Modelos Lineares Generalizados

Gilberto A. Paula

Instituto de Matemática e Estatística

Universidade de São Paulo

e-mail:giapaula@ime.usp.br

home-page:http://www.ime.usp.br/∼giapaula/mlgs.html

Dados Binários Agrupados 1

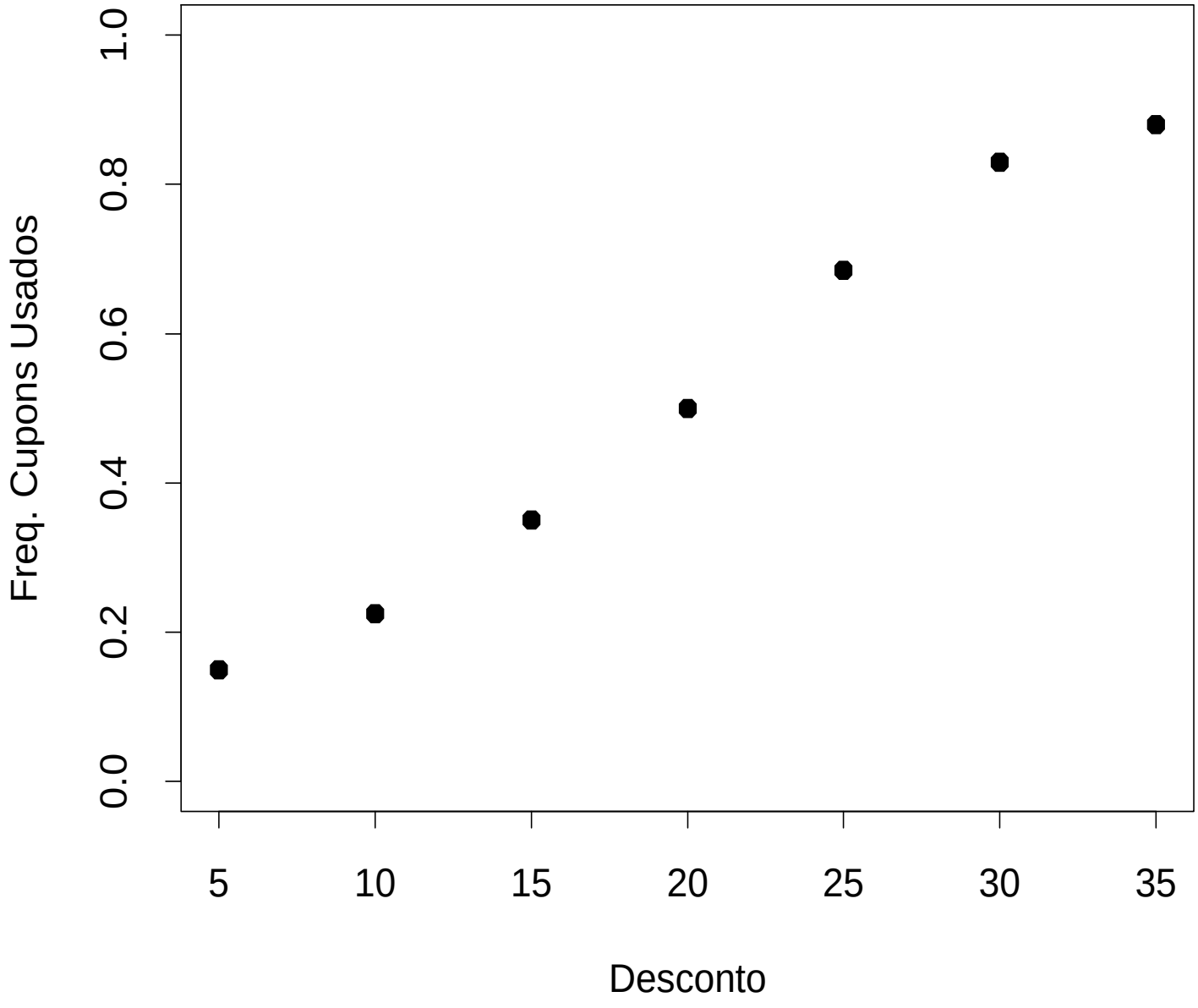
Como ilustração neste tópico vamos considerar os dados sobre o uso de cupons com descontos, enviados para clientes de uma rede de supermercados. Cupons com descontos de 5, 10, 15, 20, 25, 30 e 35 reais são enviados a clientes da rede de supermercados escolhidos aleatoriamente e deseja-se estimar a probabilidade 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.

