Evolução histórica dos modelos estatísticos de predição

Luiz R. Nakamura

Universidade Federal de Santa Catarina
Departamento de Informática e Estatística
luiz.nakamura@ufsc.br

www.gamlss.com

Webinários do PGMAC
Universidade Estadual de Londrina
18 de Setembro de 2020

Sumário

Dados sobre aluguel

Os dados sobre aluguel são provenientes de uma pesquisa conduzida em Abril de 1993 pela Infratest Sozialforschung, em que uma amostra aleatória sobre imóveis com novos contratos de locação nos últimos quatro anos em Munich foi selecionada. Para este conjunto de dados temos 1.967 observações e cinco variáveis:

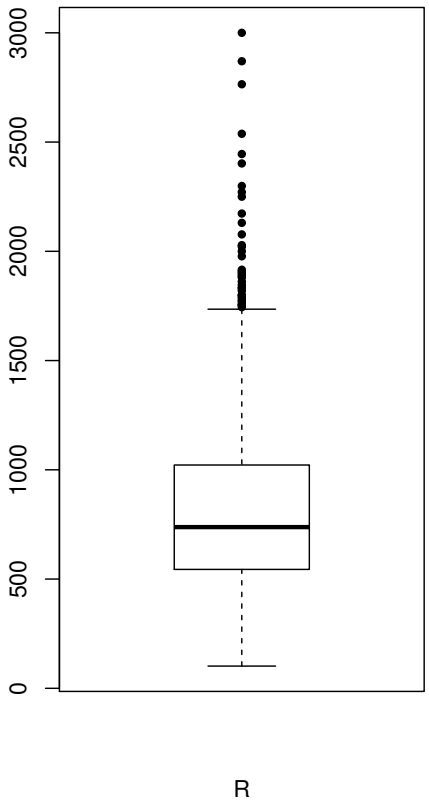
Dados sobre aluguel

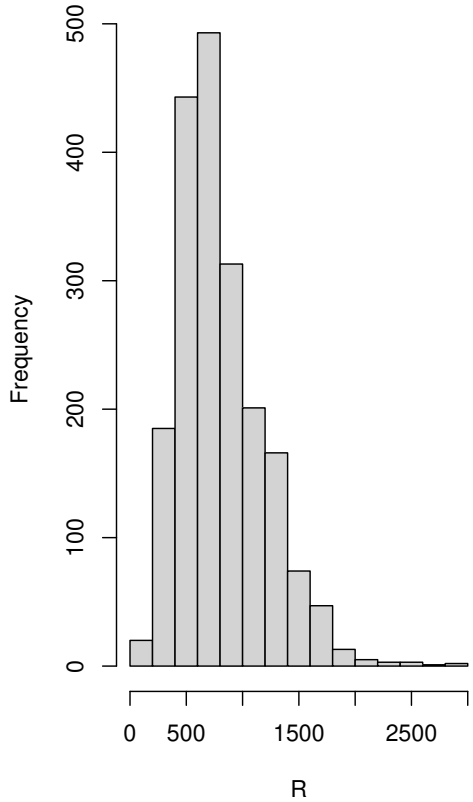
- `R`: valor do aluguel mensal (variável resposta)
- `Fl`: tamanho do imóvel em metros quadrados
- `A`: ano de construção
- `H`: fator de dois níveis, se existe (0), ou não (1), aquecimento central no imóvel
- `loc`: fator que indica a qualidade da localização do imóvel sendo: abaixo da média (1), na média (2) e acima da média (3)

