Modelos de regressão para dados de contagem inflacionados de zeros

Prof. Caio Azevedo

Exemplo 16: Número de artigos produzidos por alunos de Doutorado em bioquímica

- Corresponde aos dados relacionados à 915 alunos de pós-graduação em bioquímica.
- Disponível no pacote "'pscl' do R, sob o nome "bioChemists".
- Variaveis medidas:
  - art: quantidade de artigos produzidos nos últimos 3 anos de doutorado pelo aluno (Ph.D.) (resposta).
  - fem: gênero do aluno - masculino ou feminino (explicativa).
  - mar: estado civil - solteiro ou casado (explicativa).

Exemplo 16: Número de artigos produzidos por alunos de Doutorado em bioquímica

- kid5: número de filhos com 5 ou menos anos de idade (explicativa).
- phd: prestígio do departamento onde o aluno desenvolveu seus estudos (valor observado entre 0 e 5) (explicativa).
- ment: quantidade de artigos produzidos nos últimos de 3 anos pelo orientador (explicativa)

Comentários

- Espera-se uma concentração (inflacionamento) de zeros, dado que, em geral, os alunos tendem a publicar artigos depois de ter finalizado o Doutorado.
- Neste caso, os modelos de regressão para dados de contagem Poisson e Binomial-negativo, por exemplo, podem não ser apropriados.
- Alternativa: modificar esses modelos a fim de contemplar o inflacionamento de zeros.

Modelos probabilísticos

- Seja Y a vad (variável aleatória discreta) que representa a contagem de interesse.
- $P(Y = y) = g_Y(y) =$
  $[\pi + (1-\pi)f_Z(0)]\,\mathbb{1}_{\{0\}}(y) + (1-\pi)f_Z(y)\mathbb{1}_{\{1,2,\ldots\}}$, em que $f_Z(.)$ representa a função de probabilidade de uma vad discreta de interesse (Poisson, geométrica, binomial negativa, etc).
- Notação: $Y \sim IZ\{\pi, f_Z(y)\}$ em que $f_Z(.)$ representa o modelo probabilístico original. Por exemplo $Y \sim IZ\{\pi, \text{Poisson}(\mu)\}$.
- Exemplo:
  $$g_y(y) = [\pi + (1-\pi)e^{-\mu}]\,\mathbb{1}_{\{0\}}(y) + (1-\pi)\frac{e^{-\mu}\mu^y}{y!}\mathbb{1}_{\{1,2,\ldots\}}$$

Modelos probabilísticos

- Neste caso π representa a probabilidade de inflacionamento.
- Note que

$$\begin{aligned}\sum_{y=0}^{\infty} g(y) &= \pi + (1-\pi)f(0) + (1-\pi)\sum_{y=1}^{\infty}[f(y)] \\ &= \pi + (1-\pi)f(0) + (1-\pi)(1-f(0)) \\ &= \pi + (1-\pi)\end{aligned}$$

Modelos probabilísticos

- Valor esperado

$$\mathcal{E}(Y) = \sum_{y=0}^{\infty} yg(y) = (1-\pi)\sum_{y=1}^{\infty} yf(y) = (1-\pi)\mathcal{E}(Z)$$

em que $\mathcal{E}(Z)$ representa a esperança da variável original (não inflacionada).

- Poisson & binomial-negativa: $\mathcal{E}(Y) = (1-\pi)\mu$.

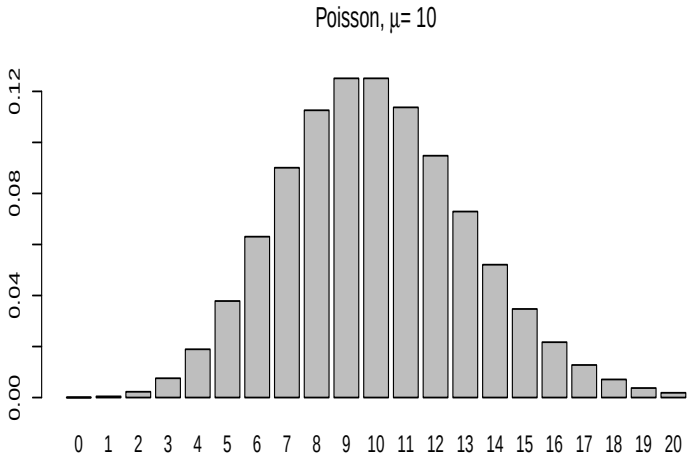
Modelos probabilísticos

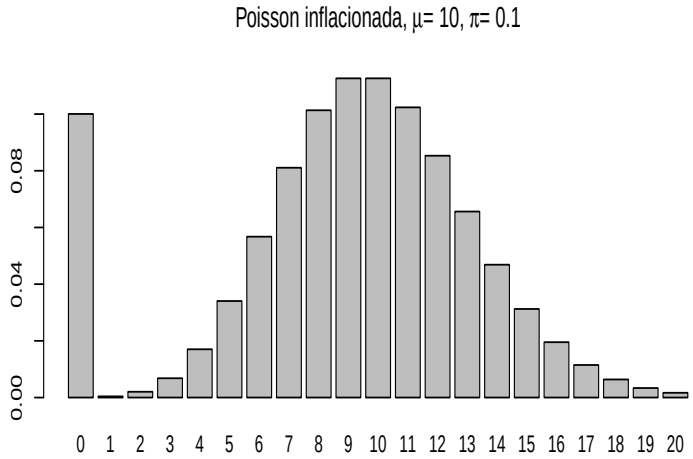
- Variância

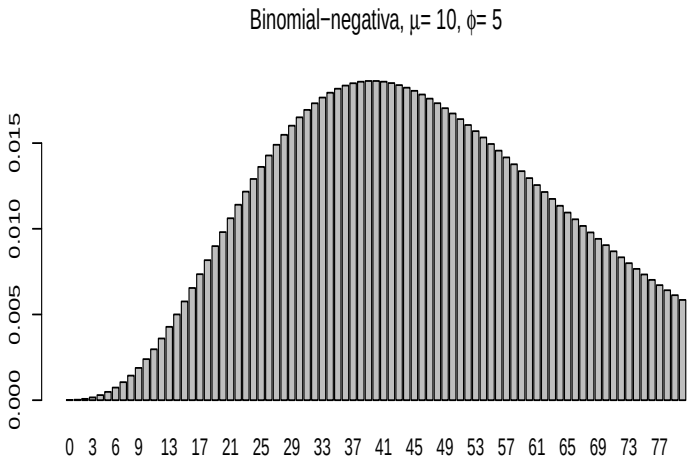
$$\begin{aligned}
\mathcal{V}(Y) &= \mathcal{E}(Y^2) - \mathcal{E}^2(Y) = \sum_{y=0}^{\infty} y^2 g(y) - (1-\pi)^2 \mathcal{E}^2(Z) \\
&= (1-\pi) \sum_{y=1}^{\infty} y^2 g(y) - (1-\pi)^2 \mathcal{E}^2(Z) \\
&= (1-\pi)\mathcal{E}(Z^2) - (1-\pi)^2 \mathcal{E}^2(Z) \\
&= (1-\pi)\left[\mathcal{E}(Z^2) - (1-\pi)\mathcal{E}^2(Z)\right]
\end{aligned}$$

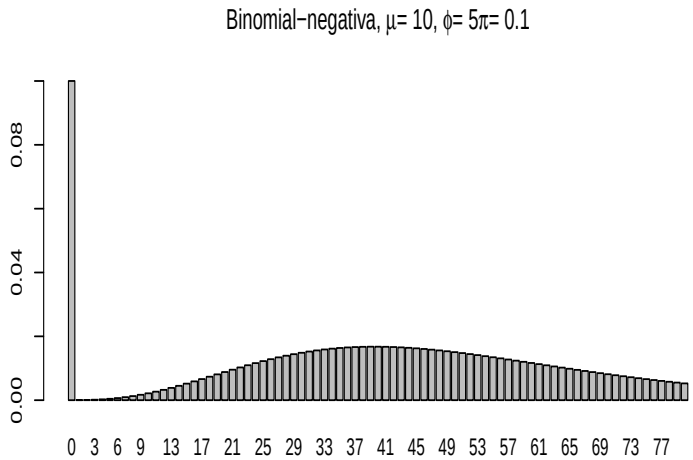
- Poisson: $\mathcal{V}(Y) = (1-\pi)\mu(1+\mu\pi)$.
- Binomial-negativa: $\mathcal{V}(Y) = (1-\pi)\mu\left(1 + \frac{\mu}{\phi} + \mu\pi\right)$.

