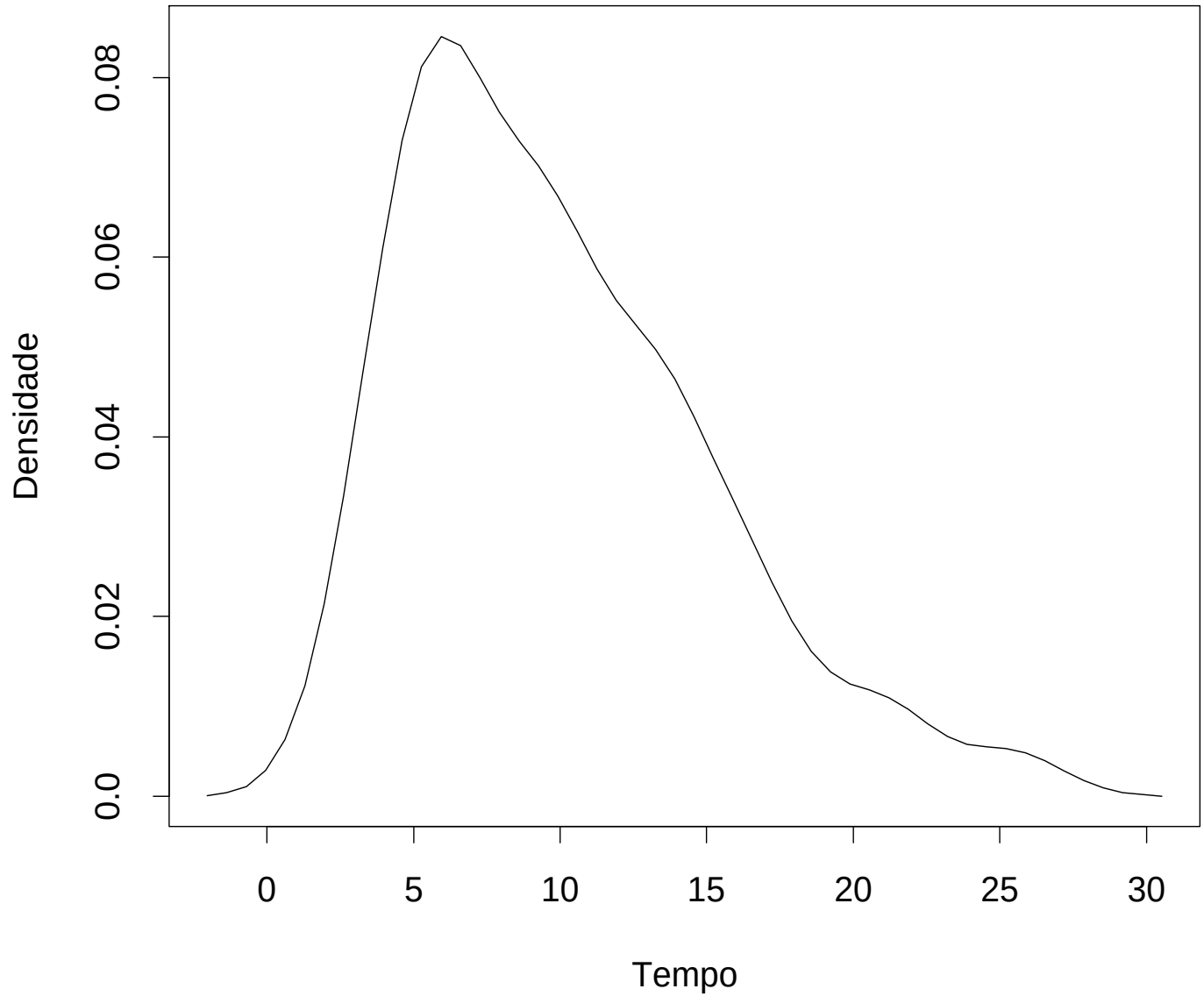