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 estimada do cupom com desconto x ser usado. O desvio do modelo é dado por $\mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2,16$ (5 g.l.), obtendo-se o P-valor 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 estimada do cupom com desconto x ser usado. O desvio do modelo é dado por $\mathrm{D}(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2,16$ (5 g.l.), obtendo-se o P-valor 0,83 que indica que o modelo está bem ajustado.

Define-se $\frac{\mu(x)}{1-\mu(x)}$ como sendo a chance do cupom com desconto x ser usado.

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:

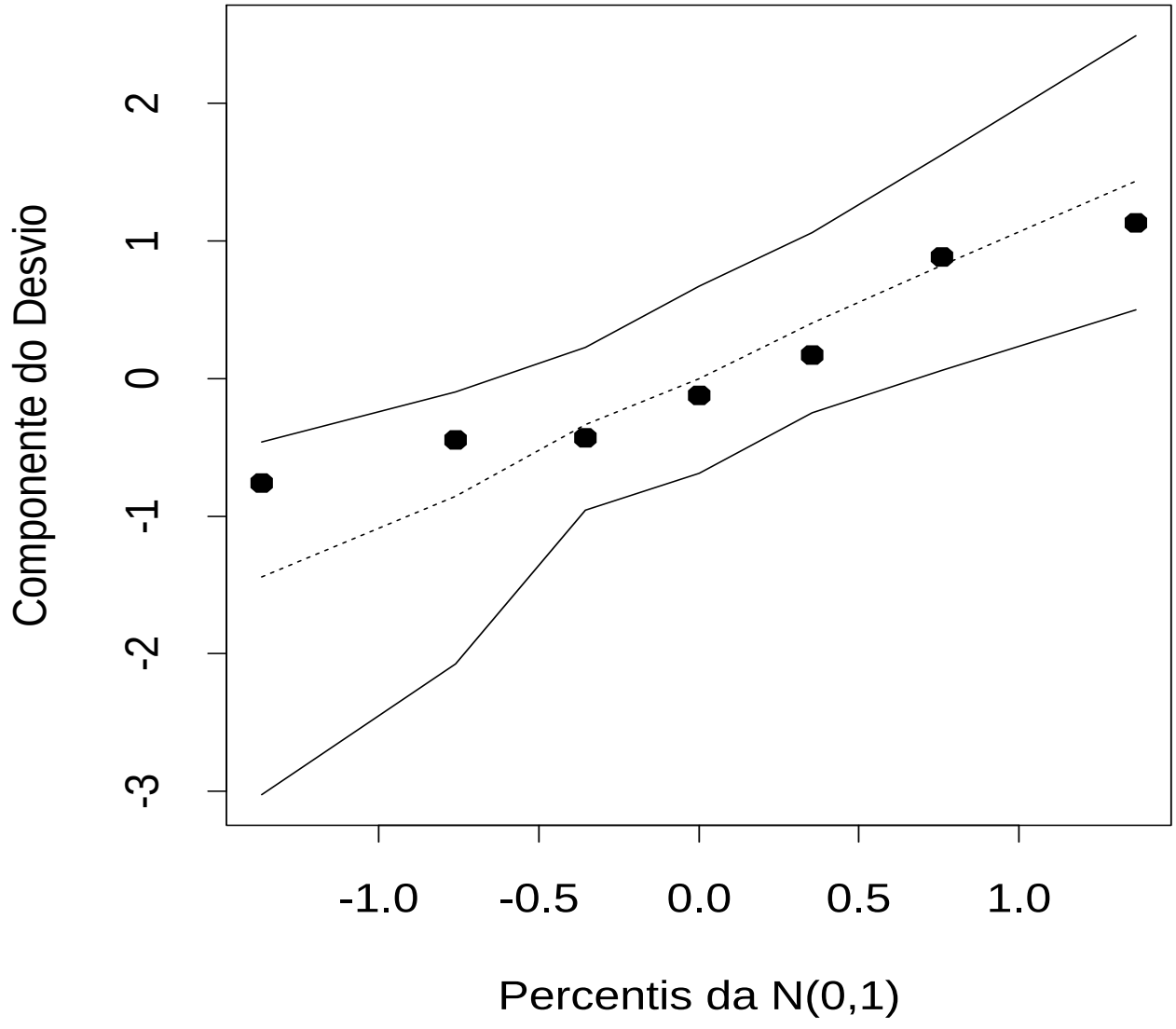
$$\begin{aligned}
\hat{\psi}(x) &= \frac{\exp\{-2,535 + 0,132(x+1)\}}{\exp\{-2,535 + 0,132x\}} \\
&= \exp(0,132) \\
&= 1,14.
\end{aligned}$$

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.