Exemplos de funções de probabilidade

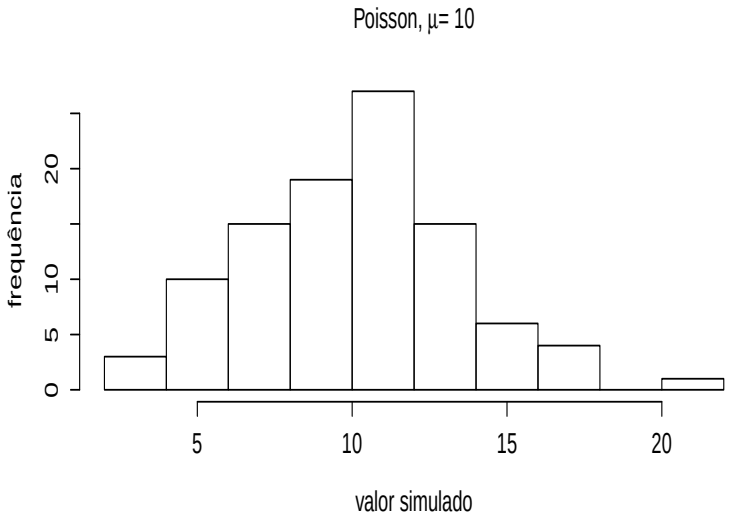


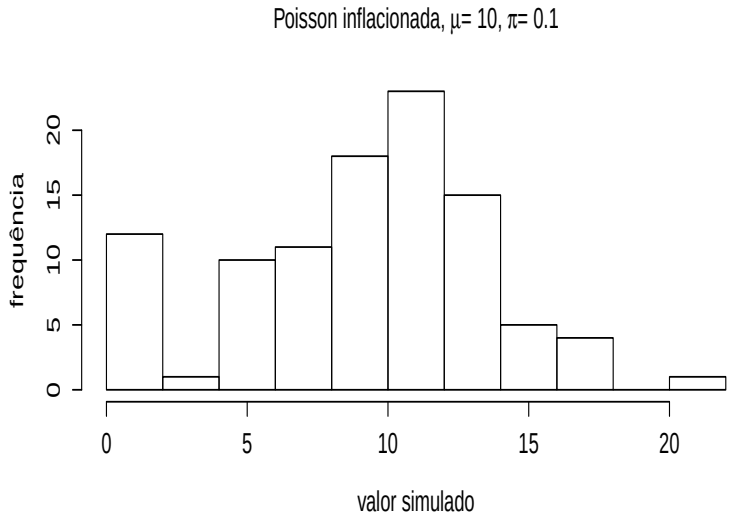


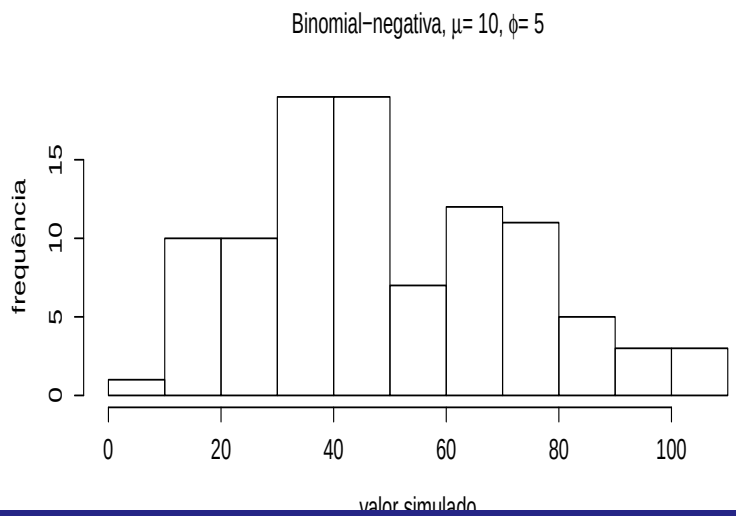


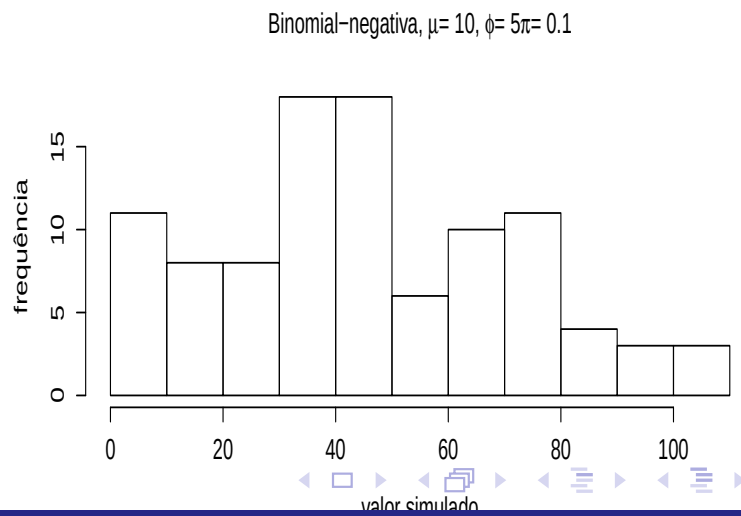


Histograma de valores simulados









Forma alternativa de representação

- Seja $U \sim \text{Bernoulli}(\pi)$, em que:
- $Y|U = 1 \sim X_1$, em que $P(X_1 = 0) = 1$, ou seja $P(Y = 0|U = 1) = 1$.
- e $Y|U = 0 \sim X_2$, em que X_2 segue a distribuição à contagem original (Poisson, geométrica, binomial-negativa). Assim

$$P(Y = y, U = u) = \pi^u \left[(1 - \pi) f_Z(y)\right]^{1-u} \tag{1}$$

- Pode-se provar que $P(Y = y) = \sum_{u=0}^{1} P(Y = y, U = u) = [\pi + (1 - \pi) f(0)] \mathbb{1}_{\{0\}}(y) + (1 - \pi) f(y) \mathbb{1}_{\{1,2,\ldots\}}$ (exercício).

Modelo de regressão para contagens inflacionadas de zeros

- Consideramos $Y_1, ..., Y_n \overset{ind.}{\sim} IZ\{\pi_i, f_{Z_i}(y_i)\}$.
- Podemos modelar tanto os parâmetros relativos à distribuição $f_Z(.)$ quanto o parâmetro π.
- Por exemplo, se $Z \sim \text{Poisson}(\mu_i)$ então podemos considerar $\ln(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ e $\text{logito}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{w}_i'\boldsymbol{\gamma}$.

Inferência via MV

- Estruturas similares podem ser utilizadas considerando-se outras funções de ligação e/ou outras distribuições para a contagem original.
- Se houver mais parâmetros na distribuição de Z (caso da distribuição binomial negativa), eles também podem ser modelados através de uma estrutura de regressão.
- Estimação paramétrica:
  - Otimização direta da log-verossimilhança.
  - Algoritmo EM, otimizando-se a log-verossimilhança aumentada definida pela distribuição conjunta de (Y, U).

log-verossimilhança

- Sejam $\boldsymbol{\theta}$ os parâmetros da contagem original (Z) e $\boldsymbol{\gamma}$ os parâmetros associados à probabilidade de inflacionamento (ou seja, $\pi = h(\boldsymbol{\gamma})$).
- Seja ainda v_i, que assume o valor 1 se $y_i = 0$ e 0 caso contrário. Então

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \left\{ [\pi_i + (1 - \pi_i) f_{Z_i}(0; \boldsymbol{\theta})]^{v_i} [(1 - \pi_i) f_{Z_i}(y_i; \boldsymbol{\theta})]^{1 - v_i} \right\}$$

$$\begin{aligned} l(\boldsymbol{\theta}, \boldsymbol{\gamma}) &= \sum_{i=1}^{n} \{ v_i \ln [\pi_i + (1 - \pi_i) f_{Z_i}(0; \boldsymbol{\theta})] \\ &+ (1 - v_i) [\ln(1 - \pi_i) + \ln f_{Z_i}(y_i; \boldsymbol{\theta})] \} \end{aligned} \quad (2)$$

log-verossimilhança

- Exemplo: Poisson com ligação log sem modelar π

$$\begin{aligned} l(\boldsymbol{\theta}, \boldsymbol{\gamma}) &= \sum_{i=1}^{n} \Big\{ v_i \ln \left[\pi + (1-\pi) e^{-\boldsymbol{x}_i' \boldsymbol{\beta}} \right] \\ &+ (1-v_i) \left[\ln(1-\pi) - e^{\boldsymbol{x}_i' \boldsymbol{\beta}} + y_i \boldsymbol{x}_i' \boldsymbol{\beta} - \ln(y_i!) \right] \Big\} \end{aligned}$$

- Exemplo: Poisson com ligação log modelando π (logito)

$$\begin{aligned} l(\boldsymbol{\theta}, \boldsymbol{\gamma}) &= \sum_{i=1}^{n} \Big\{ v_i \ln \left[\frac{e^{\boldsymbol{w}_i' \boldsymbol{\gamma}} + e^{-\boldsymbol{x}_i' \boldsymbol{\beta}}}{1 + e^{\boldsymbol{w}_i' \boldsymbol{\gamma}}} \right] \\ &+ (1-v_i) \left[-\ln \left(1 + e^{\boldsymbol{w}_i' \boldsymbol{\gamma}} \right) - e^{\boldsymbol{x}_i' \boldsymbol{\beta}} + y_i \boldsymbol{x}_i' \boldsymbol{\beta} - \ln(y_i!) \right] \Big\} \end{aligned}$$

log-verossimilhança (completada/aumentada)

- Consiste em utilizar as observações (y_i) e as variáveis lantes (u_i), ou seja, a distribuição conjunta de (y_i, u_i), i=1,2,...,n.
- De (1) temos que a verossimilhança aumentada é dada por

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \left\{ {\pi_i}^{u_i} \left[f_{Z_i}(y_i; \boldsymbol{\theta})(1 - \pi_i) \right]^{1-u_i} \right\}$$

- Log-verossimilhança aumentada

$$l(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \{ u_i \ln \pi_i + (1 - u_i) \left[\ln(1 - \pi_i) + \ln f_{Z_i}(y_i; \boldsymbol{\theta}) \right] \} \qquad (3)$$

- Note que a expressão (3) é mais tratável do que a expressão (2). Contudo, não observamos as variáveis u_i. Neste caso, devemos usar, por exemplo, o algoritmo EM para obtermos as estimativas de máxima verossimilhança.
- A obtenção das emv através da maximização direta da logverossimilhança (3) é computacionalmente mais complicada. Podemos utilizar a informação de Fisher (observada ou esperada) para calcularmos os erros-padrão
- A obtenção das emv através via algoritmo EM é mais simples (2), mas requer o cálculo da esperança condicional $\mathcal{E}(U_i|y_i, \boldsymbol{\theta}, \boldsymbol{\gamma})$. Além disso, para calcularmos os erros-padrão devemos utilizar a abordagem de Louis (1982) ou a de Meilijson (1989).

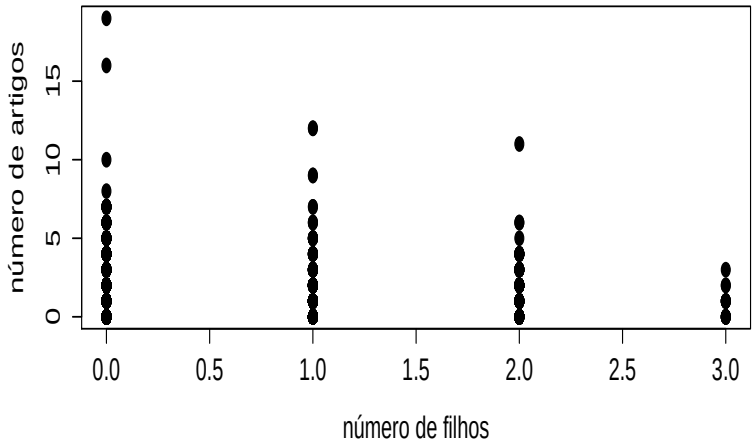
Análise residual

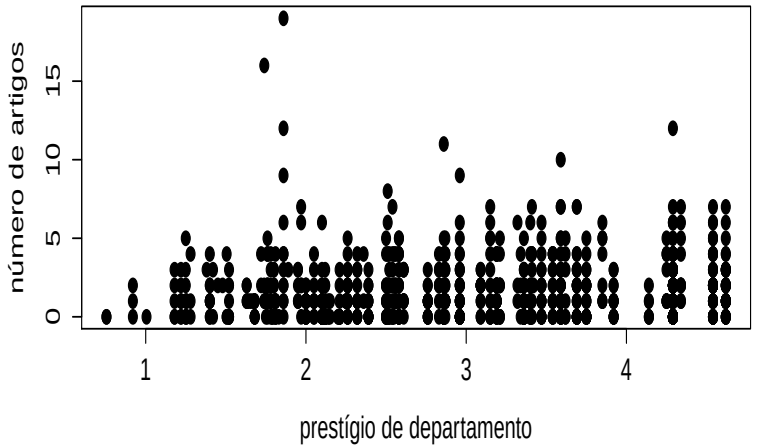
- Resíduo padronizado: $R_i = \frac{Y_i - \widehat{\mu}_i}{\widehat{\mathcal{V}(Y_i)}}$.
- Construir os cinco gráficos usuais: resíduo x índice, resíduo x valores preditos, histograma, valores preditos x valores observados e gráfico de envelopes.
- Mesmo sob o bom ajuste do modelo não, necessariamente, espera-se que os resíduos apresentem distribuição normal (mesmo que aproximadamente).
- Voltaremos agora ao Exemplo 16.

