Exemplos MLPAGs

Gilberto A. Paula

Departamento de Estatística
IME-USP, Brasil
giapaula@ime.usp.br

2^{o} Semestre 2022

Sumário

Motivação

Os MLPAGs combinam duas classes conhecidas de modelos de regressão:

- modelos lineares generalizados (Nelder e Wedderburn, 1972; McCullagh e Nelder, 1989)
- modelos aditivos generalizados (Hastie e Tibshirani, 1990)

Definição

Os MLGs são definidos pelos componentes

- $Y_i|\mathbf{x}_i \overset{\text{ind}}{\sim} \text{FE}(\mu_i, \phi)$
- $\mu_i = g^{-1}(\eta_i)$, com $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$,

em que $\mu_i = \text{E}(Y_i)$, $\phi > 0$ é o parâmetro de precisão, η_i denota o preditor linear e $g(\cdot)$ a função de ligação, $i = 1, \ldots, n$.

Definição

Os MAGs são definidos pelos componentes

- $Y_i|\mathbf{t}_i \overset{\text{ind}}{\sim} \text{FE}(\mu_i, \phi)$
- $\mu_i = g^{-1}(\eta_i)$, com $\eta_i = \alpha + \sum_{j=1}^{q} f_j(t_{ij})$,

em que $f_1(t_1), \ldots, f_q(t_q)$ são funções duas vezes diferenciáveis, integráveis na segunda derivada e cujas derivadas são contínuas, enquanto $t_{i1}, \ldots, t_{iq}$ são valores de variáveis explicativas contínuas, para $i = 1, \ldots, n$.

Definição

Os MLPAGs são definidos pelos componentes

- $Y_i|(\mathbf{x}_i, \mathbf{t}_i) \overset{\text{ind}}{\sim} \text{FE}(\mu_i, \phi)$
- $\mu_i = g^{-1}(\eta_i)$, com $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{j=1}^{q} f_j(t_{ij})$,

para $i = 1, \ldots, n$.

Sumário

Definição

Os MLPGs são definidos pelos componentes

- $Y_i|(\mathbf{x}_i, t_i) \overset{\text{ind}}{\sim} \text{FE}(\mu_i, \phi)$
- $\mu_i = g^{-1}(\eta_i)$, com $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + f(t_i)$,

para $i = 1, \ldots, n$.

B-splines

Os beta-splines (B-splines, de Boor, 1978; Wood, 2017, Seção 5.3.3) formam uma classe flexível de splines que podem ser expressos conforme abaixo

$$f(t,k) = \sum_{j=1}^{q} N_{j,k}(t)\gamma_j,$$

em que k é o grau do spline (por exemplo, $k = 2$ spline cúbico), q a dimensão da base e $m = q + k + 2$ o número de nós equidistantes. As funções da base $N_{j,k}(t)$ são expressas de forma recursiva.

P-splines

P-splines é um método eficiente de suavização que utiliza os B-splines com uma penalização discreta, definida abaixo

$$\int_a^b \{f^{(d)}(t)\}^2 dt \cong \gamma^\top \mathbf{D}_d^\top \mathbf{D}_d \gamma,$$

em que $\mathbf{D}_d$ denota uma matriz de diferenças de ordem d nos coeficientes $\gamma_1, \ldots, \gamma_r$ do B-spline.

Problema de Otimização

Em geral $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{N}\boldsymbol{\gamma}$ e tem-se o seguinte problema de maximização:

$$\mathrm{L}_p(\boldsymbol{\theta}, \lambda) = \mathrm{L}(\boldsymbol{\theta}) - \frac{\lambda}{2}\boldsymbol{\gamma}^\top \mathbf{K}\boldsymbol{\gamma},$$

em que $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \phi)^\top$ e $\mathbf{K}$ é uma matriz positiva semidefinida que depende do spline utilizado.

Sumário

Início da Menstruação

Descrição dos Dados

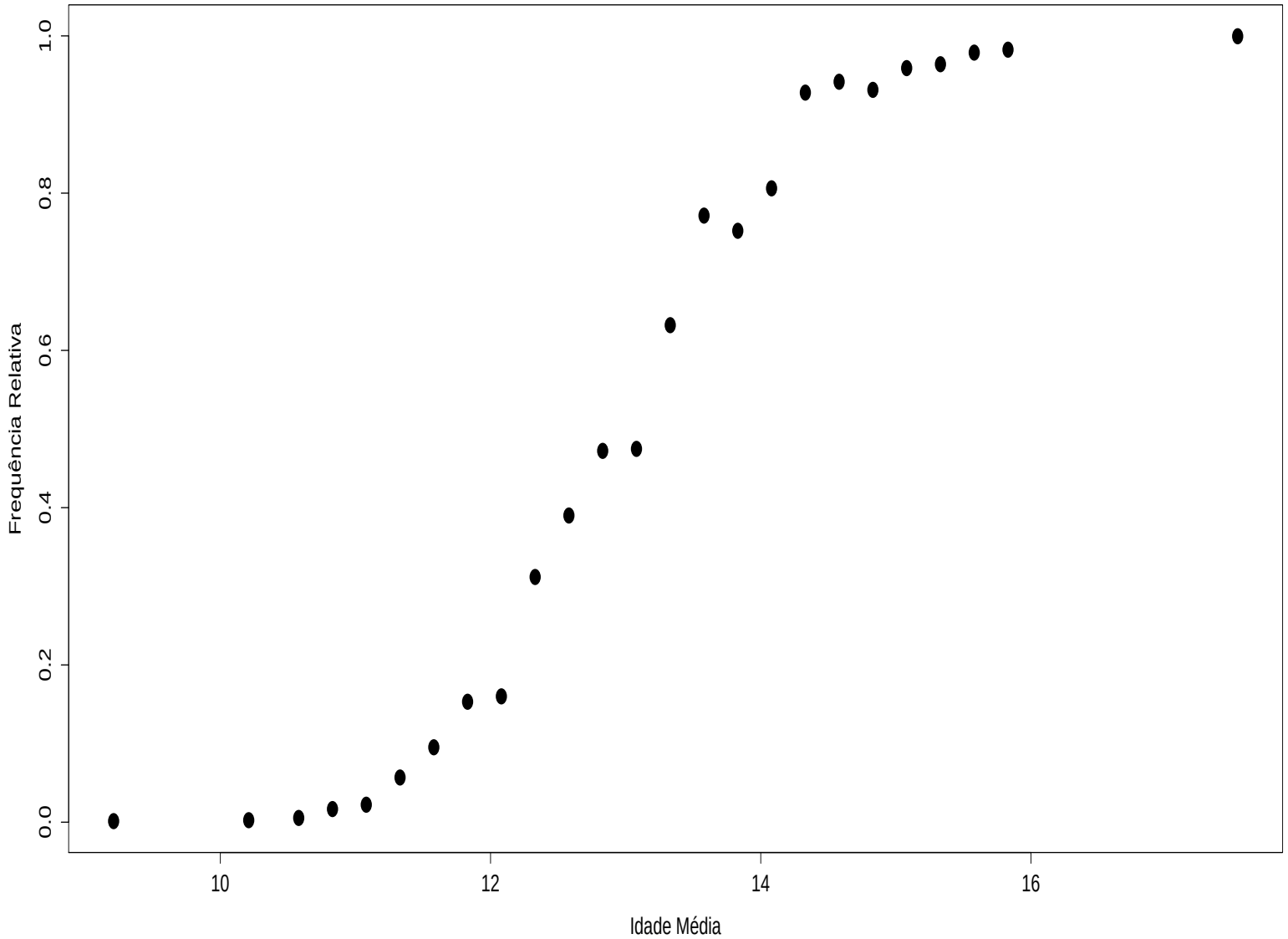
Milecer e Szczotka (1966) investigaram a idade do início da menstruação em 3918 garotas de Varsóvia. Para 25 médias de idade foram observadas:

- total de adolescentes
- total de adolescentes que afirmaram estarem menstruando.