Estudo de Seguimento

Distribuição dos camundongos segundo o sexo e a ocorrência ou não de tumor após as 85 semanas:

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

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/D.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. 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) (Foster, Stine & Waterman, 1998, pp. 338-339).

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. 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) (Foster, Stine & Waterman, 1998, pp. 338-339). A seguir tem-se algumas análises descritivas.

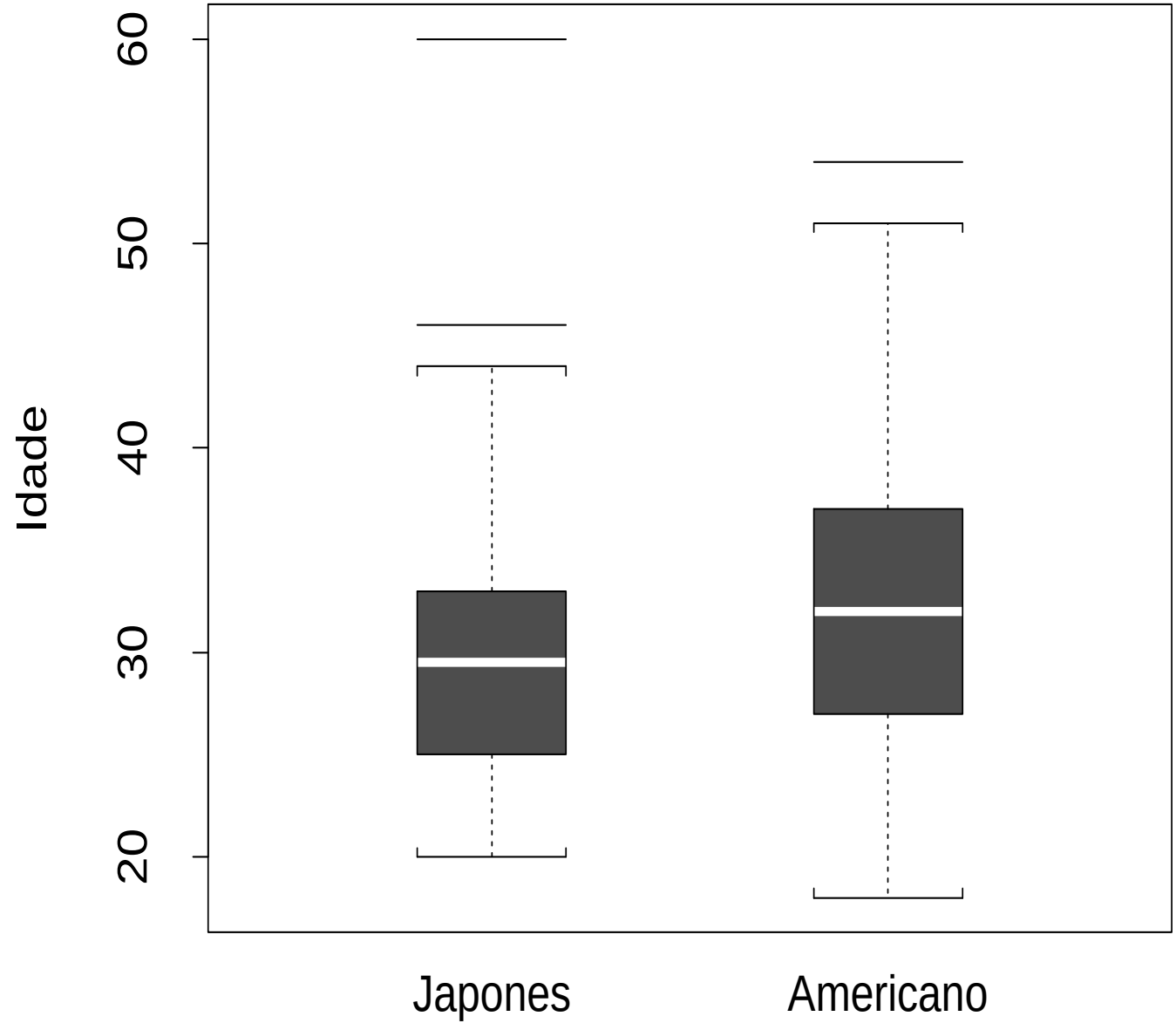
Preferência por Sexo e E. Civil

	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.Padrão	z-valor
Constante	-1,559	0,701	-2,22
Idade	0,050	0,022	2,31
Sexo	-0,094	0,256	-0,37
E.Civil	-0,518	0,272	-1,90