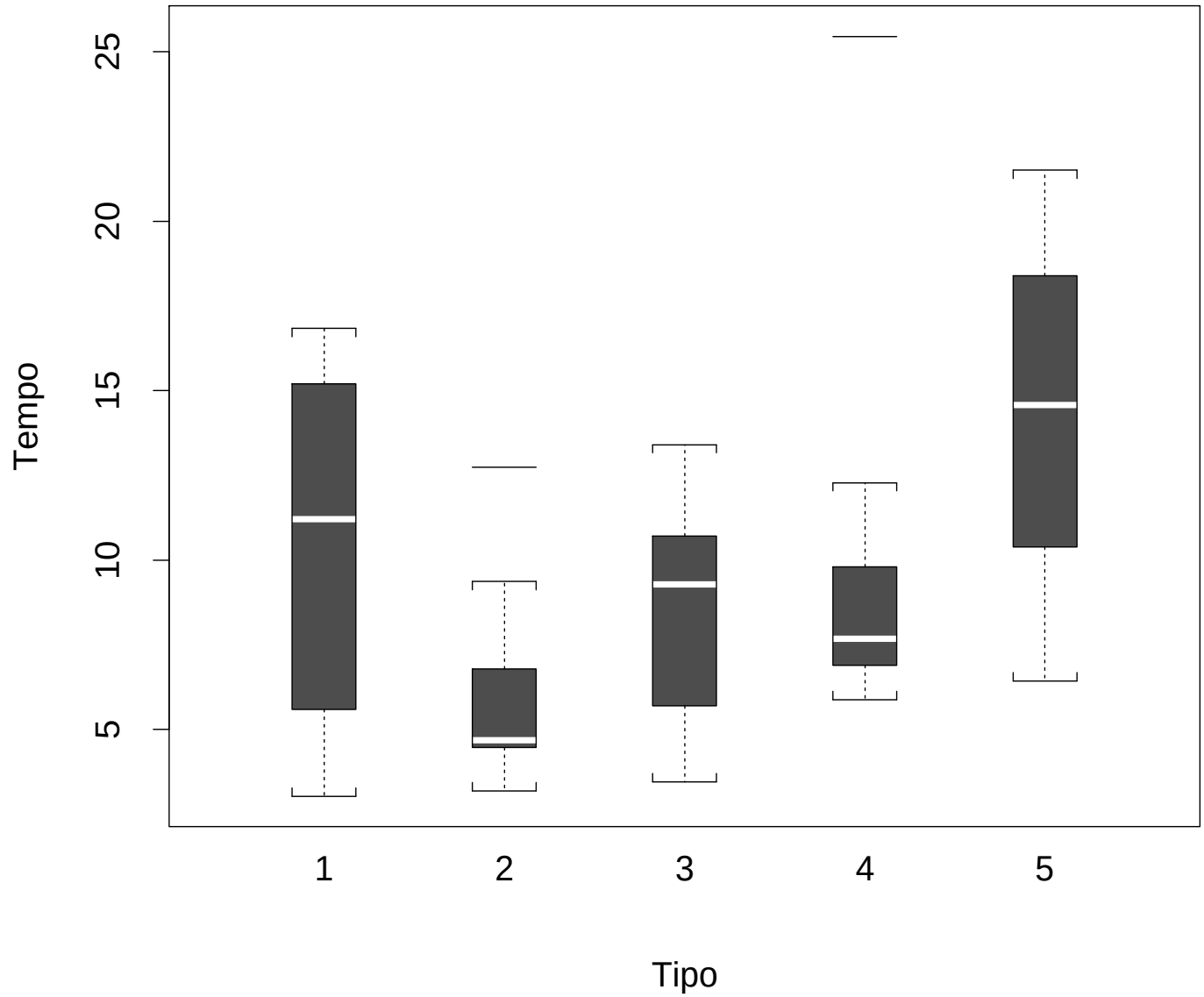