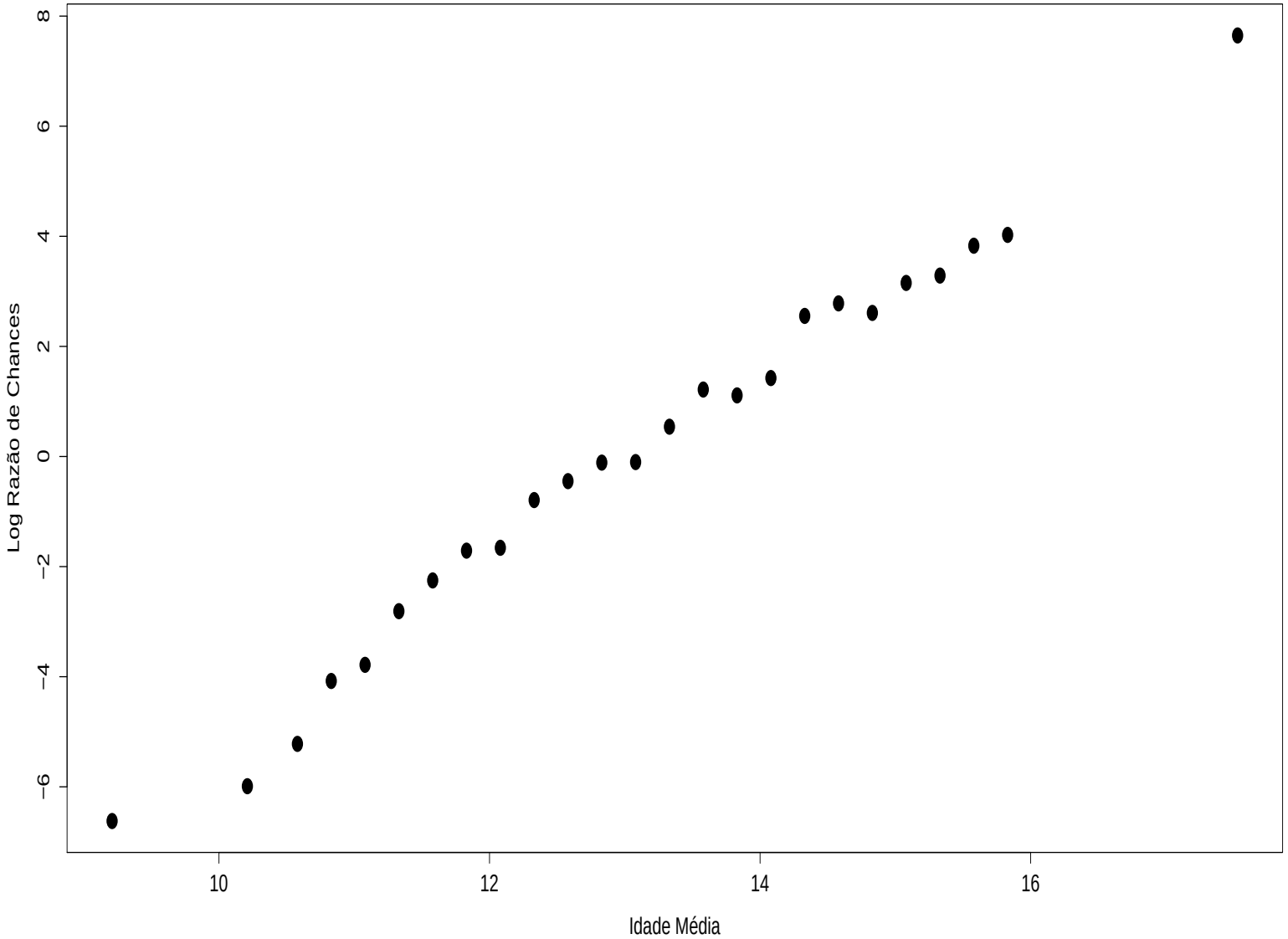
Descrição dos Dados

IdadeM	Mens	Total	IdadeM	Mens	Total
9,21	0	376	13,08	47	99
10,21	0	200	13,33	67	106
10,58	0	93	13,58	81	105
10,83	2	120	13,83	88	117
11,08	2	90	14,08	79	98
11,33	5	88	14,33	90	97
11,58	10	105	14,58	113	120
11,83	17	111	14,83	95	102
12,08	16	100	15,08	117	122
12,33	29	93	15,33	107	111
12,58	39	100	15,58	92	94
12,83	51	108	15,83	112	114
			17,53	1049	1049

Frequência Relativa versus Idade Média



Log Razão de Chances versus Idade Média



MAG Binomial

Seja y_i o número de adolescentes que afirmaram estarem mestruando dentre as n_i adolescentes entrevistadas no grupo com idade média x_i, para $i = 1, \ldots, 25$. Supor o seguinte modelo:

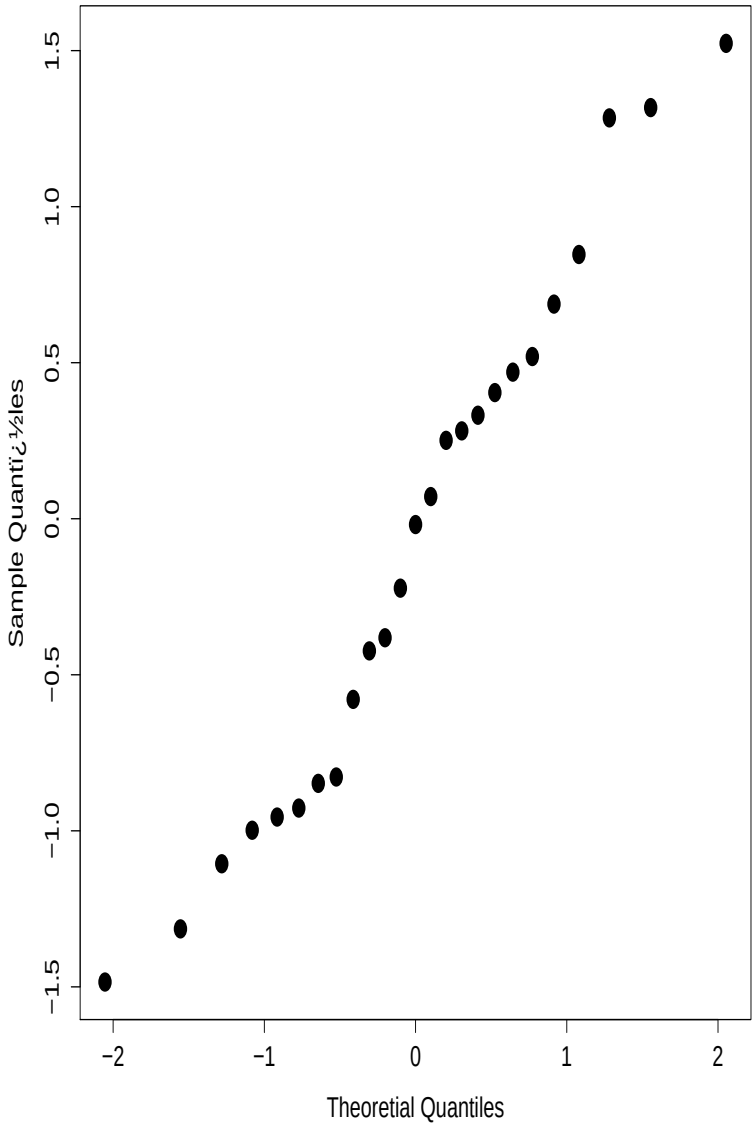
- $Y_i | x_i \overset{\text{ind}}{\sim} \text{B}\{n_i, \pi(x_i)\}$,
- $\log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} = f(x_i)$,