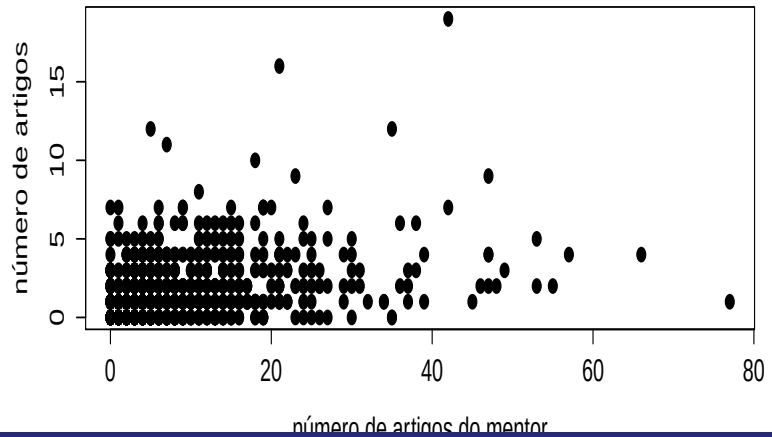
Medidas resumo

Gênero	Estado civil	Média	DP	Var.	CV(%)	Min.	Max.	n
masculino	solteiro	1,95	2,01	4,05	103,38	0	7	113
	casado	1,86	2,23	4,97	119,58	0	19	381
feminino	solteiro	1,39	1,51	2,28	108,80	0	7	196
	casado	1,54	1,59	2,52	102,88	0	10	225

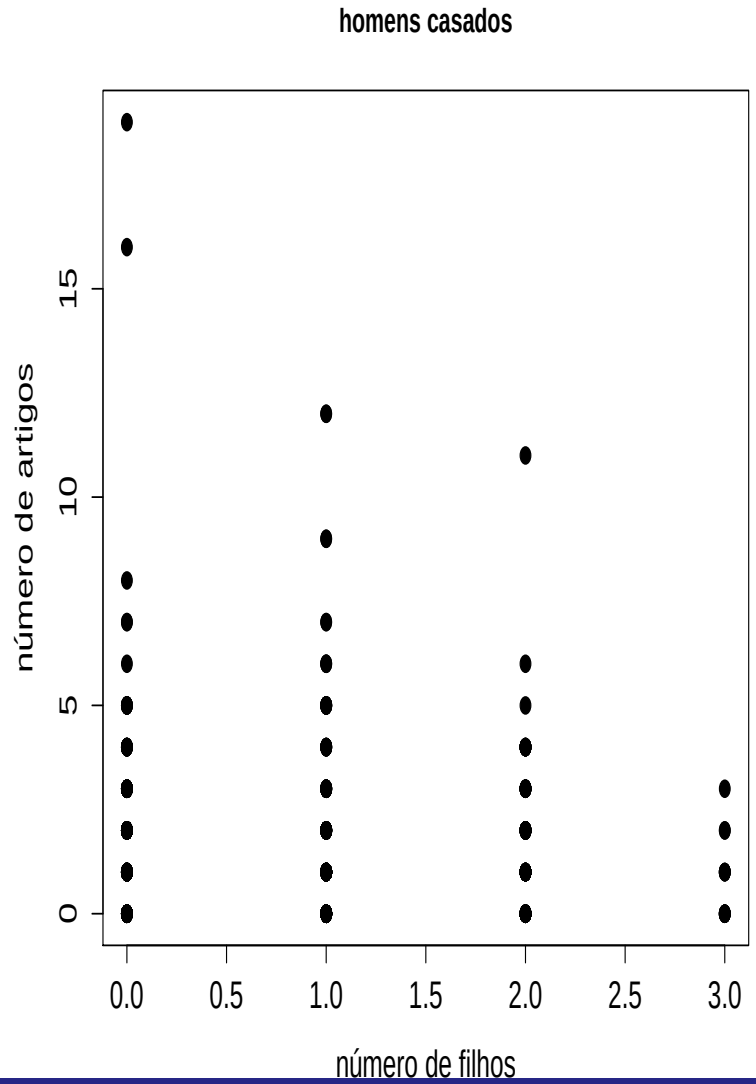
Gráficos de dispersão

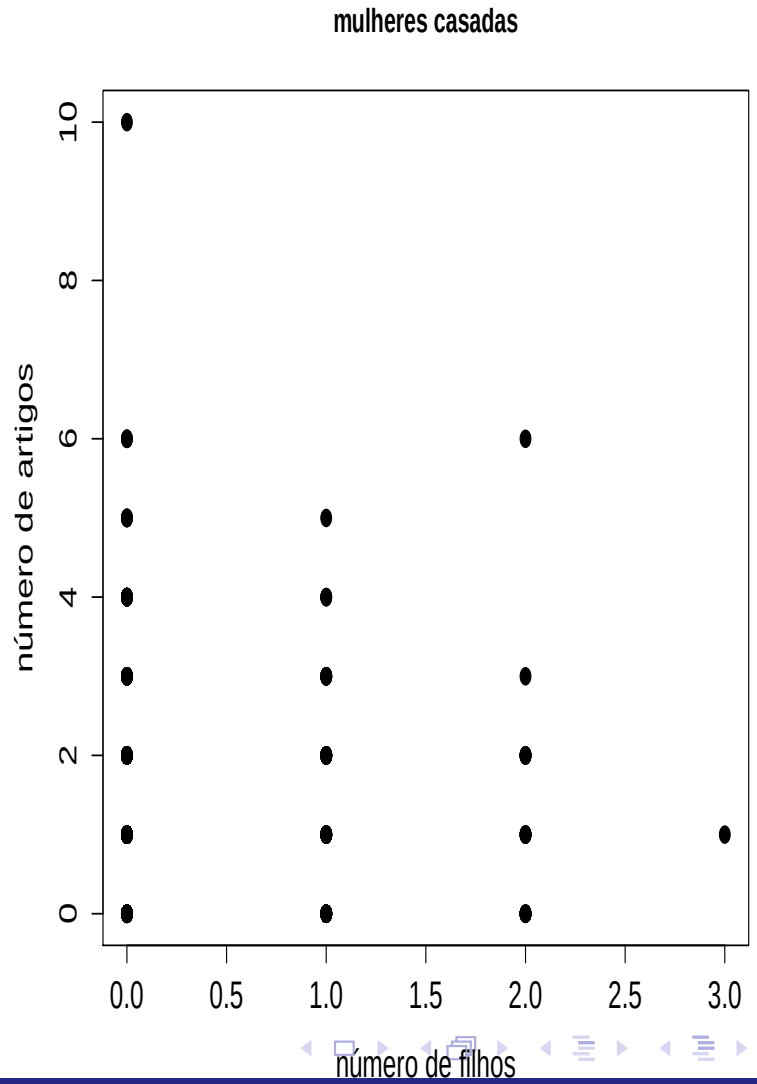




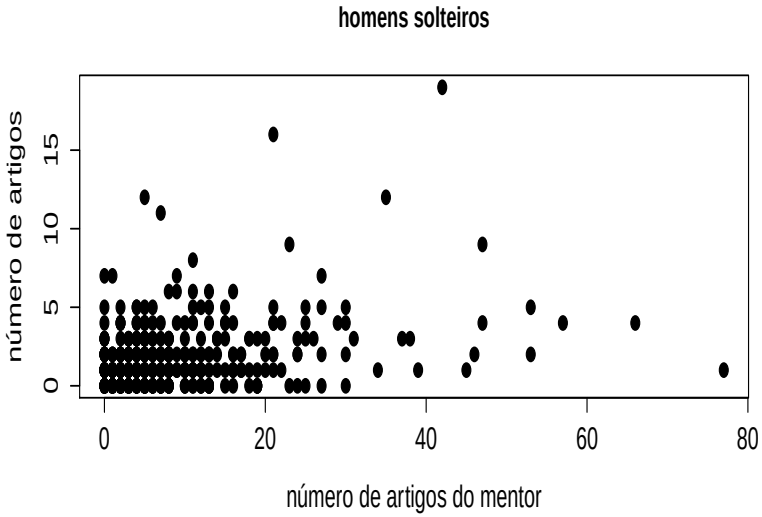


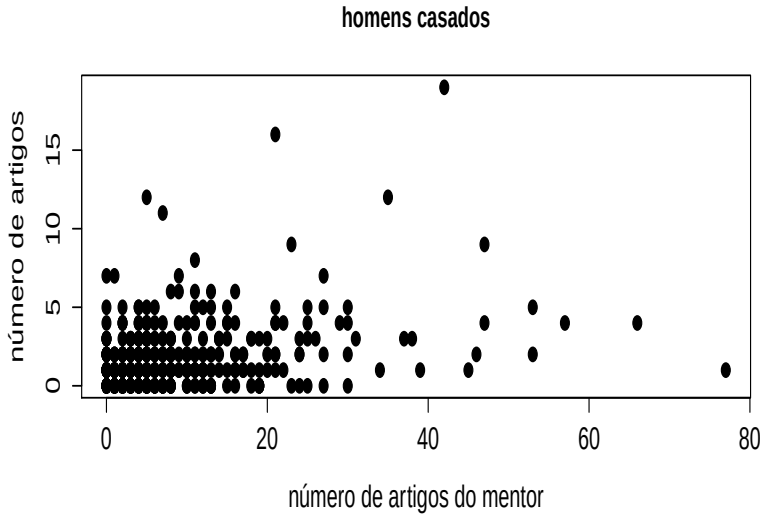
Gráficos de dispersão (número de filhos)

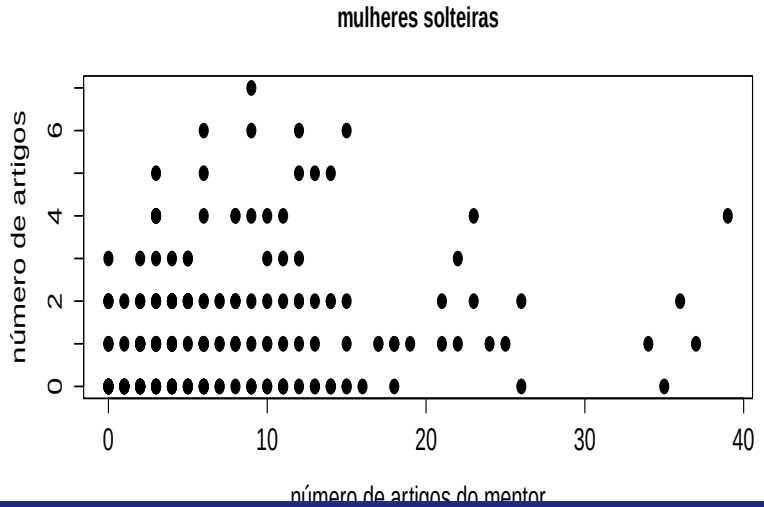


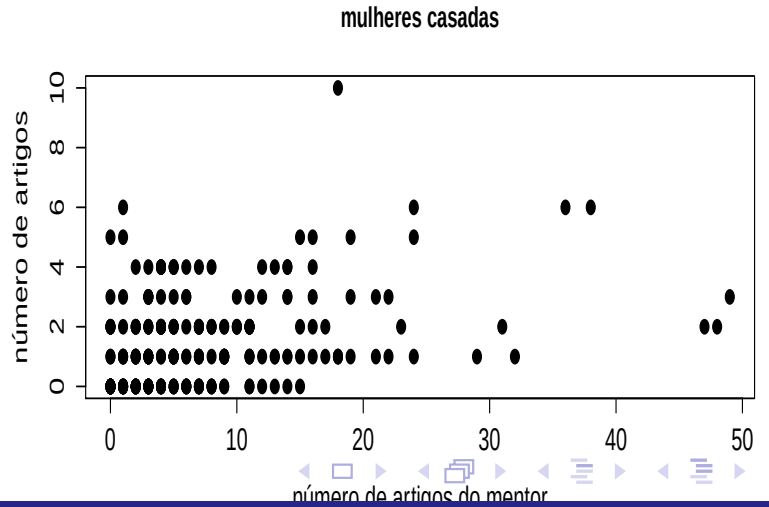


Gráficos de dispersão (número de artigos do orientador)