Figura 2. BoxPlots Tempo de Duração.



Análises Descritivas

Estatística	Tipo I	Tipo II	Tipo III	Tipo IV	Tipo V
Média	10,69	6,05	8,64	9,80	14,71
D.Padrão	4,82	2,91	3,29	5,81	4,86
C. Variação	45,09%	48,10%	38,08%	59,29%	33,04%

Nota-se pela Figura 1 uma assimetria no tempo de duração do motor da turbina. Pelos boxplots da Figura 2 nota-se que as variâncias dos cinco grupos são muito heterogêneas. Já pela tabela acima nota-se que os coeficientes de variação são menos heterogêneos.

Modelos Gama

Seja Y_{ij} o tempo até a perda da velocidade da j-ésima turbina do i-ésimo tipo. Vamos supor inicialmente que $Y_{ij} \sim \mathrm{G}(\mu, \phi)$, ou seja, vamos ignorar o efeito tipo.

Obtém-se as seguintes estimativas:

$$\hat{\mu} = 9,98(0,73) \text{ e } \hat{\phi} = 4,01(0,77).$$

Portanto, confirma-se pela estimativa de ϕ um certa assimetria na distribuição empírica do tempo de duração do motor das trubinas.

Vamos incluir os grupos no modelo gama:

- $Y_{ij} \sim \mathrm{G}(\mu_i, \phi)$ $(i = 1, \ldots, 5)$ $(j = 1, \ldots, 10)$
- $\mu_1 = \alpha$
- $\mu_i = \alpha + \beta_i$,

em que $i = 2, 3, 4, 5$.

Isto é, temos um modelo casela de referência em que

$\beta_2, \beta_3, \beta_4$ e β_5 são incrementos nas médias dos tipos II, III,

IV e V em relação à média do tipo I.

Estimativas dos Parâmetros

Parâmetro	Estimativa	E. Padrão	valor-Z
α	10,69	1,54	6,93
β_2	-4,64	1,77	-2,62
β_3	-2,06	1,98	-1,04
β_4	-0,89	2,09	-0,43
β_5	4,01	2,62	1,53
ϕ	5,80	1,13	5,13

O desvio do modelo é dado por

$$\mathrm{D}^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = \hat{\phi} \times \mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 5,80 \times 8,86 = 51,39 (45\ g.l.).$$

valor-P dado por P=0,24 (não rejeitamos o modelo).

Pelas estimativas dos parâmetros nota-se que os tempos médios até a perda de velocidade dos tipos I, III e IV parecem não diferir. O tipo II é aquele que apresenta o menor tempo médio, enquanto que o tipo V parece ter o maior tempo médio.

Vamos verificar através de um teste F, se é possível agrupar os tipos I, III e IV. Ou seja, vamos testar as hipóteses $\mathrm{H}: \beta_3 = \beta_4 = 0$ contra $\mathrm{A}:$ pelo menos um parâmetro diferente de zero.

Teste F

A estatística F é dada por:

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

em que $\mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}^0)$ e $\mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}})$ denotam, respectivamente, os desvios (não escalonados) sob as hipóteses H e A.

Sob a hipótese H e para ϕ grande F segue uma distribuição $\mathrm{F}_{q,(n-p)}$. Observando-se o valor u para a estatística F o nível descritivo fica dado por $\mathrm{P} = \mathrm{Pr}\{F_{q,(n-p)} \geq u\}$.

No exemplo temos que n=50, p=5 e q=2.

Aplicando o teste F encontramos o seguinte resultado:

$$\begin{aligned} \mathrm{F} &= \frac{(9,09 - 8,86)/2}{8,86/45} \\ &= 0,58. \end{aligned}$$

O valor-P neste caso, obtido numa distribuição $F_{2,45}$ é dado por P=0,56.

Portanto, não rejeitamos a hipótese H e podemos agrupar os tipos de turbina I, III e IV. Vamos então considerar apenas três grupos.

Estimativas dos Parâmetros

Vamos assumir então que $\mu_1 = \mu_3 = \mu_4 = \alpha$, $\mu_2 = \alpha + \beta_2$ e $\mu_5 = \alpha + \beta_5$. As novas estimativas são dadas abaixo:

Parâmetro	Estimativa	E. Padrão	valor-Z
α	9,71	0,81	12,01
β_2	-3,66	1,19	-3,08
β_5	5,00	2,27	2,20
ϕ	5,66	1,10	5,14

O desvio do modelo foi de

$$\mathrm{D}^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = \hat{\phi} \times \mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 5,66 \times 9,09 = 51,45 \ (47 g.l.).$$

valor-P dado por P=0,30 (não rejeitamos o modelo).

Figura 3. Distância Cook Modelo Gama Turbina.