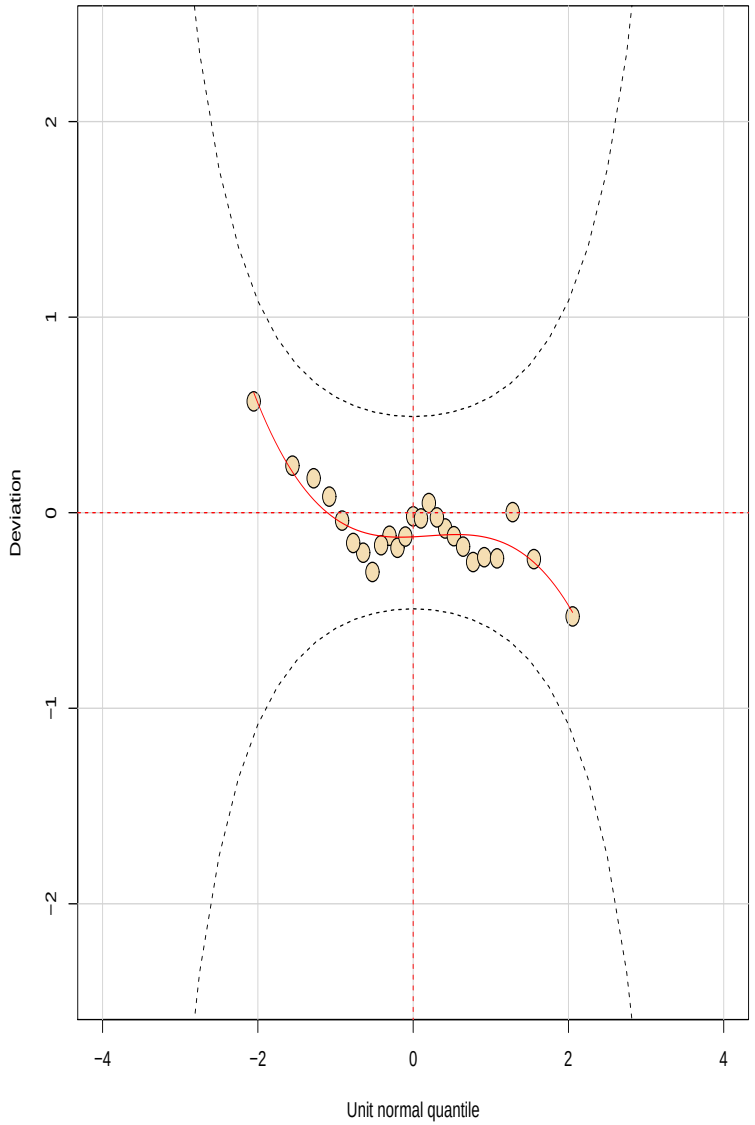
em que $f(x)$ é um B-spline.

Solução através de P-splines (k=2,d=2)

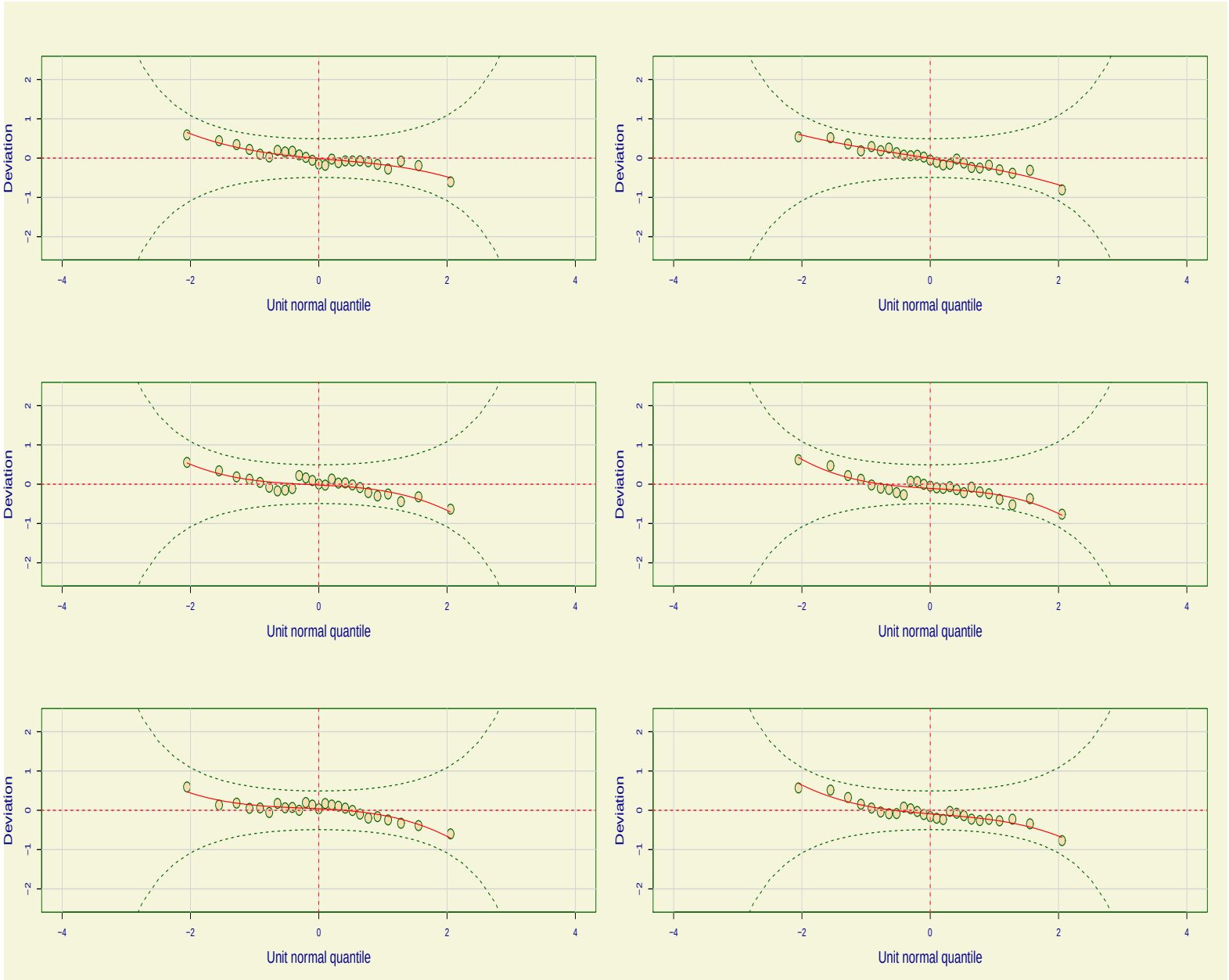
Efeito	Estimativa	E.Padrão	Valor-z	Valor-P
Constante	-20,994	0,813	-25,82	0,00
f(Idade)	1,614	0,062	26,12	0,00
df(λ)	8,764			