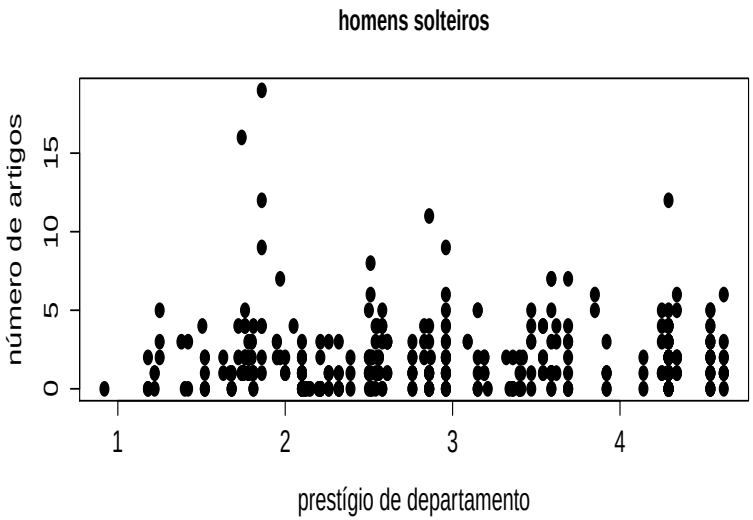


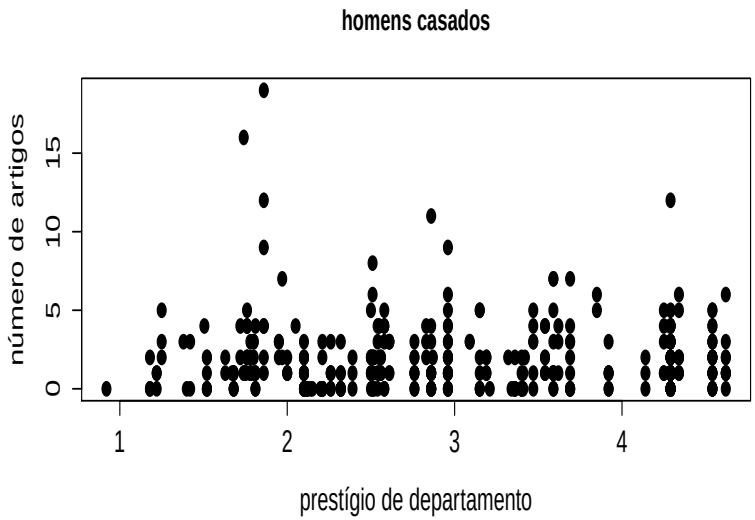


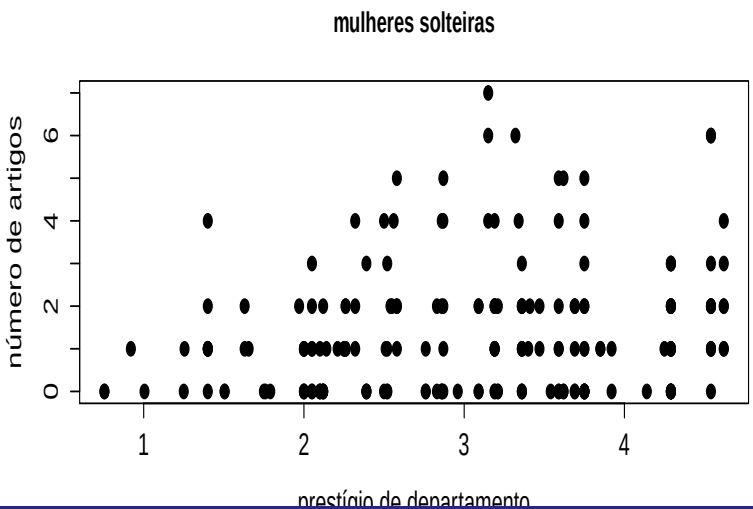


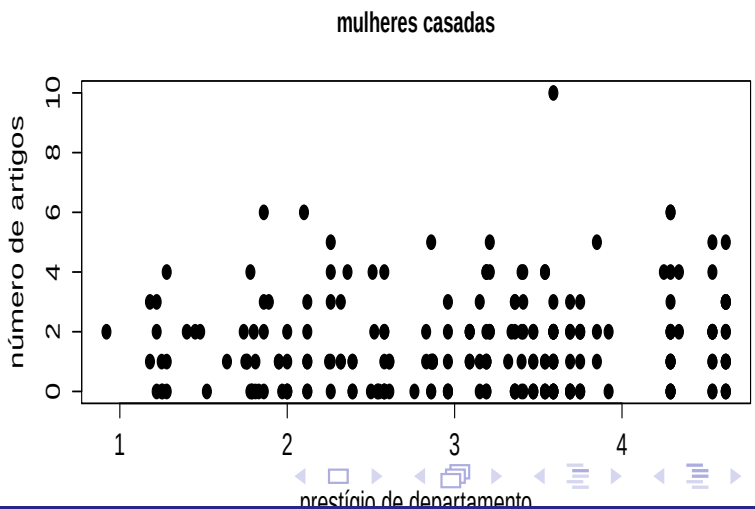


Gráficos de dispersão (prestígio do departamento)