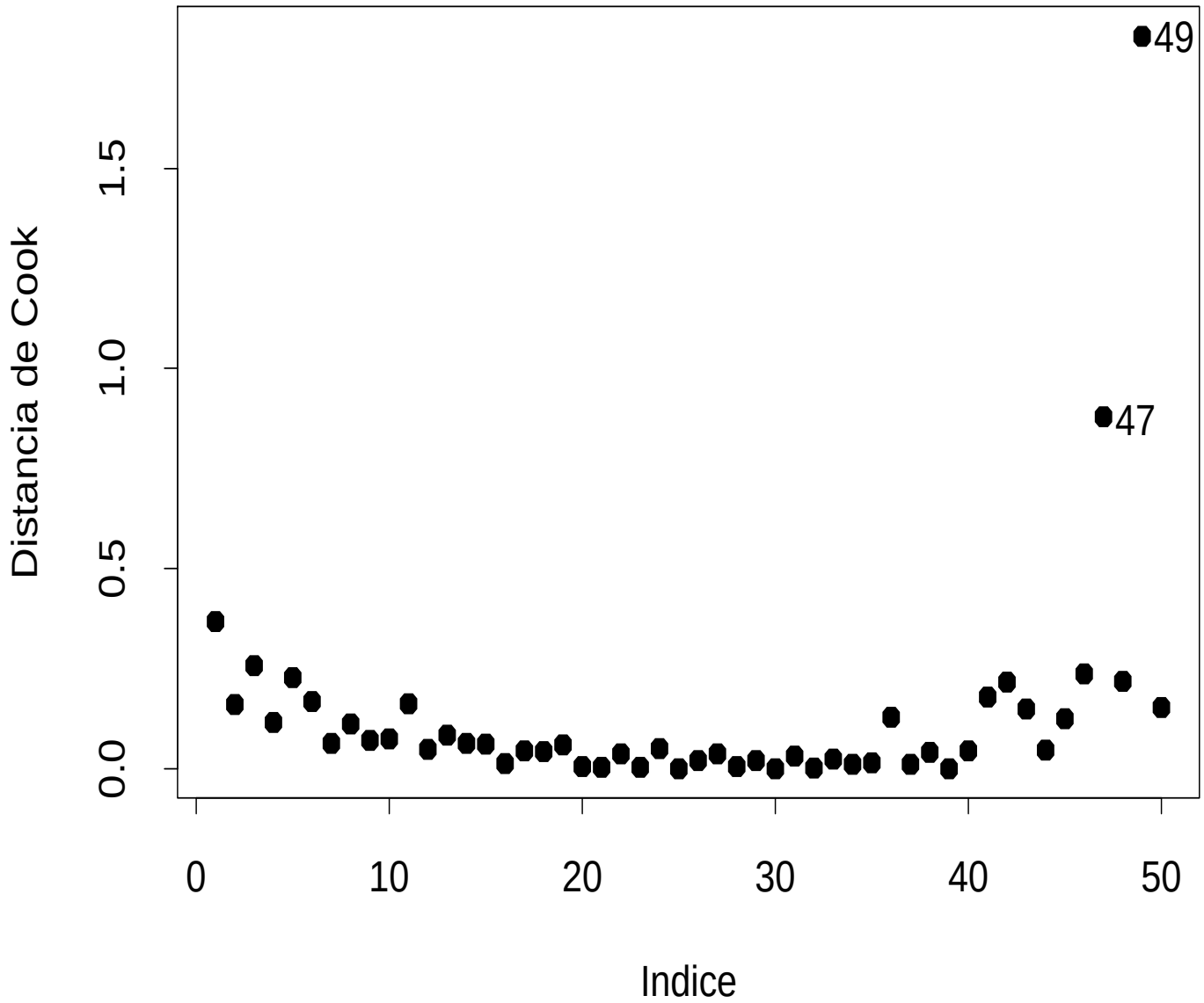