Gráficos Resíduo Quantílico

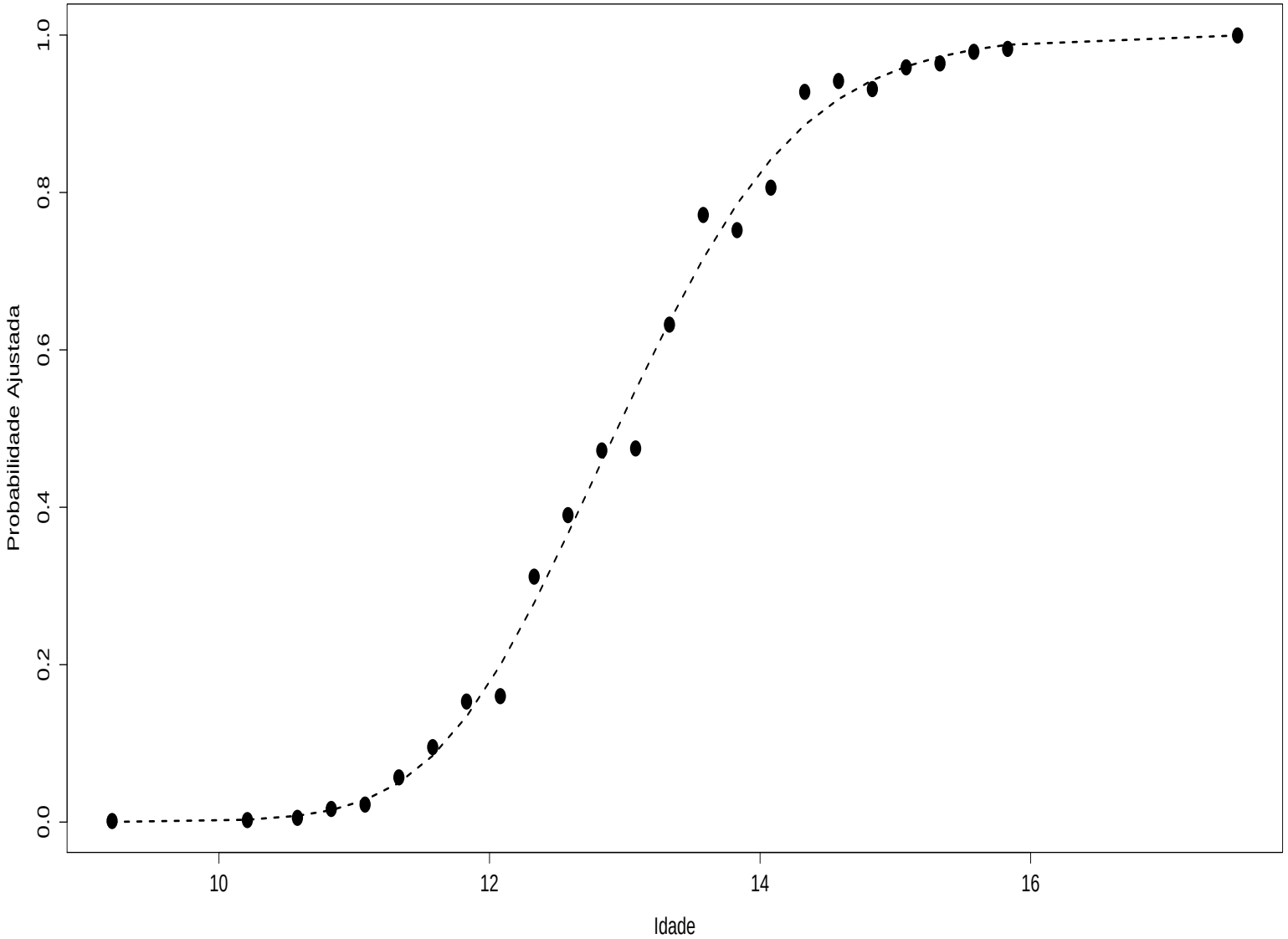




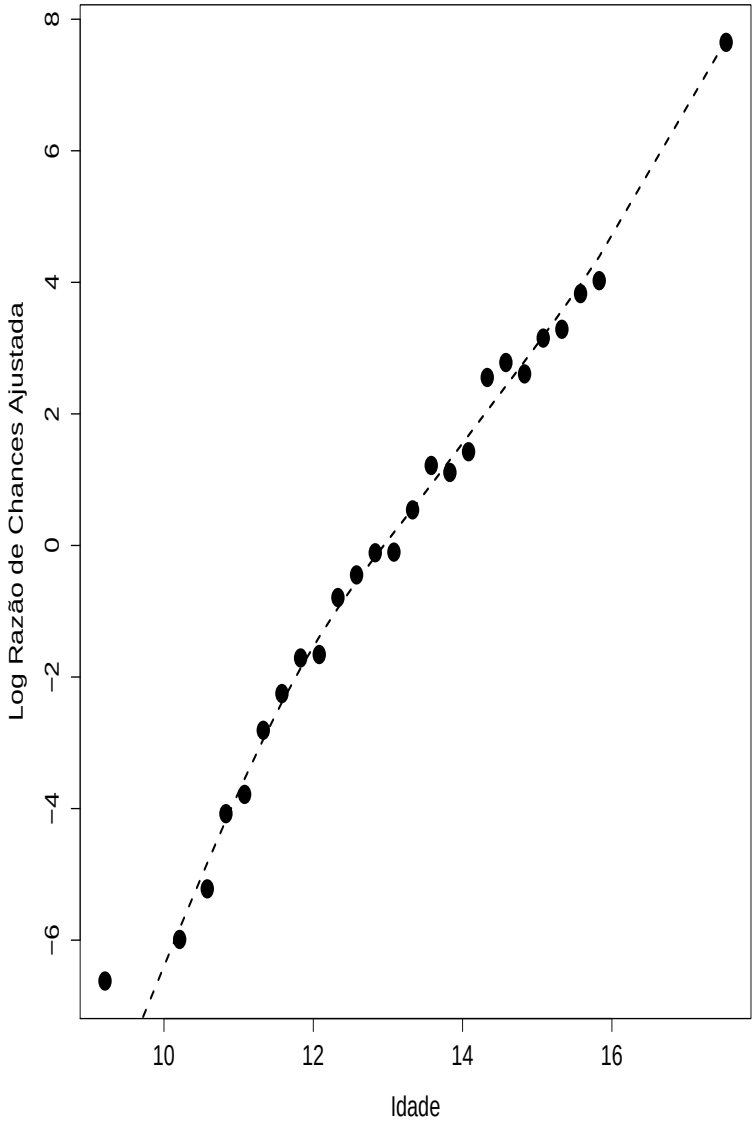
Gráficos Resíduo Quantílico

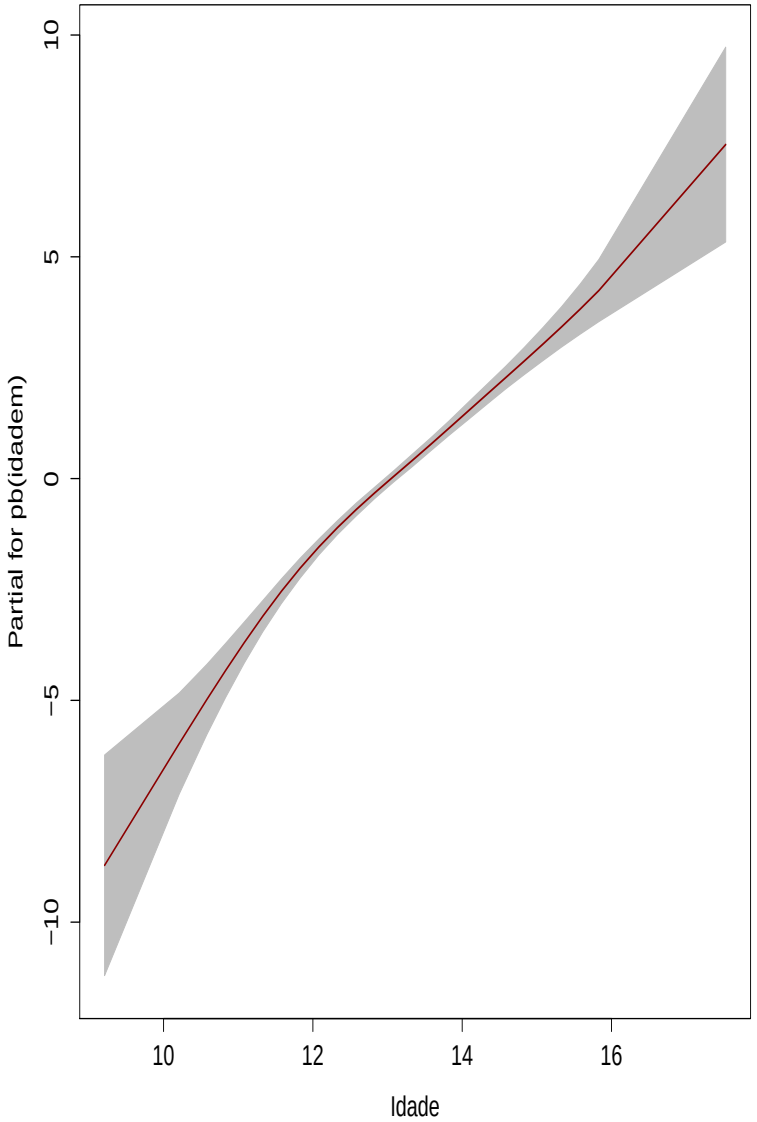


Probabildade Ajustada



Log Razão de Chances Ajustada





Descrição dos Dados

Experimento desenvolvido no Depto de Nutrição da Faculdade de Saúde Pública da USP e analisado pelo CEA (Paula, 2013):