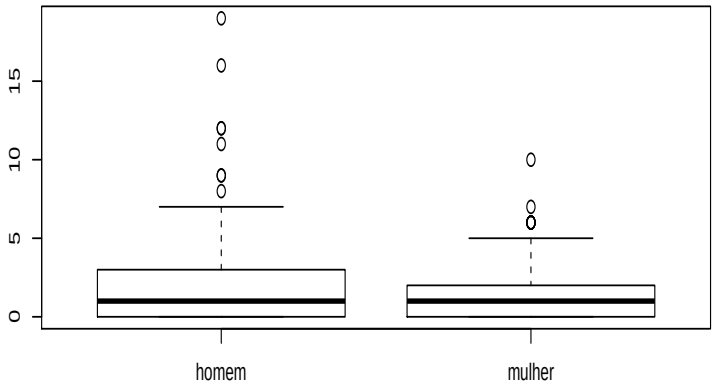


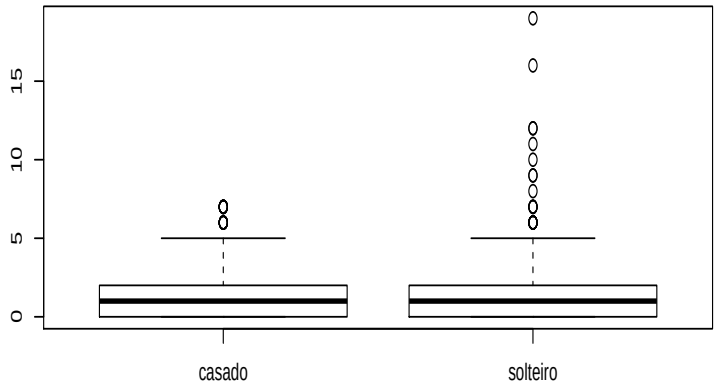


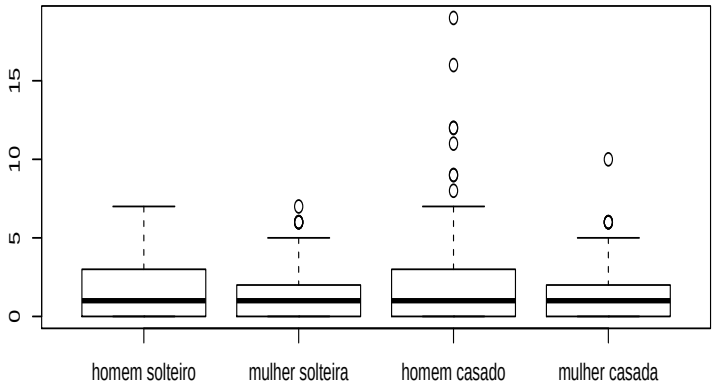




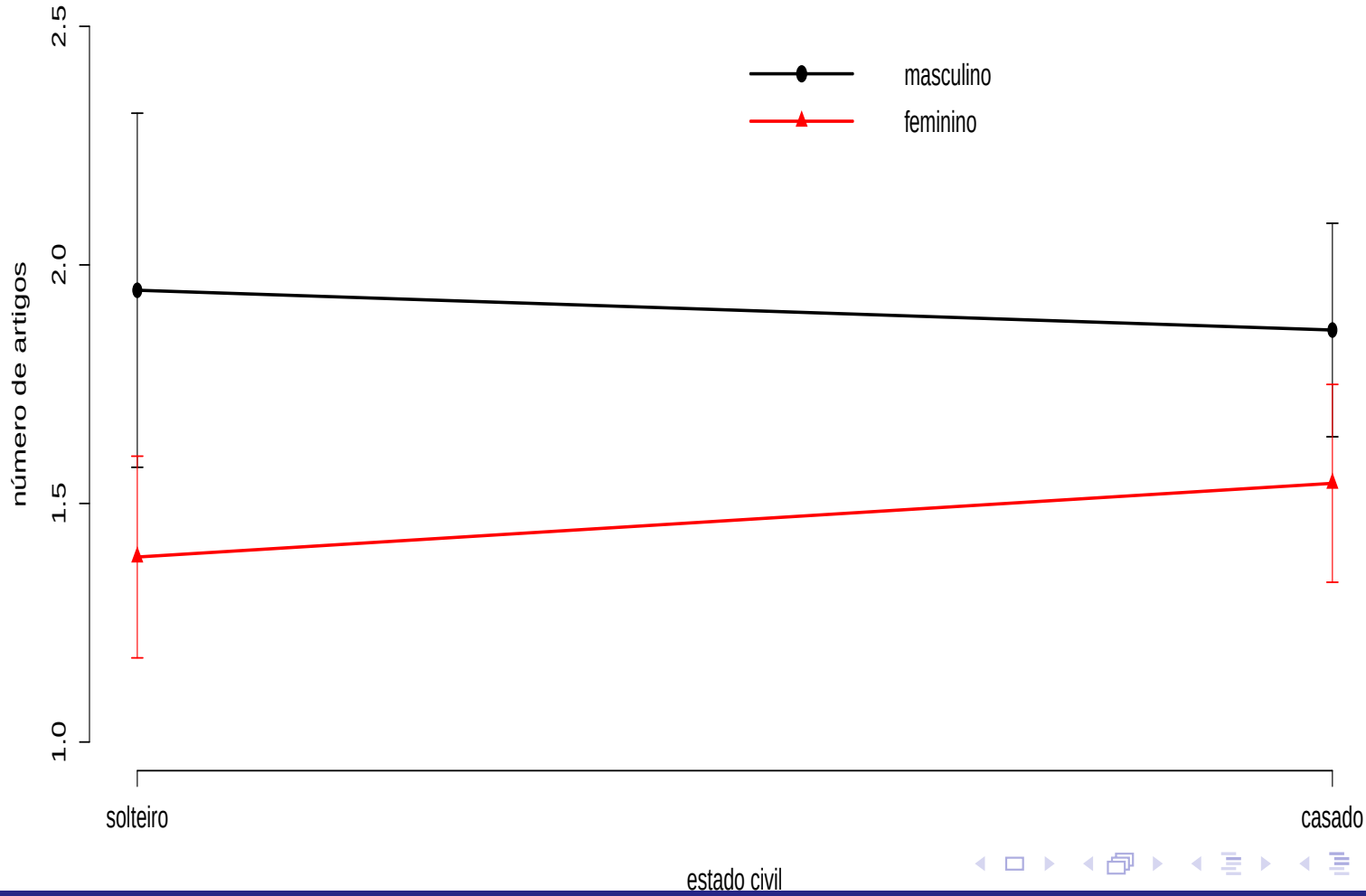
Box-plots


0
5
10
15
homem
mulher


0
5
10
15
casado
solteiro


0
5
10
15
homem solteiro
mulher solteira
homem casado
mulher casada

Gréficos de perfis



Voltando ao Exemplo 16

Modelos

$Y_{ijk} \overset{ind.}{\sim} \text{Poisson}(\mu_{ijk})(1)$; $Y_{ijk} \overset{ind.}{\sim} \text{BN}(\mu_{ijk}, \phi)(2)$;

$Y_{ijk} \overset{ind.}{\sim} IZ\{\pi, \text{Poisson}(\mu_{ijk})\}(3)$; $Y_{ijk} \overset{ind.}{\sim} IN\{\pi, \text{BN}(\mu_{ijk}, \phi)\}(4)$

gênero (i = 1 (masculino),2 (feminino)), estado civil (j=1 (solteiro), 2 (casado)), $k = 1, 2, ..., n_{ij}$ (aluno)

- Em todos os modelos

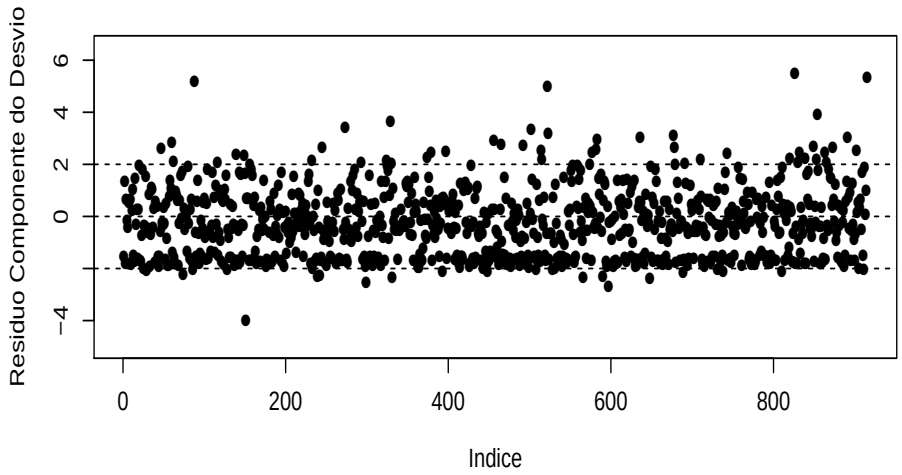
  $$\ln(\mu_{ijk}) = \alpha + \beta_i + \gamma_j + \delta_1(x_{1ijk} - \overline{x}_1) + \delta_2(x_{2ijk} - \overline{x}_2) + \delta_3(x_{3ijk} - \overline{x}_3),$$

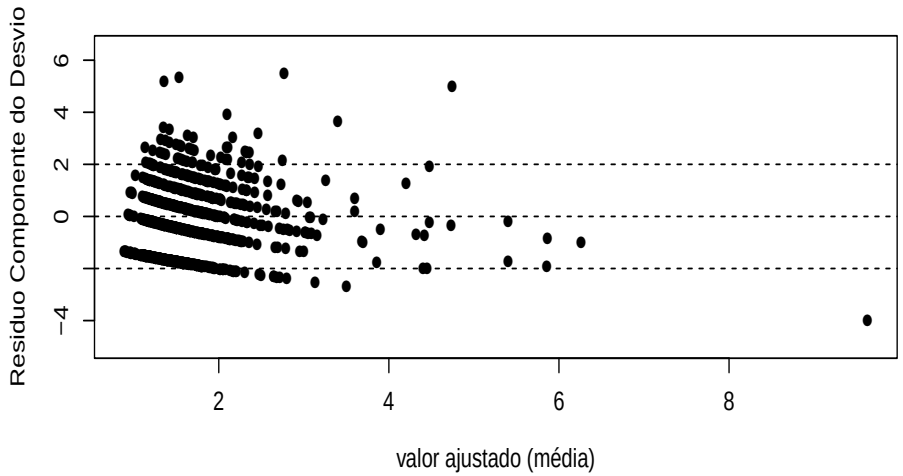
  em que $x_r = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} x_{rijk}$, $r = 1, 2, 3$ e $\beta_1 = \gamma_1 = 0$.

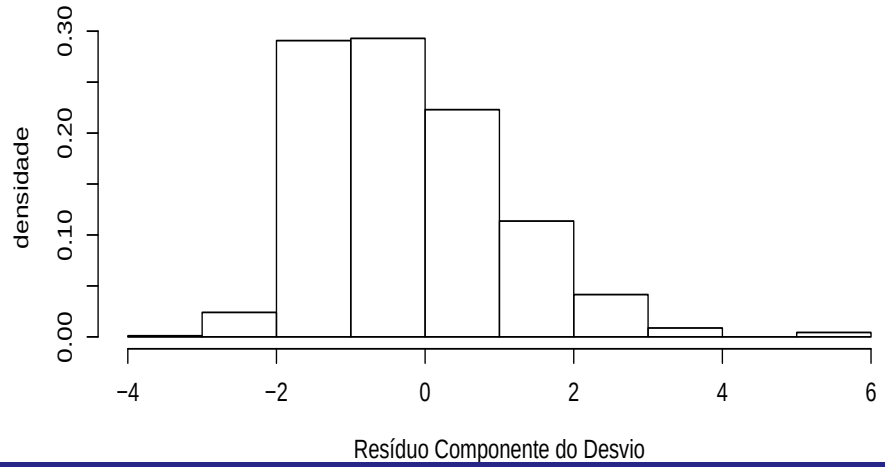
- Nos modelos (3) e (4) $\ln\left(\frac{\pi}{1-\pi}\right) = \theta$.

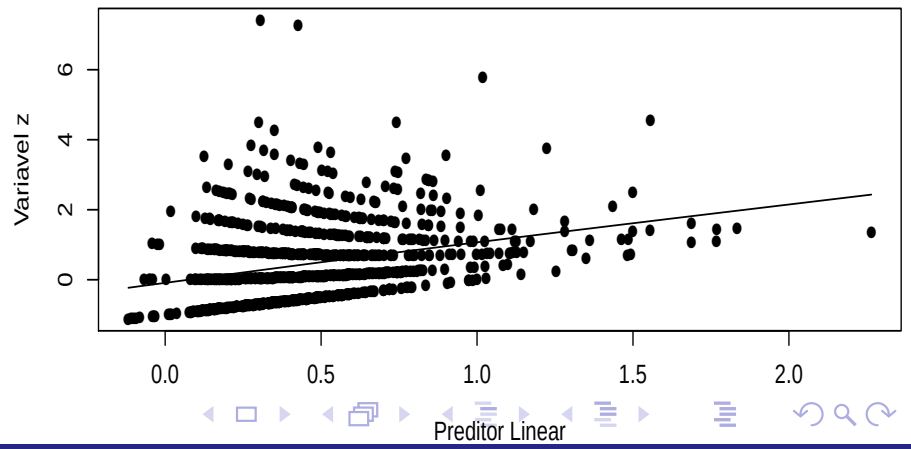
- Exercício: interpretar os parâmetros adequadamente.

Gráficos de diagnóstico (M1)