$$\text{input} \rightarrow \text{output}$$

$$X \rightarrow Y$$

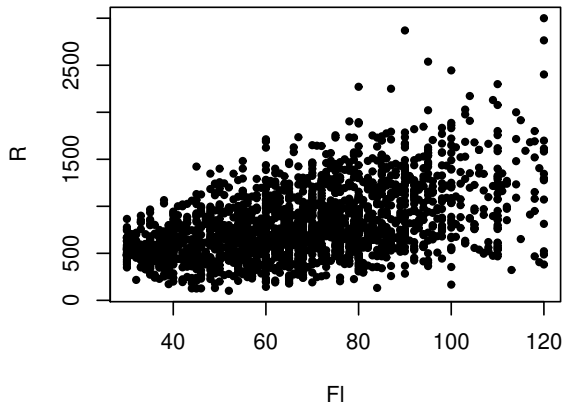
Variável resposta: valor do aluguel

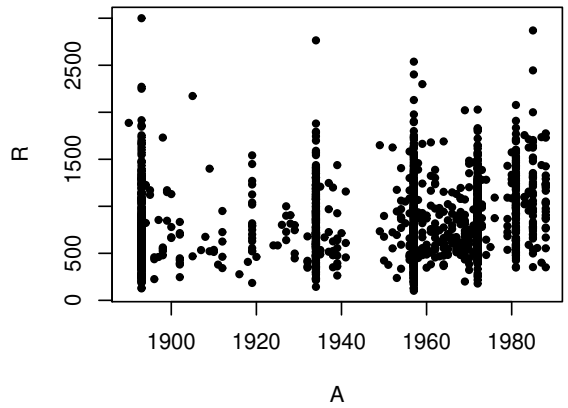


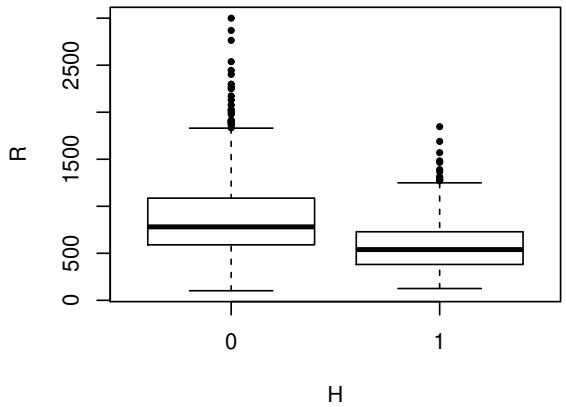


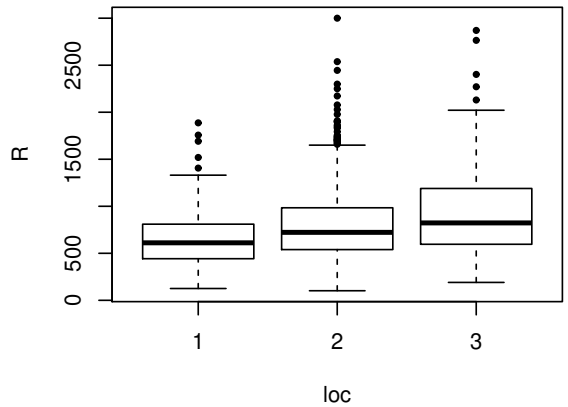
Mínimo	101.7
Máximo	3000.0
Mediana	737.8
Média	811.9
Desvio Padrão	379.00
Assimetria	1.07
Curtose	4.91

Gráficos de dispersão









Como explicar o relacionamento entre o valor do aluguel mensal e as variáveis explicativas disponíveis?

Como explicar o relacionamento entre o valor do aluguel mensal e as variáveis explicativas disponíveis?