- 5 formas diferentes de um novo tipo de snack, com baixo teor de gordura saturada e de ácidos graxos, foram comparados ao longo de 20 semanas;
- o agente responsável pela fixação do aroma do produto, a gordura vegetal hidrogenada, foi substituído totalmente ou parcialmente por óleo de canola;
- como ilustração vamos analisar a variável textura.

Descrição dos Dados

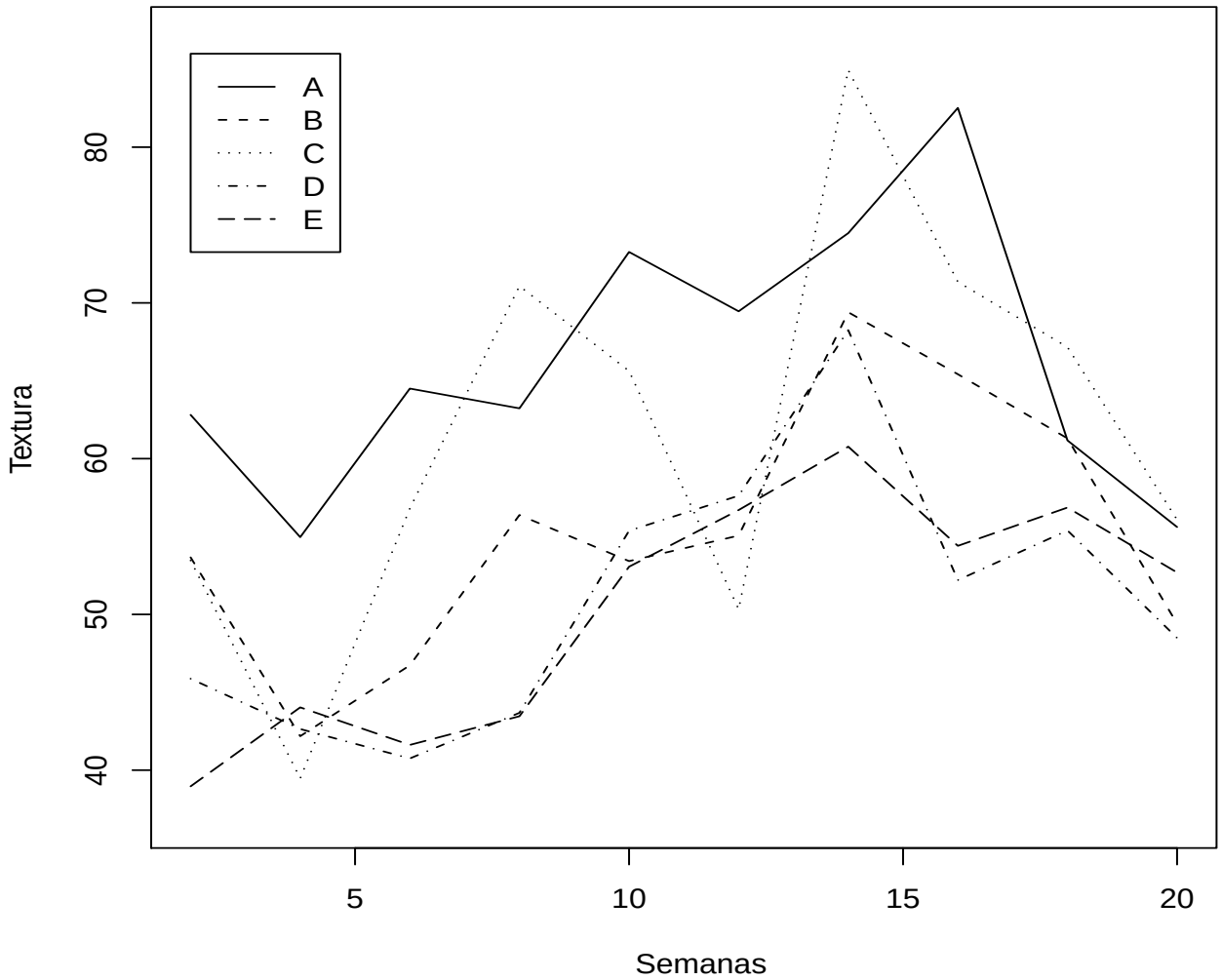
As 5 formas do novo tipo de snack foram as seguintes:

- A (22% de gordura, 0% de óleo de canola)
- B (0% de gordura, 22% de óleo de canola)
- C (17% de gordura, 5% de óleo de canola)
- D (11% de gordura, 11% de óleo de canola)
- E (5% de gordura, 17% de óleo de canola).

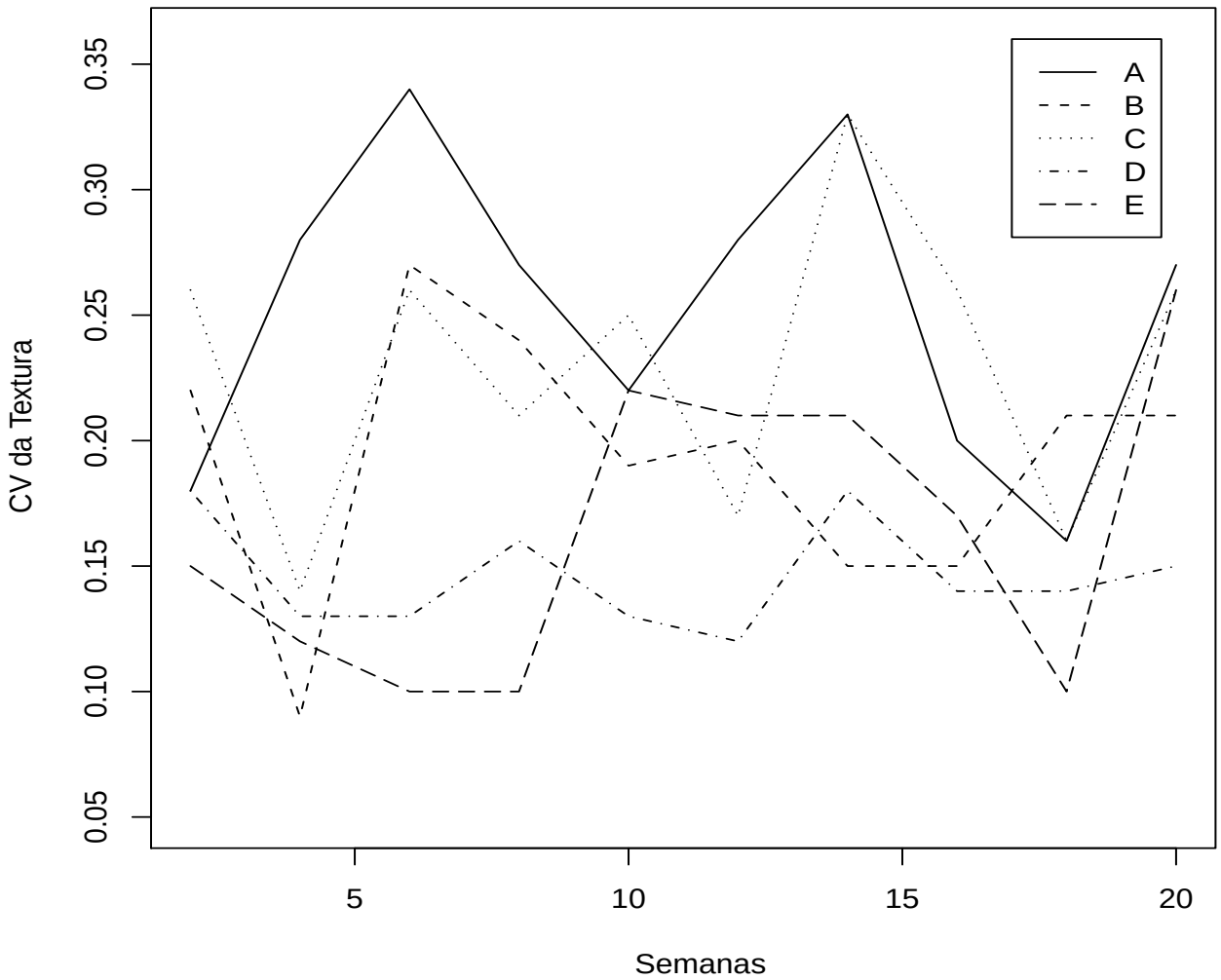
O experimento foi conduzido de modo que nas semanas pares 15 embalagens de cada um dos produtos A, B, C, D e E fossem analisadas em laboratório.