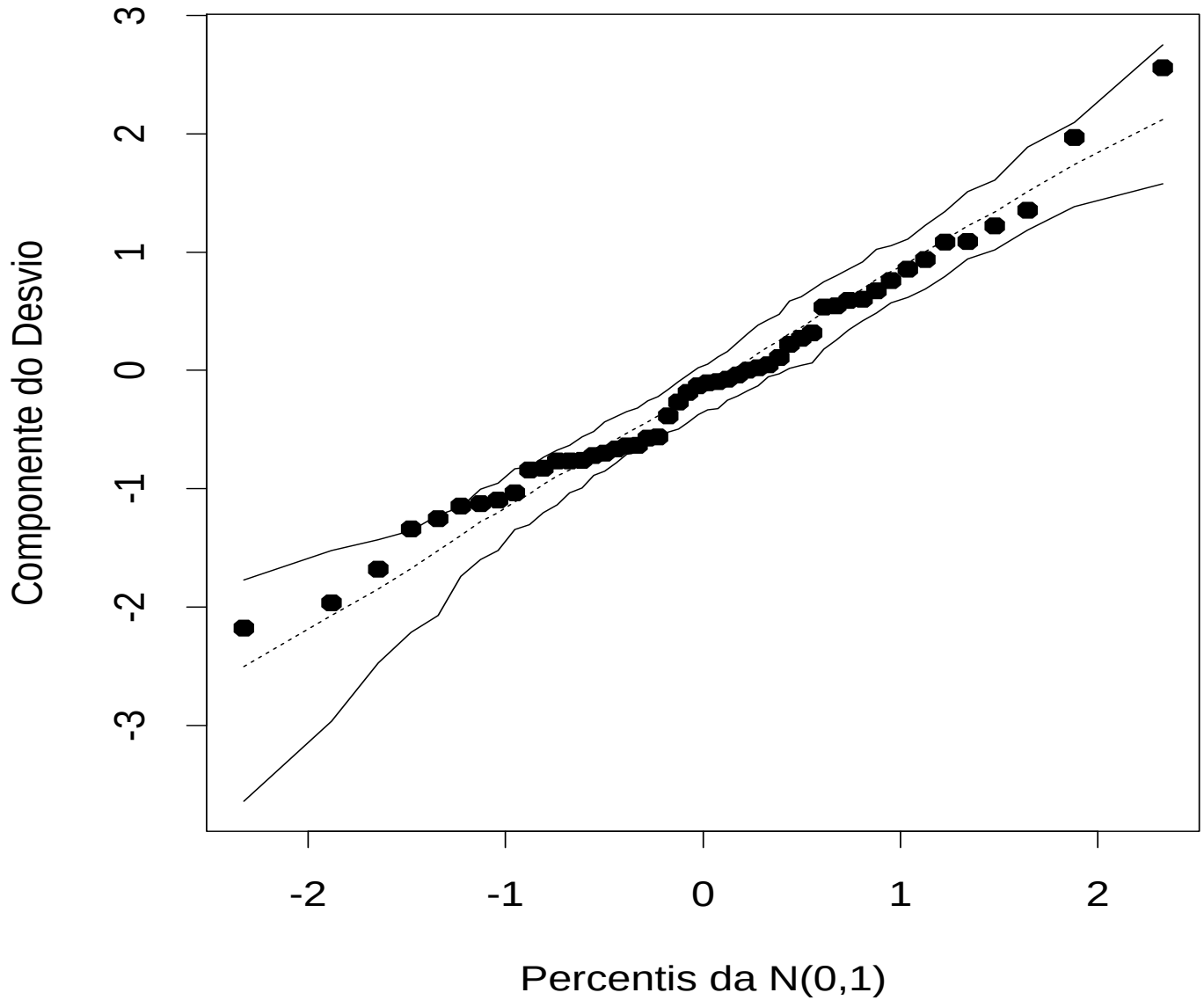