"All models are wrong, but some are useful"
(George E.P. Box)

Modelo de regressão linear clássico

Modelo de regressão linear clássico

Definição

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_r X_{ir} + \epsilon_i, \; i = 1, \ldots, n$$

em que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Modelo de regressão linear clássico

Definição

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_r X_{ir} + \epsilon_i, \ i = 1, \ldots, n$$

em que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Matricialmente

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Modelo de regressão linear clássico

Definição

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_r X_{ir} + \epsilon_i, \; i = 1, \ldots, n$$

em que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

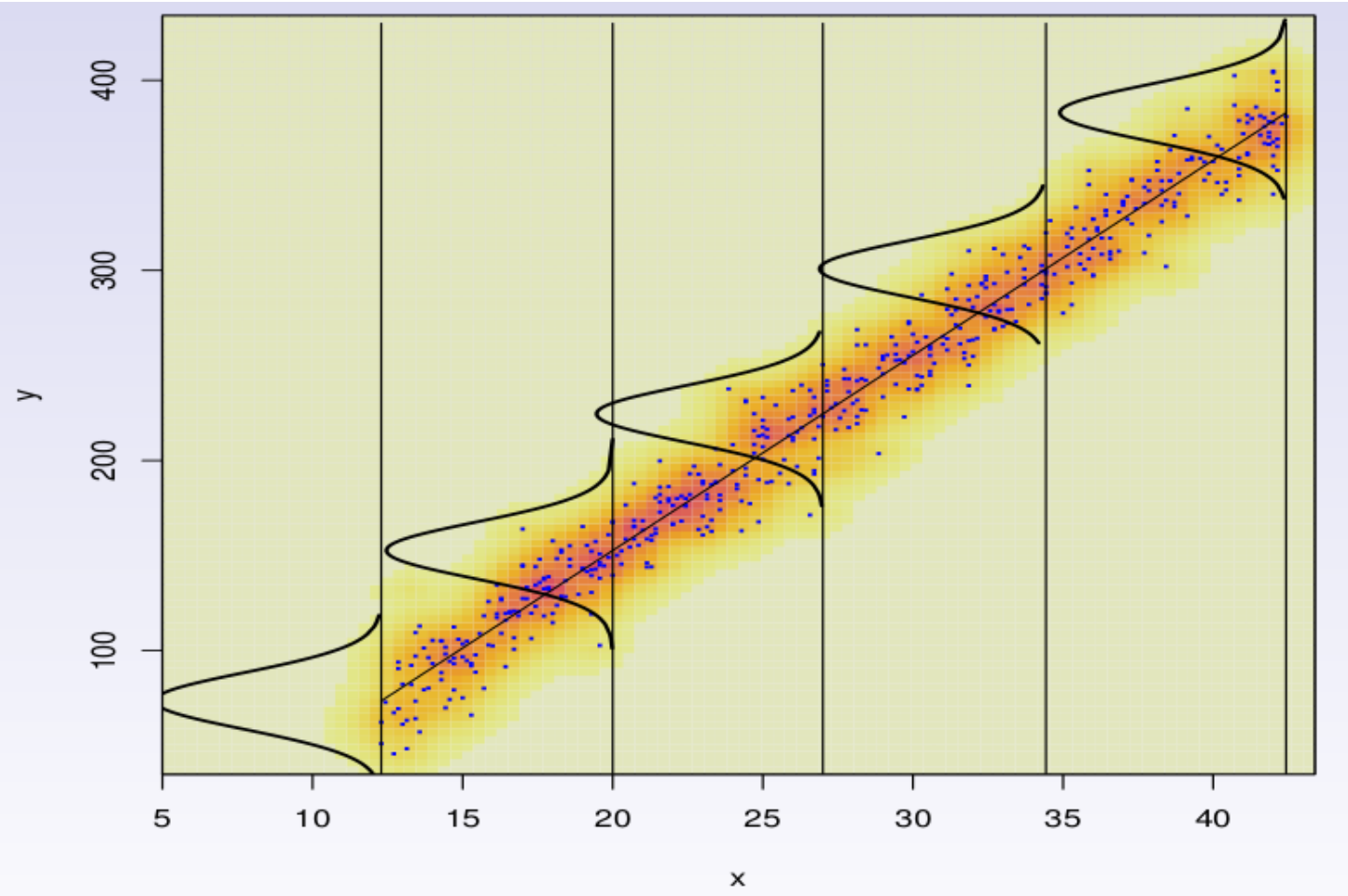
Matricialmente

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

ou ainda,

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}, \text{ em que } \boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2)$$