Sumário

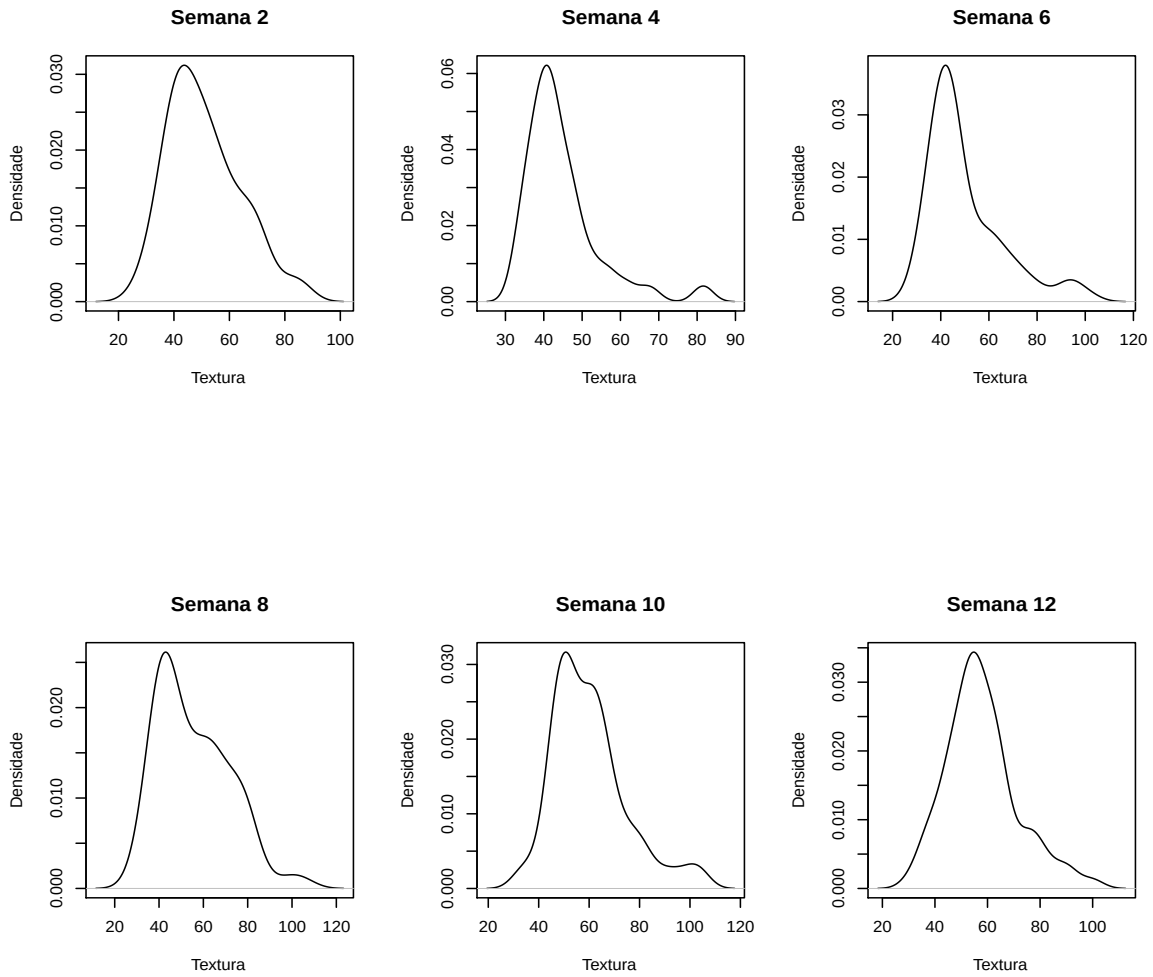
Perfil Textura Média



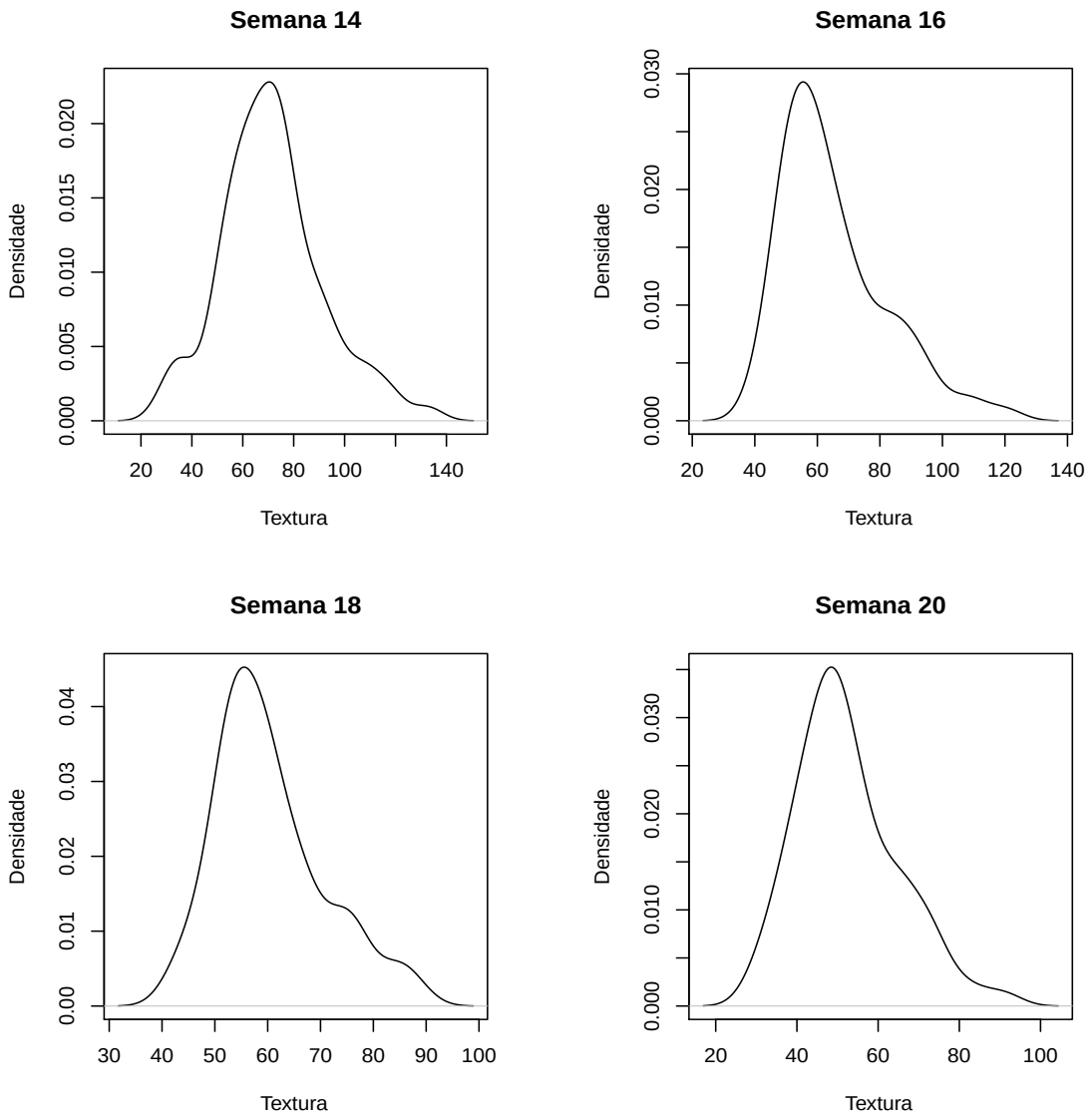
Perfil CV da Testura



Densidade Textura segundo Semana



Densidade Textura segundo Semana



Comparação de Snacks

Conclusões Preliminares

- Distribuição assimétrica à direita para a textura em cada semana.
- Tendência não linear para a textura média ao longo das semanas.
- Tendência não linear para o coeficiente de variação ao longo das semanas.

MLPAG Gama Duplo