Figura 4. Envelope Modelo Gama Turbina.

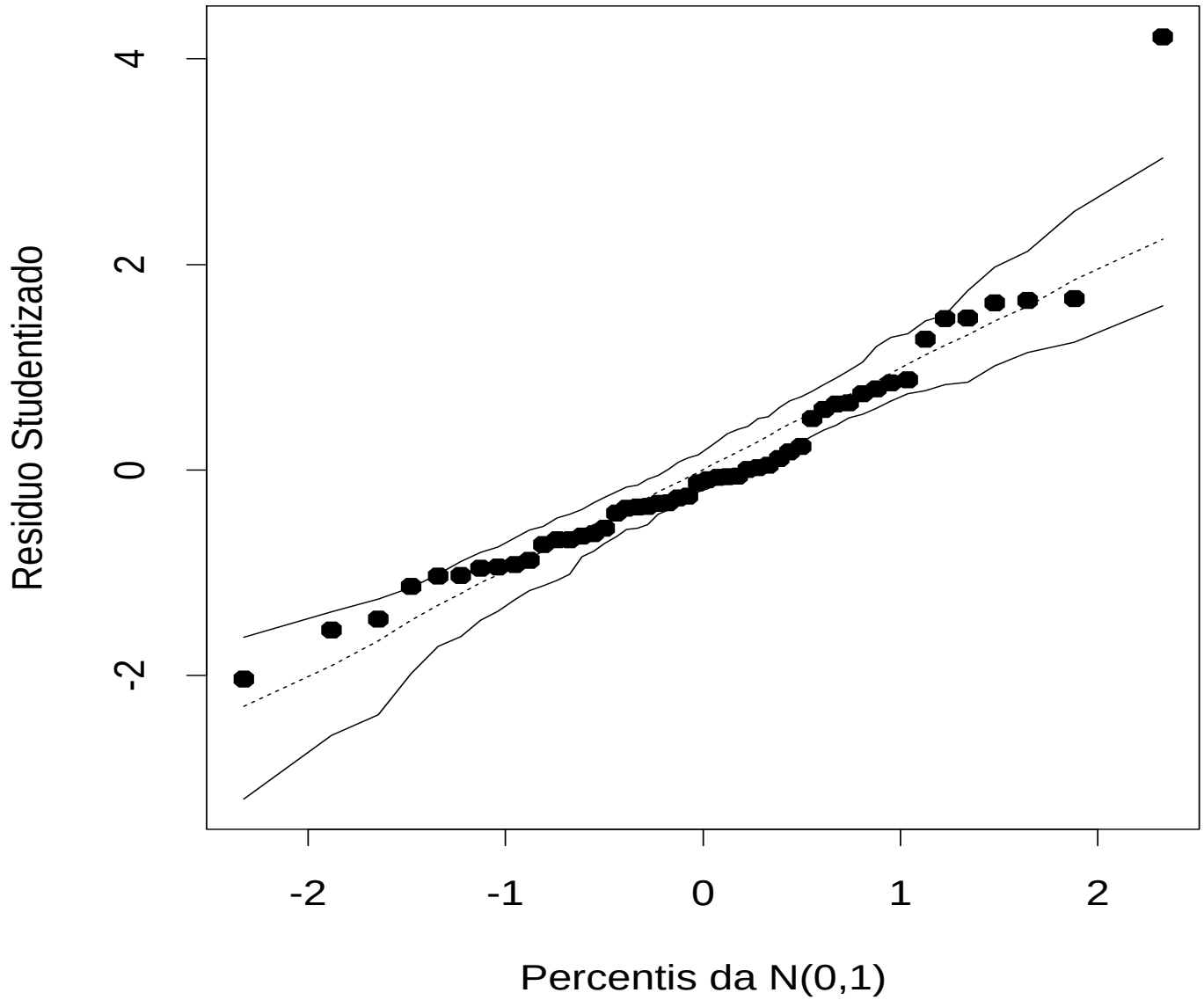


Eliminação Pontos Influentes

Os pontos 47 e 49 aparecem como influentes. Ambos referem-se aos maiores tempos dos tipos II e IV, respectivamente. Abaixo são apresentadas as variações nas estimativas eliminando-se esses pontos.

Estimativa	Sem 47	Sem 49
$\hat{\alpha}$	0%	-5%
$\hat{\beta}_2$	-22%	14%
$\hat{\beta}_5$	0%	10%
$\hat{\phi}$	8%	16%

Figura 5. Envelope Modelo Normal Turbina.



Conclusões

Neste exemplo nota-se que os 5 tipos podem ser agrupados em 3 tipos. Os tipos I, III e IV com o mesmo desempenho médio e os tipos II e V com o menor e maior desempenho médio, respectivamente. A suposição de distribuição gama para o tempo de duração parece bastante razoável e as observações detectadas como pontos influentes não mudam as conclusões quando são eliminadas do estudo. A Figura 5 mostra que a suposição de modelo normal homocedástico não é razoável.

Referência

- Goñi, R., Alvarez, F. e Adlerstein, S. (1999). Application of generalized linear modeling to cath rate analysis of western mediterranean fisheries: the Castellón trawl fleet as a case study. *Fisheries Research* **42**, 291-302.
- Paula, G. A. e Oshiro, C. H. (2001). Relatório de Análise Estatística sobre o Projeto: *Análise de Captura por Unidade de Esforço do Peixe-Batata na Frota Paulista*. RAE-CEA0102, IME-USP.