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

Efeito	Estimativa	E.Padrão	z-valor
Constante	-1,600	0,692	-2,31
Idade	0,049	0,021	2,30
E.Civil	-0,526	0,272	-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 P-valor 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.

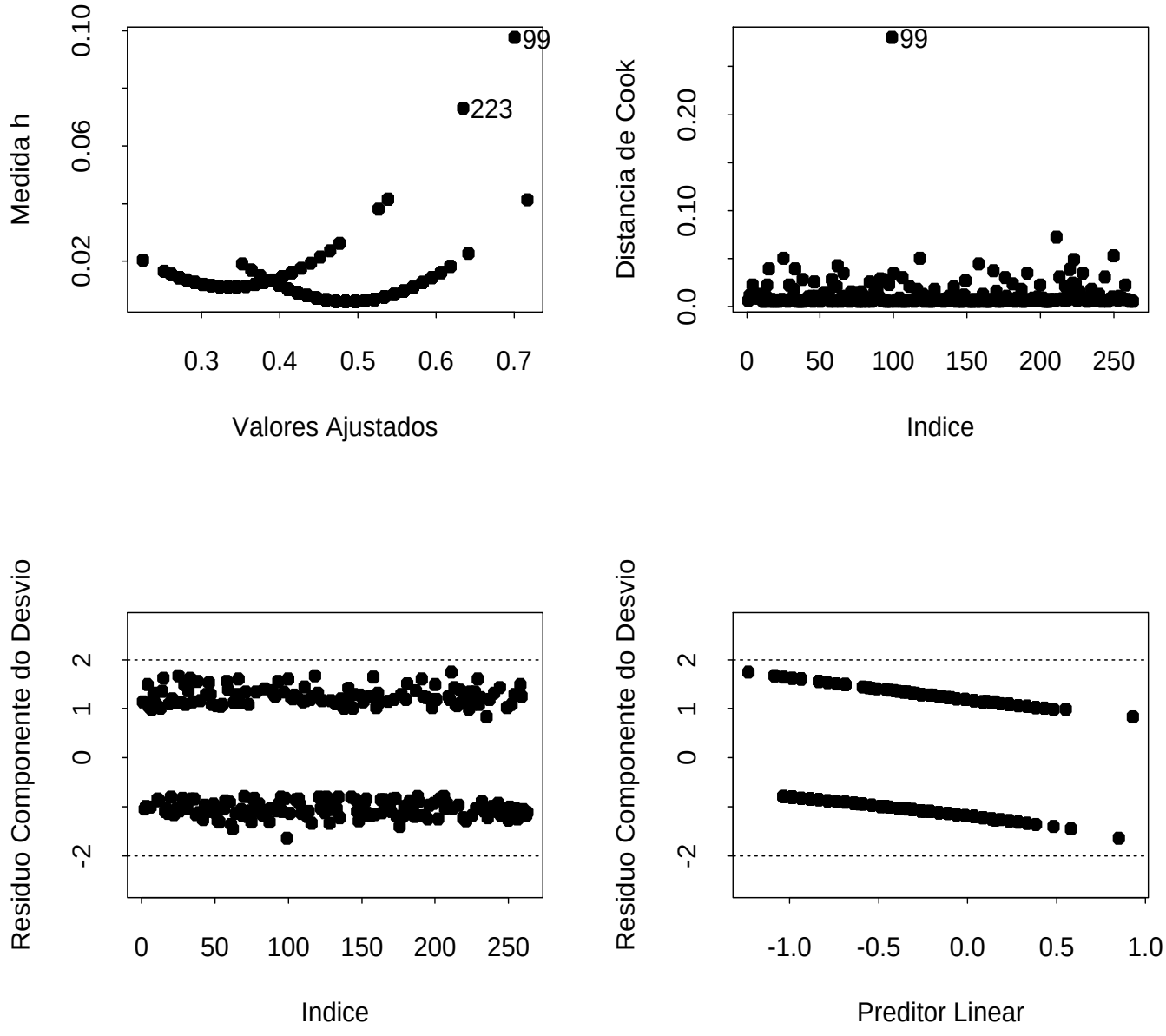
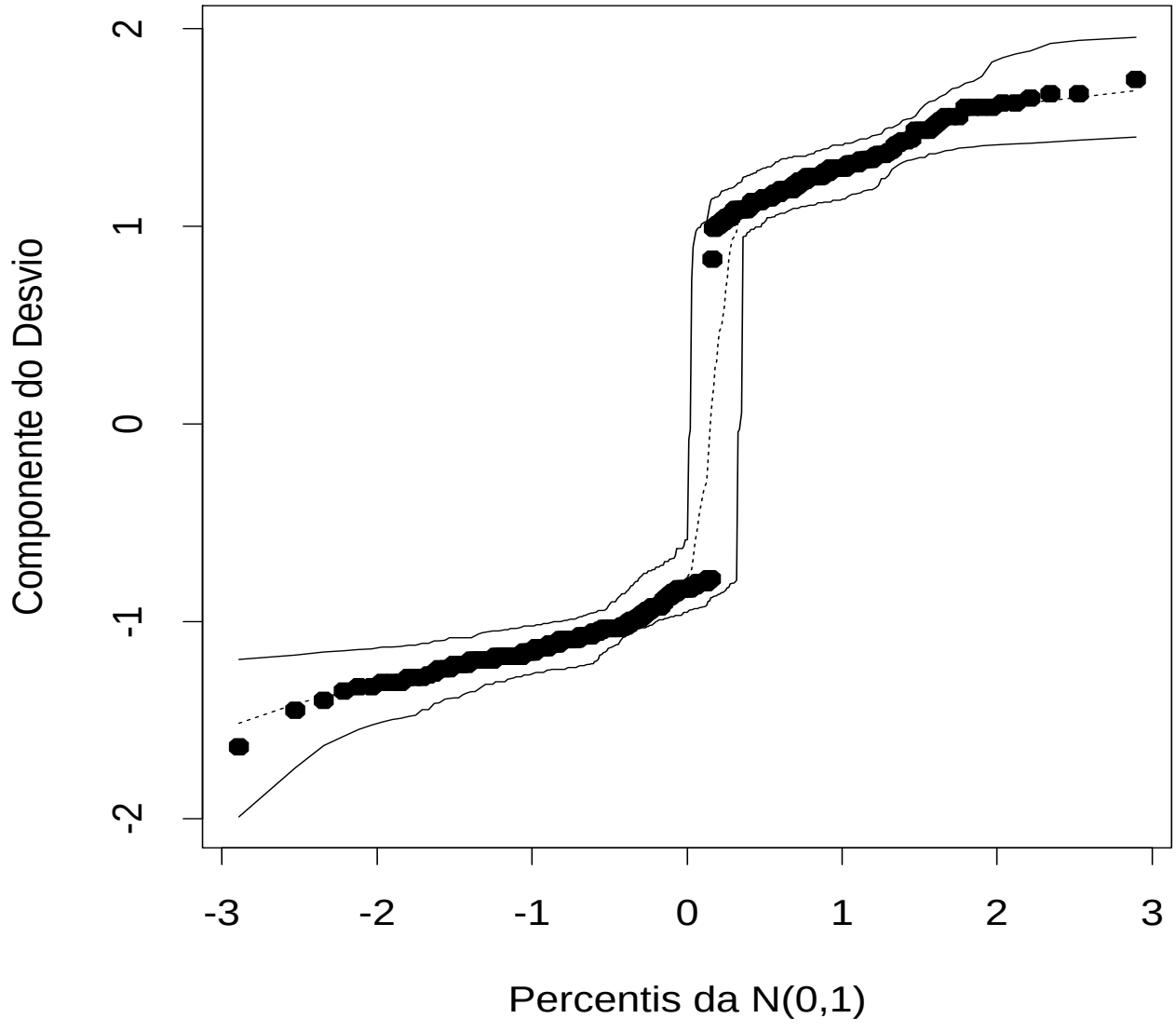


Figura 5. Envelope Exemplo Preferência.



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.