Seja y_{ijk} a textura correspondente à k-ésima réplica do i-ésimo grupo na j-ésima semana, para $i = 1(A), 2(B), 3(C), 4(D), 5(D)$, $j = 2, 4, \ldots, 20$ e $k = 1, \ldots, 15$.Considere o seguite modelo:

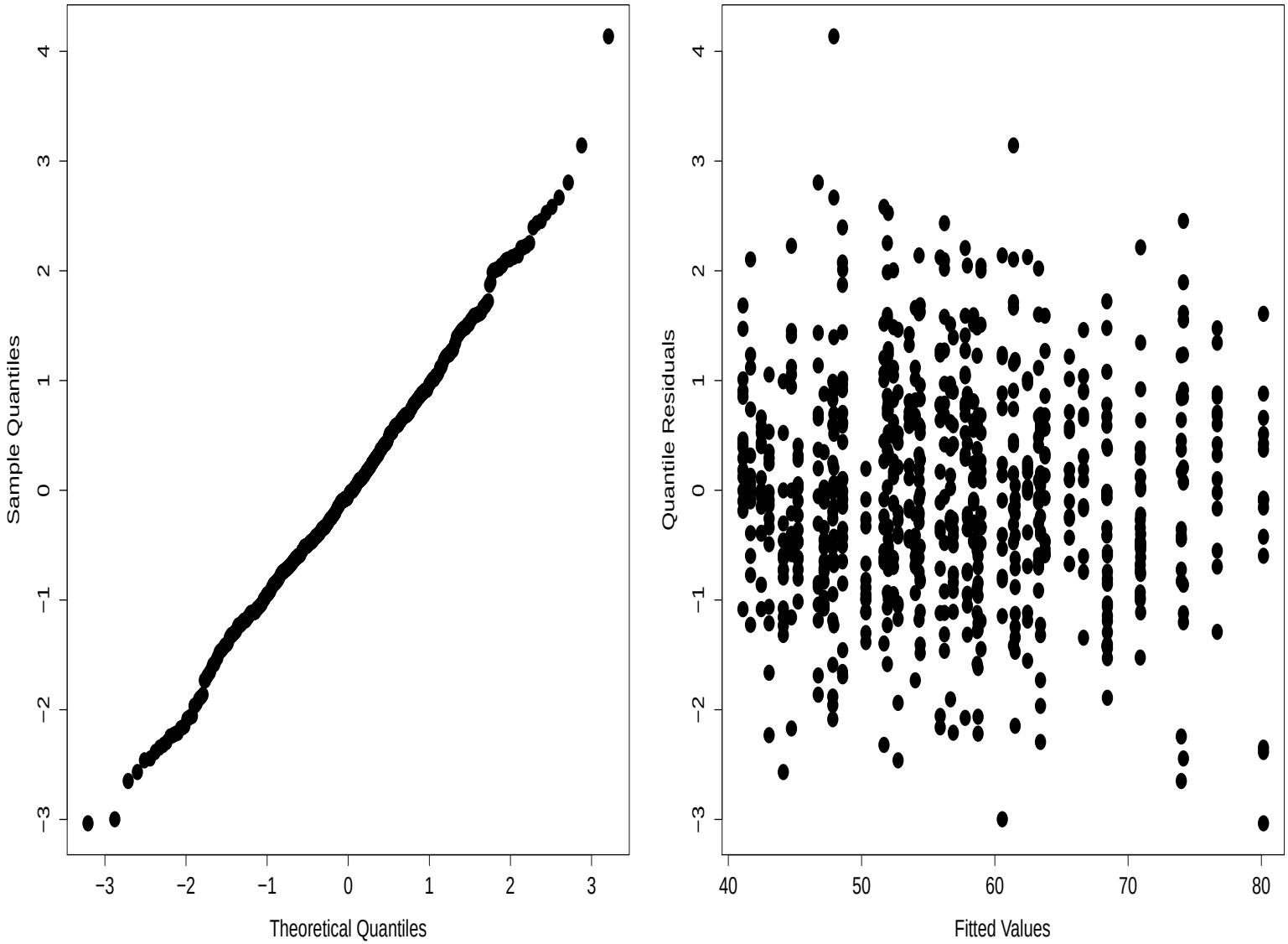
- $Y_{ijk}|(\text{grupo}, \text{semana}) \overset{\text{ind}}{\sim} \text{G}(\mu_{ij}, \sigma_{ij})$,
- $\log(\mu_{ij}) = \beta_0 + \beta_i + f_\mu(\text{semana}_j)$,
- $\log(\sigma_{ij}) = \gamma_0 + \gamma_i + f_\phi(\text{semana}_j)$,

em que $\sigma_{ij} > 0$ denota o coeficiente de variação, $f_\mu(t)$ e $f_\phi(t)$ são B-splines.