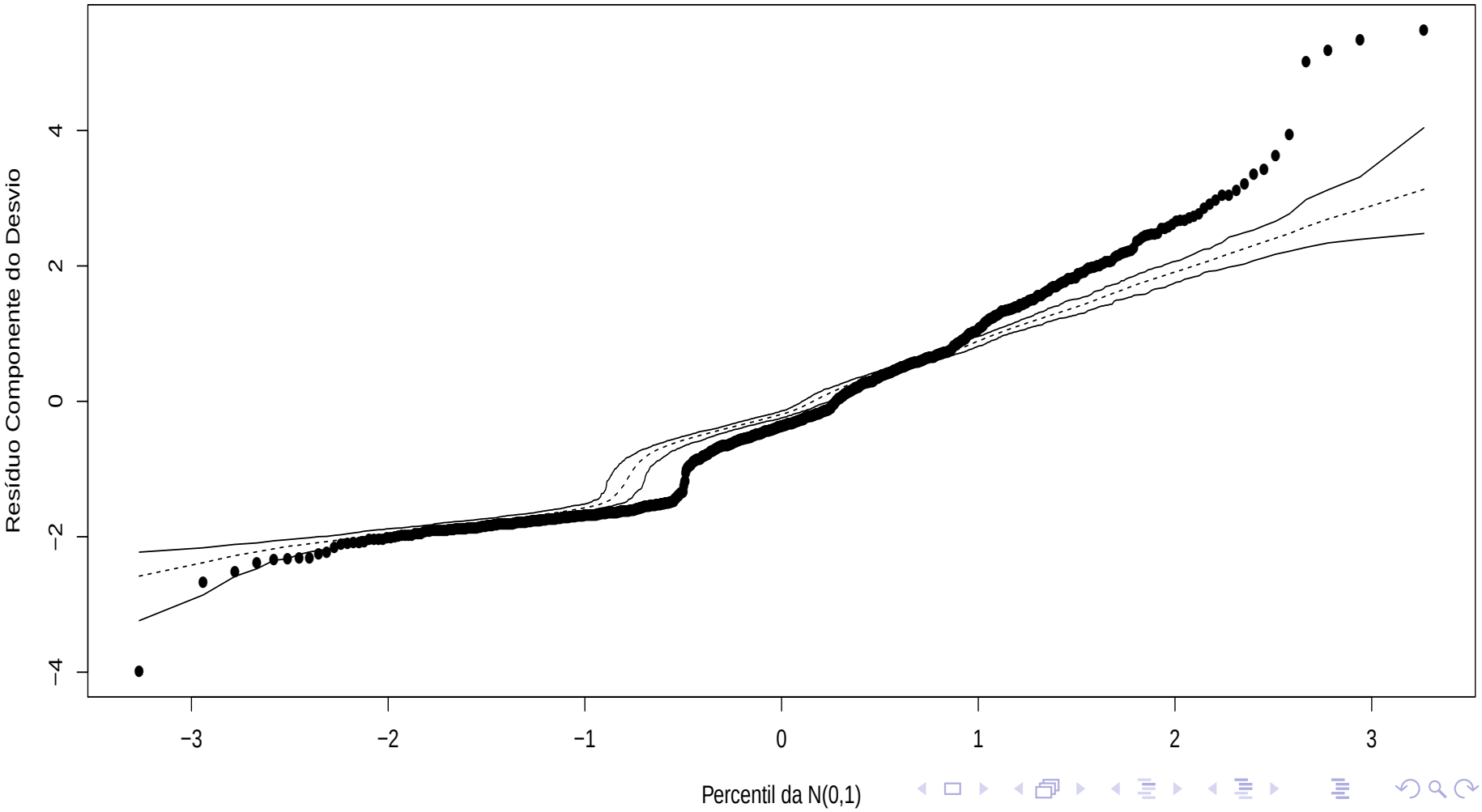




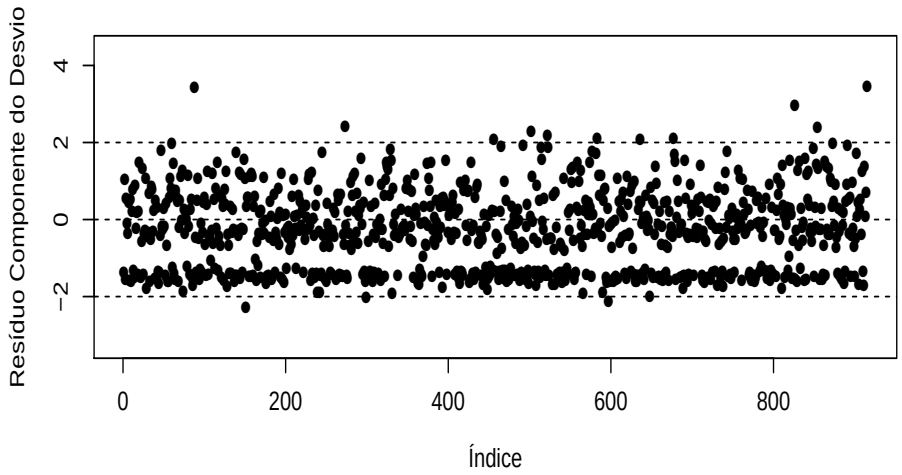


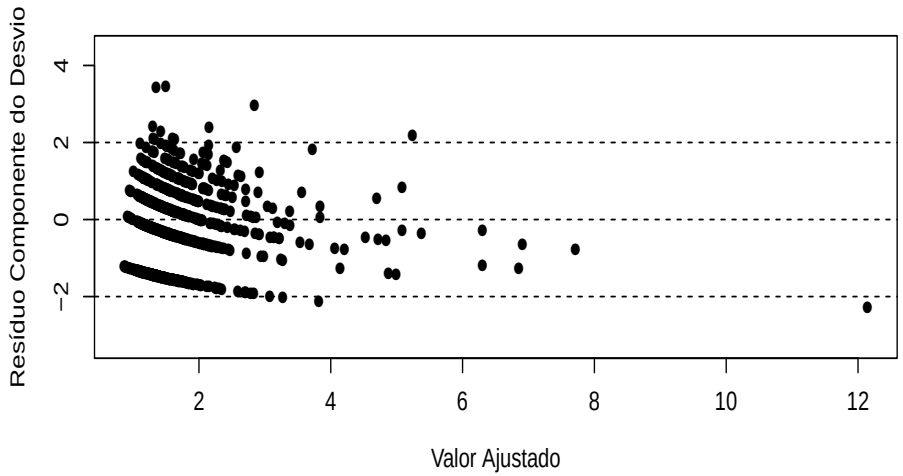


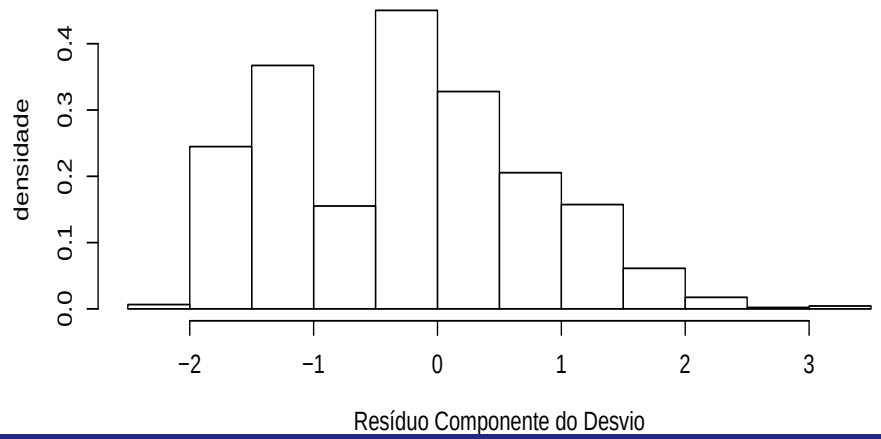
Gráficos de envelopes (M1)

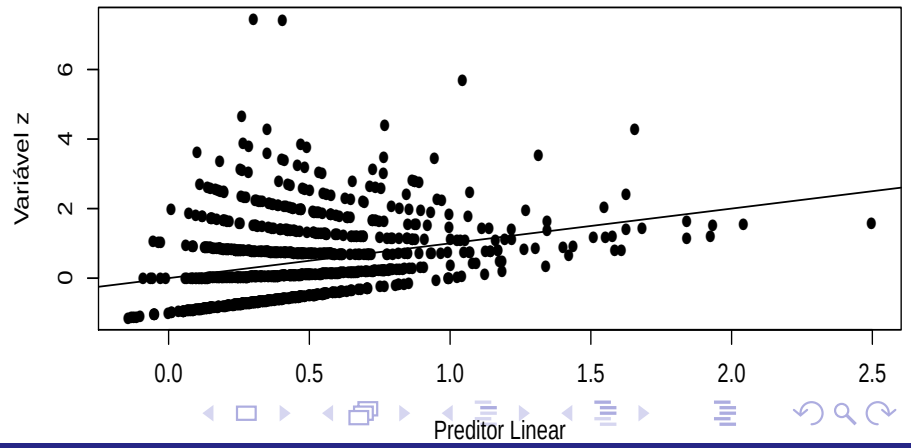


Gráficos de diagnóstico (M2)