Suposições do modelo



Exemplo: dados sobre aluguel

R: Modelo de regressão linear

Exemplo: dados sobre aluguel

R: Modelo de regressão linear

```
library(gamlss)
data(rent) #do pacote gamlss.data
r1 <- gamlss(R ~ Fl + A + H + loc, family=NO, data=rent)
summary(r1)
```

Exemplo: dados sobre aluguel

R: Modelo de regressão linear

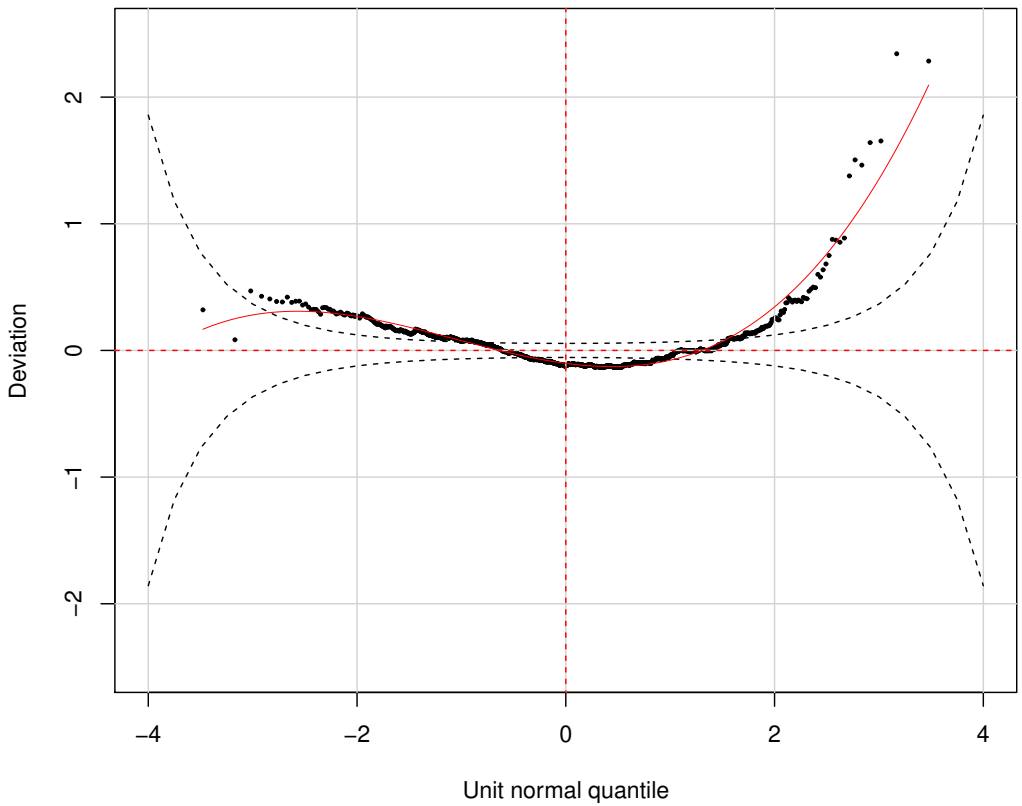
```
library(gamlss)
data(rent) #do pacote gamlss.data
r1 <- gamlss(R ~ Fl + A + H + loc, family=NO, data=rent)
summary(r1)
```

Saída

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2775.039	470.135	-5.903	4.20e-09	***
Fl	8.839	0.337	26.228	< 2e-16	***
A	1.481	0.239	6.208	6.55e-10	***
H1	-204.760	18.986	-10.785	< 2e-16	***
loc2	134.052	25.143	5.332	1.09e-07	***
loc3	209.582	27.129	7.725	1.76e-14	***

Análise de resíduos – *worm plot*

```
wp(r1)
```



Modelo linear generalizado (GLM)