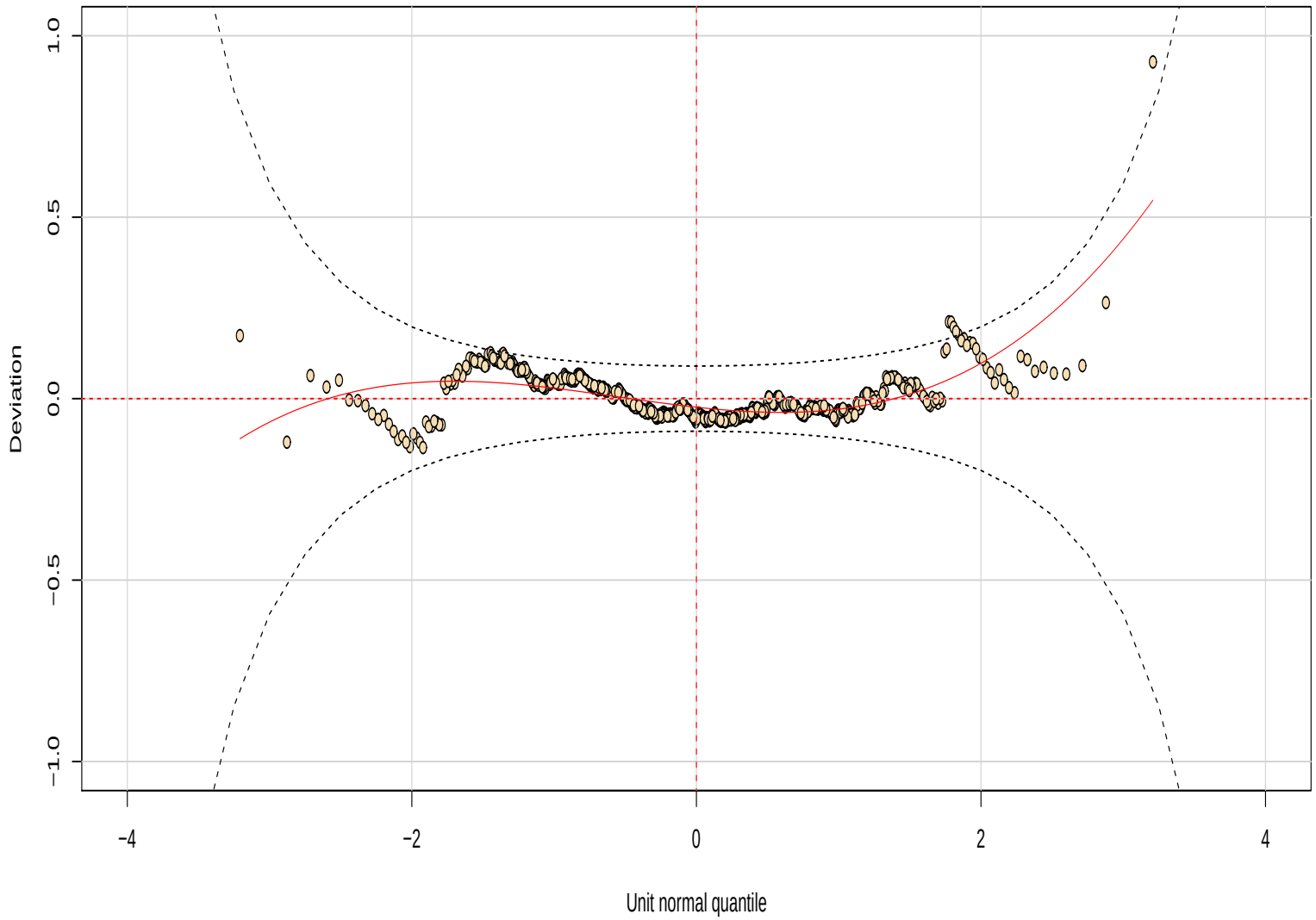
Solução através de P-splines (k=2,d=2)

	Média		Coeficiente Variação	
Efeito	Estimativa	E/E.Padrão	Estimativa	E/E.Padrão
Constante	4,023	151,76	-1,282	-22,37
Grupo B	-0,184	-6,60	-0,324	-4,01
Grupo C	-0,078	-2,46	-0,021	-0,26
Grupo D	-0,267	10,26	-0,576	-7,10
Grupo E	-0,280	-10,44	-0,452	-5,51
f(Semana)	0,015	11,66		
df(λ)	17,27			

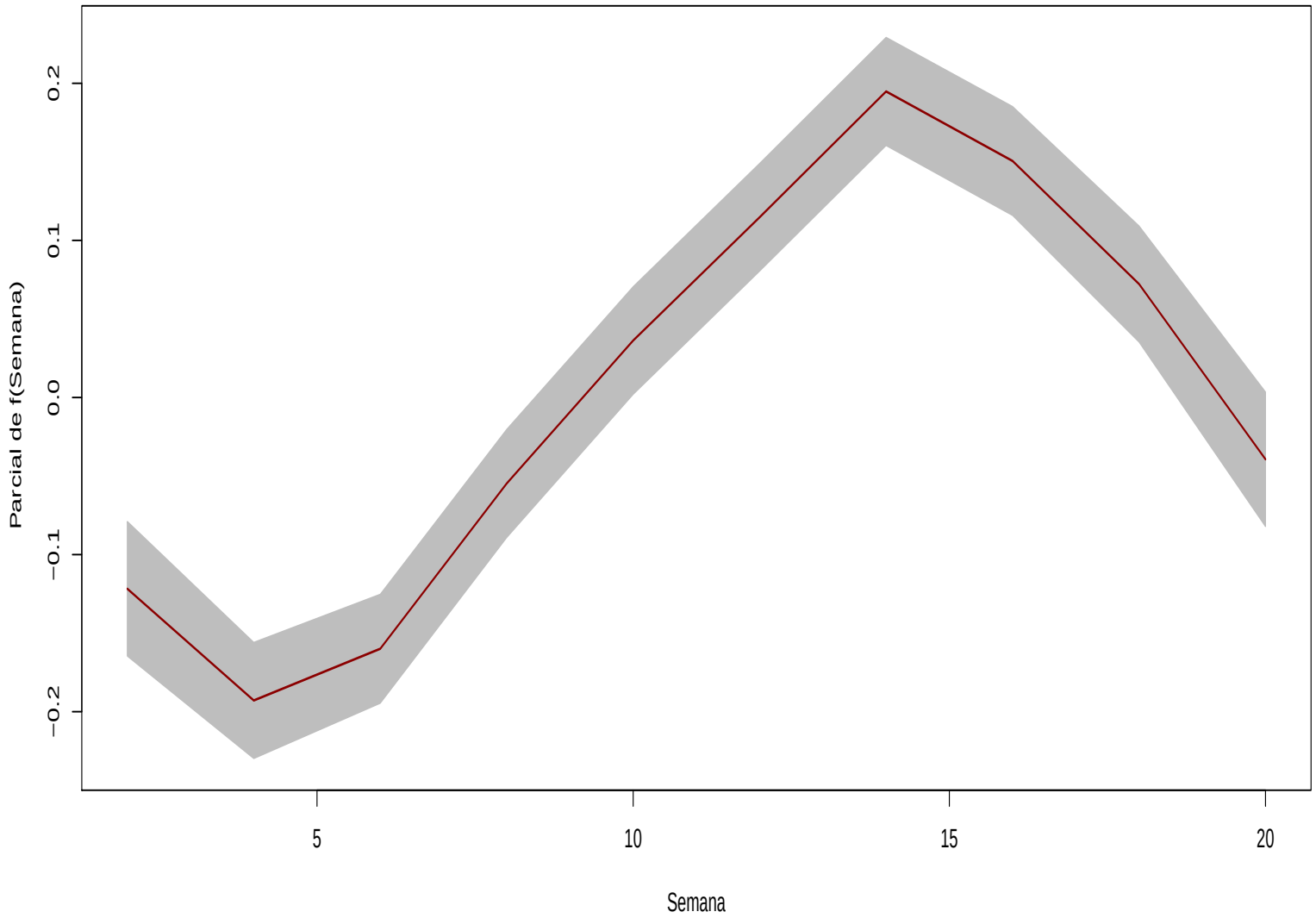
Resíduos Saída GAMLSS



Resíduos Saída GAMLSS



Banda de Confiança



Bibliotecas em R

Algumas bibliotecas que podem ser obtidas de CRAN no endereço `http://CRAN.R-project.org`:

- `gamlss`
- `mgcv`

Sumário

Referências

- De Boor C (1978). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer-Verlag, New York.
- Eilers, P. H. C., Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* 11, 89-121.
- Milicer, H. e Szczotka, F. (1966). Age at menarche in Warsaw girls in 1965. *Human Biology* **38**, 199-203.

Referências

- Paula, G. A. (2013). On diagnostics in double generalized linear models. *Computational Statistics and Data Analysis* 68, 44-51.
- Stasinopoulos, M.D., Righy, R.A., Gillian, Z.A., Voudouris, V. e de Bastiani, F. (2017). *Flexible Regression and Smoothing Using GAMLSS in R*. Chapman and Hall/CRC.
- Wood, S.N., (2017). *Generalized Additive Models: An introduction with R, Second Edition*. CRC Press, Boca Raton.