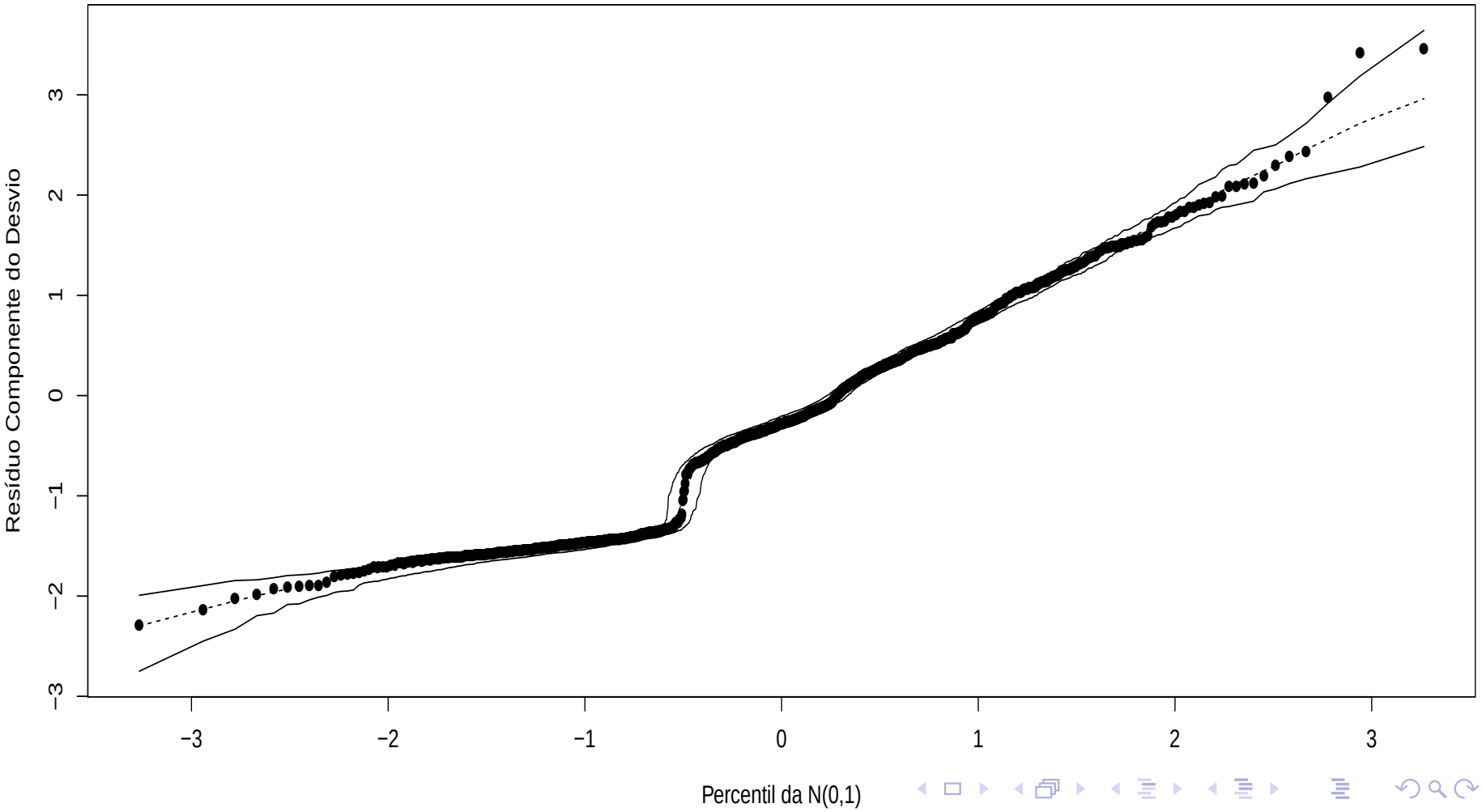




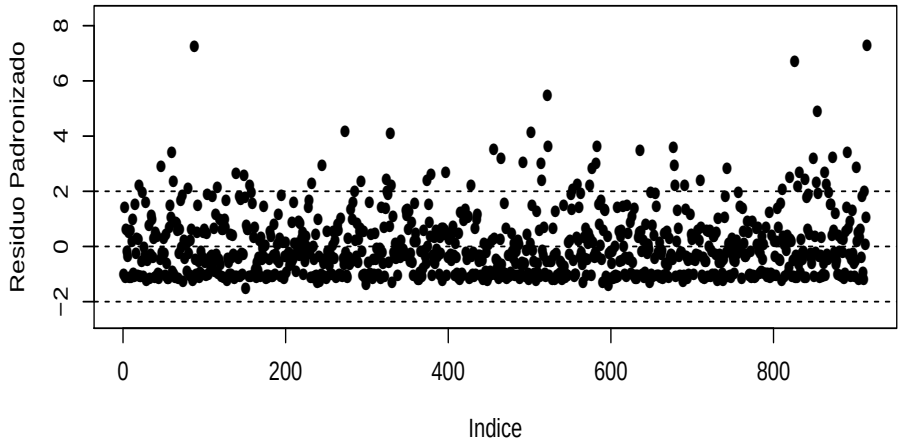


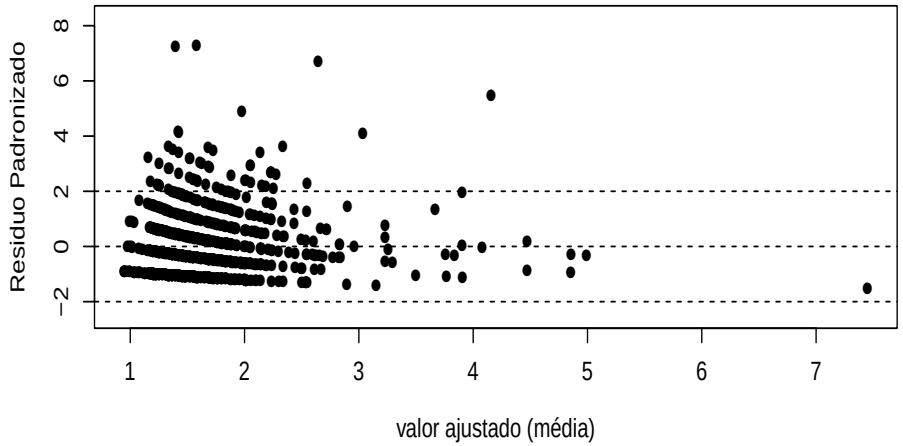


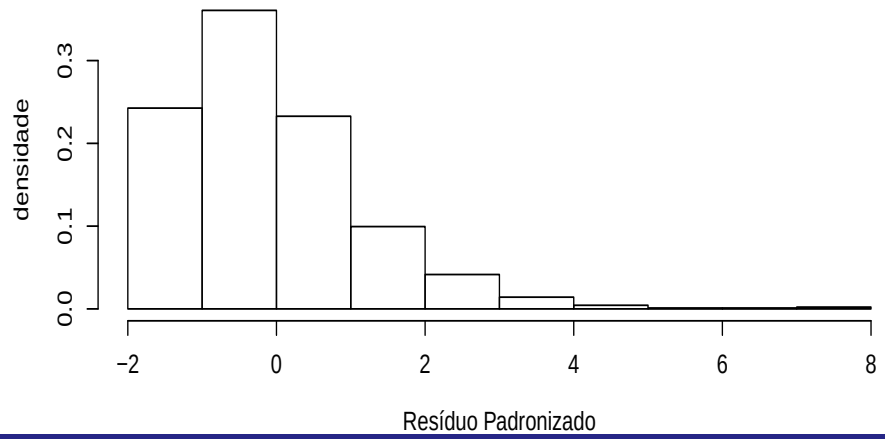
Gráficos de envelopes (M2)

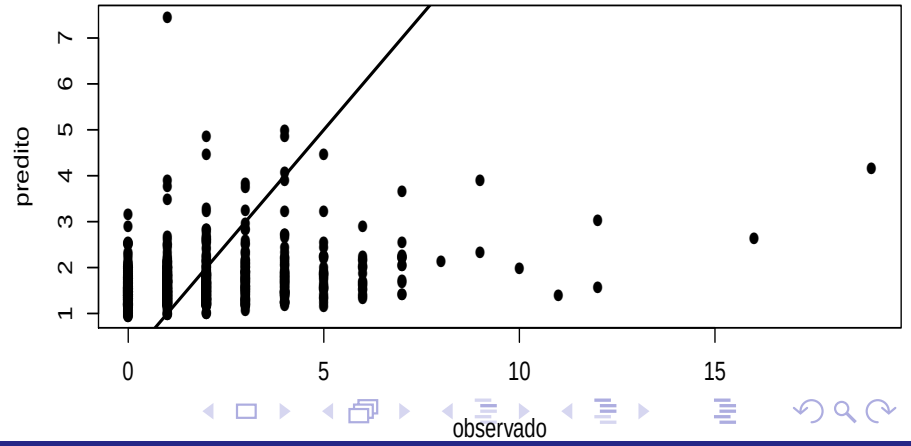


Gráficos de diagnóstico (M3)

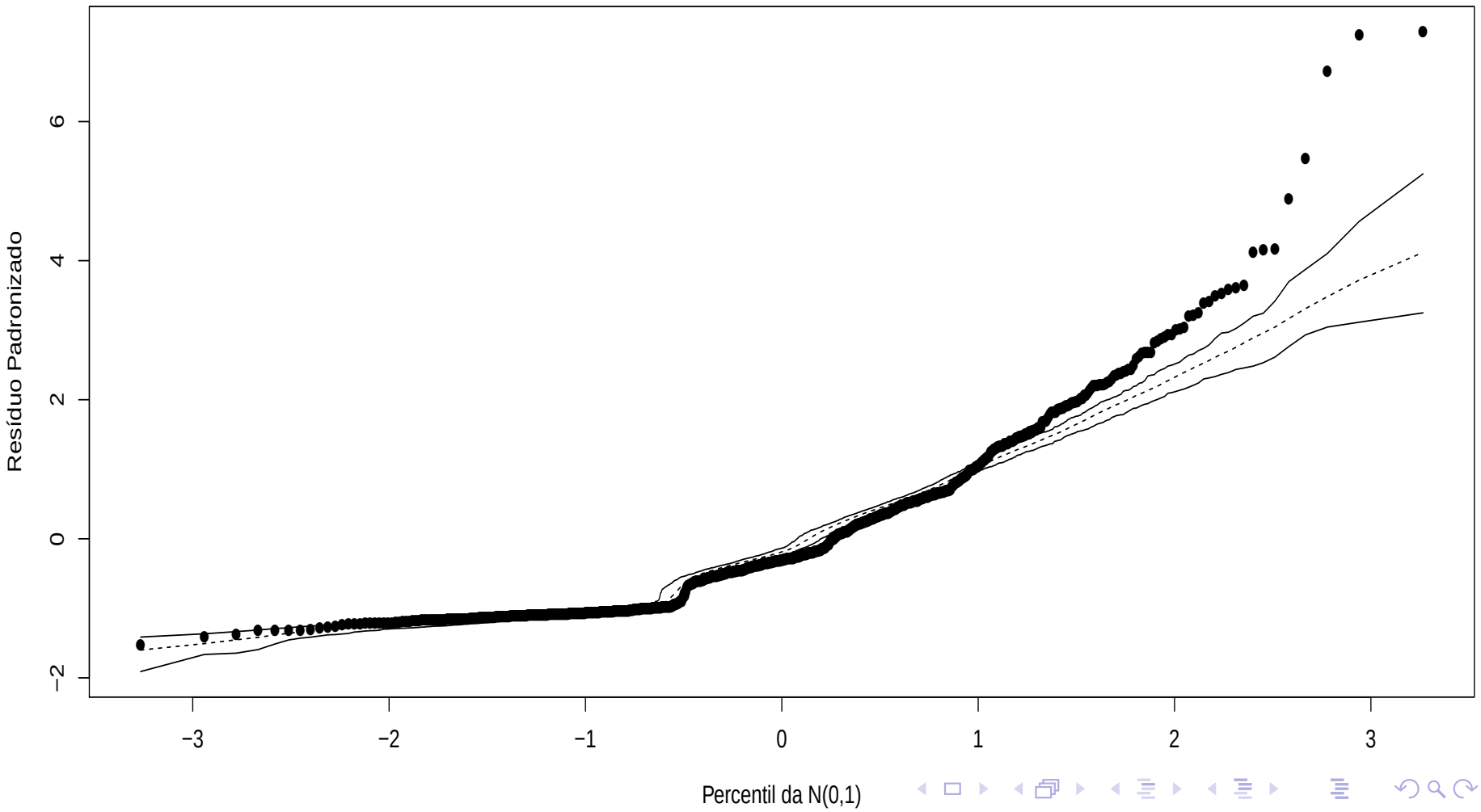




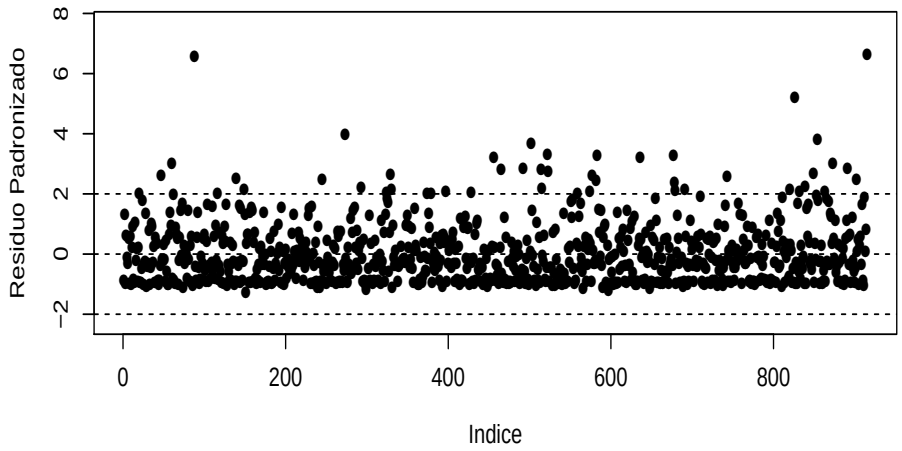


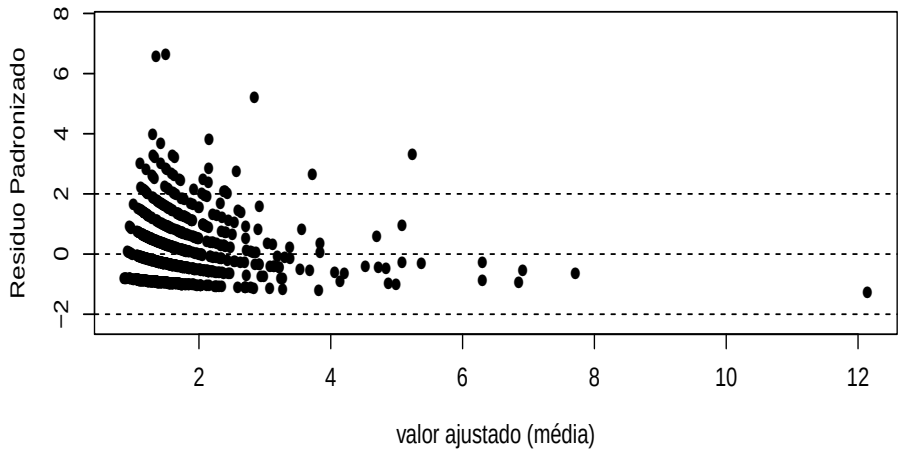


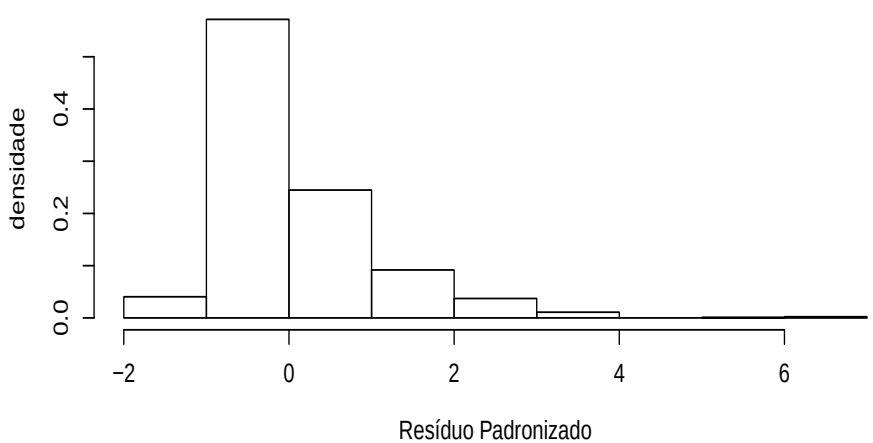
Gráficos de envelopes (M3)

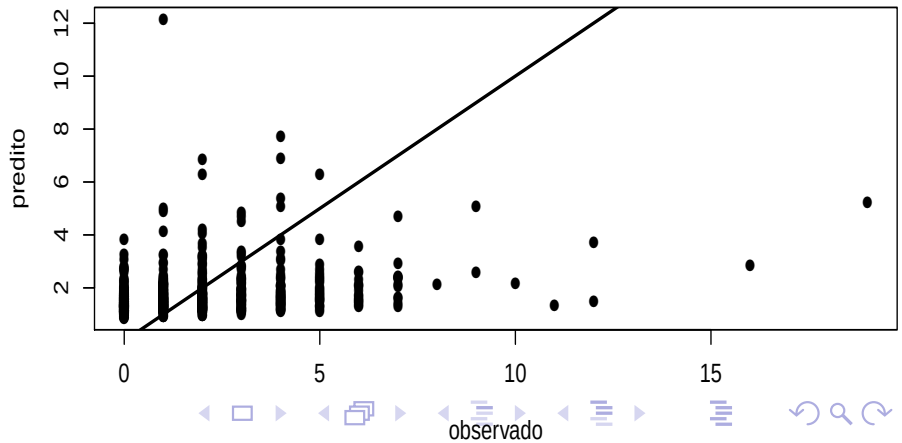


Gráficos de diagnóstico (M4)

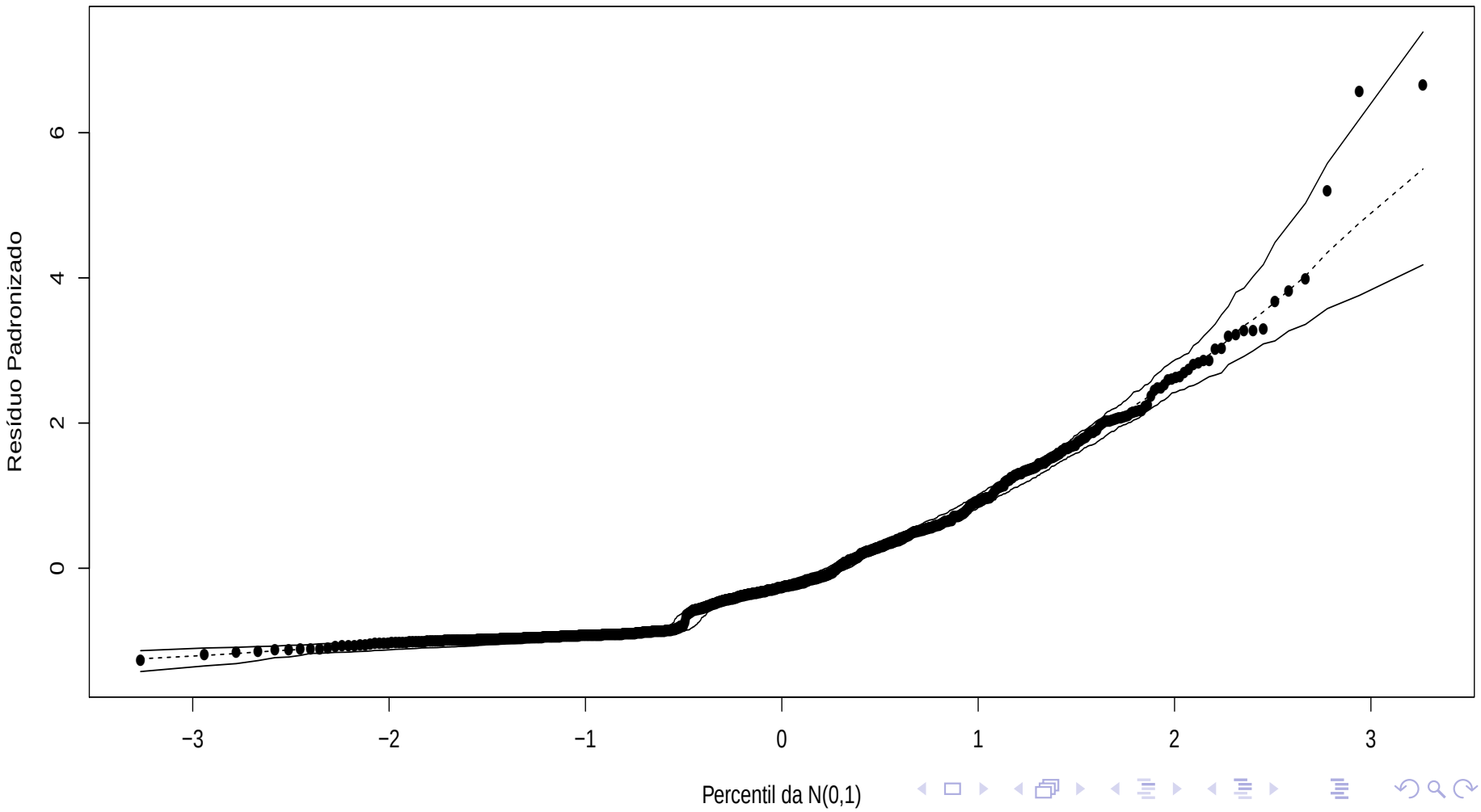








Gráficos de envelopes (M4)



Estatísticas de comparação de modelos

Modelo	AIC	BIC	desvio	p-valor (desvio)
1	3314,11	3343,03	1634,37	$< 0{,}0001$
2	3135,92	3169,65	1004,28	0,0148
3	3255,57	3289,30	-	-
4	3137,92	3176,47	-	-

Estimativas (M1)

Parâmetro	Est.	EP	IC(95%)	Estat. Z	p-valor
α	0,48	0,06	[0,37 ; 0,59]	8,36	$<$ 0,0001
β_2	-0,18	0,04	[-0,26 ; -0,11]	-4,61	$<$ 0,0001
γ_2	0,026	0,002	[0,022 ; 0,029]	12,73	$<$ 0,0001
δ_1	0,01	0,03	[-0,04 ; 0,06]	0,49	0,6271
δ_2	-0,22	0,05	[-0,33 ; -0,12]	-4,11	$<$ 0,0001
δ_3	0,16	0,06	[0,03 ; 0,28]	2,53	0,0114

Estimativas (M2)

Parâmetro	Est.	EP	IC(95%)	Estat. Z	p-valor
α	0,47	0,08	[0,32 ; 0,62]	6,22	$<$ 0.0001
β_2	-0,18	0,05	[-0,28 ; -0,07]	-3,34	0.0008
γ_2	0,03	$<$ 0,01	[0,02 ; 0,04]	9,05	$<$ 0.0001
δ_1	0,02	0,04	[-0,06 ; 0,09]	0,43	0,6703
δ_2	-0,22	0,07	[-0,36 ; -0,07]	-2,98	0,0029
δ_3	0,15	0,08	[-0,01 ; 0,31]	1,83	0,0668
ϕ	2,26	0,27	[1,73 ; 2,80]	-	-

Estimativas (M3)

Parâmetro	Est.	EP	IC(95%)	Estat. Z	p-valor
α	0,67	0,06	[0,54 ; 0,79]	10,41	$<$ 0,0001
β_2	-0,17	0,04	[-0,26 ; -0,09]	-3,94	0,0001
γ_2	0,02	$<$ 0,01	[0,02 ; 0,03]	9,97	0,0001
δ_1	0,003	0,029	[-0,053 ; 0,058]	0,09	0,9294
δ_2	-0,23	0,06	[-0,35 ; -0,12]	-3,95	0,0001
δ_3	0,13	0,07	[$<$0,01 ; 0,26]	2,00	0,0460
π	0,16	0,07	[0,02 ; 0,30]	0,00	$<$ 0,0001

Estimativas (M4)

Parâmetro	Est.	EP	IC(95%)	Estat. Z	p-valor
α	0,47	0,08	[0,32 ; 0,62]	6,24	<0,0001
β_2	-0,18	0,05	[-0,28 ; -0,07]	-3,33	0,0009
γ_2	0,03	$<$ 0,01	[0,02 ; 0,04]	8,38	<0,0001
δ_1	0,02	0,04	[-0,06 ; 0,09]	0,42	0,6716
δ_2	-0,22	0,07	[-0,36 ; -0,07]	-2,98	0,0029
δ_3	0,15	0,08	[-0,01 ; 0,31]	1,83	0,0677
ϕ	2,26	0,37	[1,54 ; 2,99]	0,00	<0,0001
π	$4,11 \times 10^{-6}$	35,17	-	-	-

Comentários

- As vezes superdispersão e excesso de zeros produzem comportamentos semelhantes nos dados.
- Isso se reflete, em algumas vezes, na semelhança dos ajustes de modelos que contemplam pelo menos uma dessas característcas.
- Os modelos M2 (superdispersão) e M4 (superdispersão + excesso de zeros) apresentaram bons e equivalentes ajustes, com uma ligeira vantagem par ao modelo M2.
- Neste caso, devido aos resultados (veja também a estimativa de baixa magnitude de π para o modelo M4), considerar a superdispersão (sem excesso de zeros) foi suficiente.