(Nelder e Wedderburn, 1972)

Modelo linear generalizado (GLM)

(Nelder e Wedderburn, 1972)

Definição

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}, \text{ em que } \boldsymbol{Y} \sim \textit{ExpFamily}(\boldsymbol{\mu}, \phi)$$

em que $g(\boldsymbol{\mu})$ é chamada de função de ligação

Modelo linear generalizado (GLM)

(Nelder e Wedderburn, 1972)

Definição

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}, \text{ em que } \boldsymbol{Y} \sim \textit{ExpFamily}(\boldsymbol{\mu}, \phi)$$

em que $g(\boldsymbol{\mu})$ é chamada de função de ligação

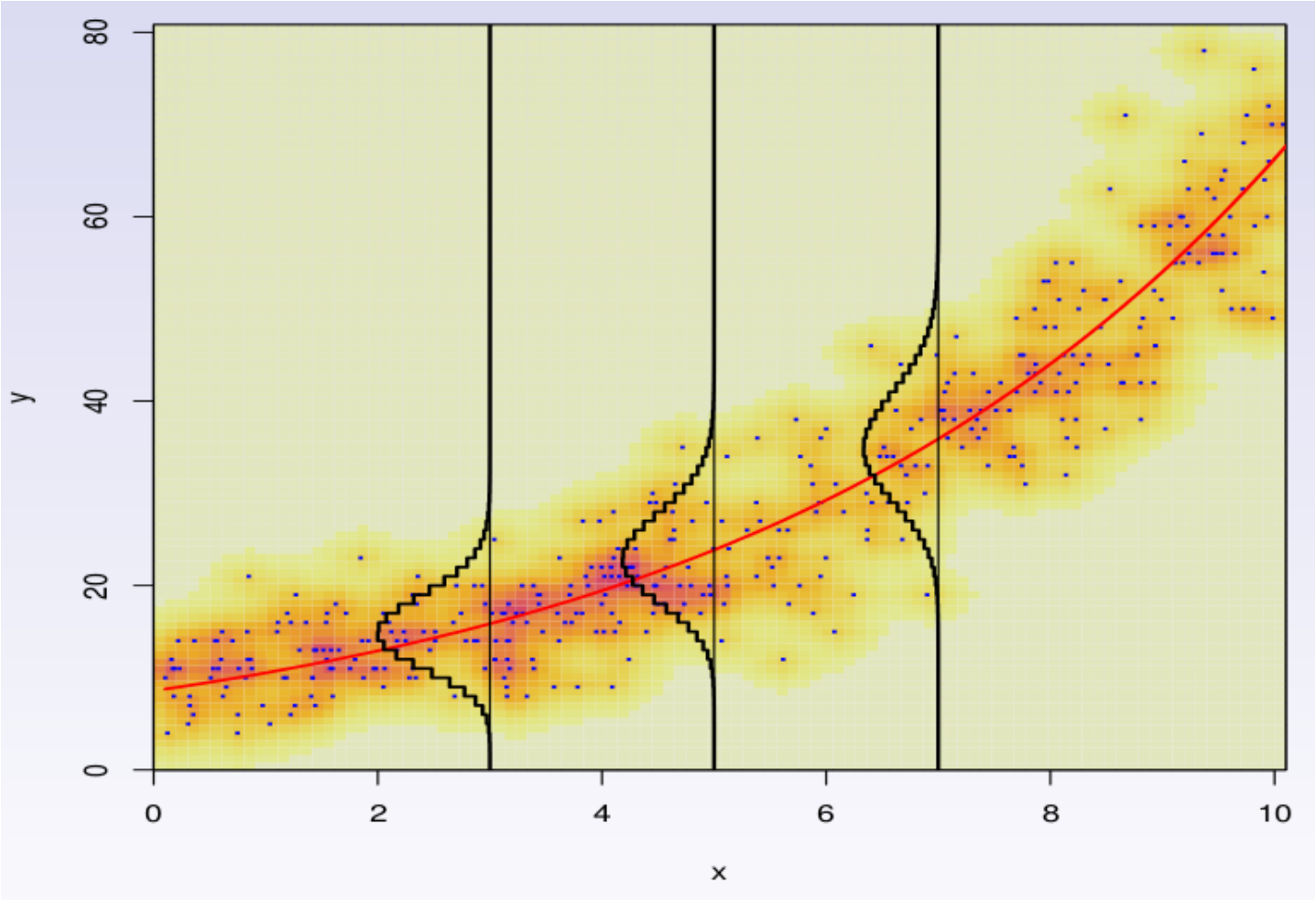
Família exponencial

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

Exemplos

- normal
- gamma
- normal inversa
- Poisson
- binomial
- entre outras...

Suposições do modelo



Exemplo: dados sobre aluguel

R: GLM (distribuição gamma)

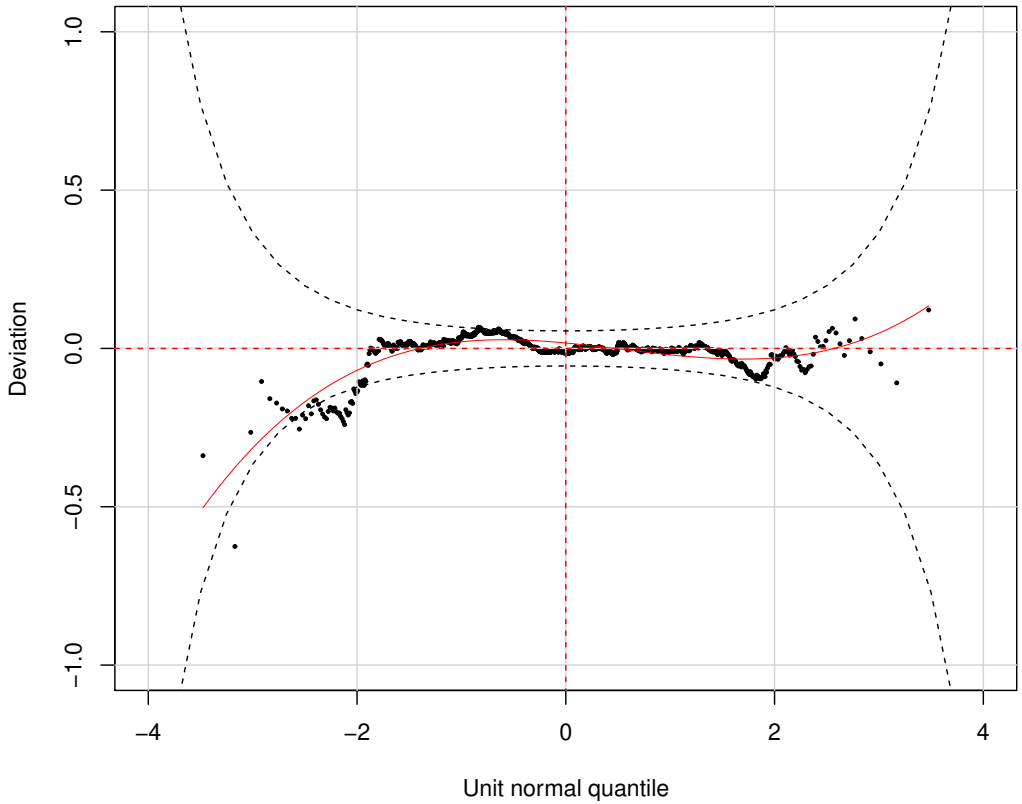
```
r2 <- gamlss(R ~ Fl + A + H + loc, family=GA, data=rent)
summary(r2)
```

Saída

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.865	0.568	5.043	5.01e-07	***
Fl	0.011	0.001	25.735	< 2e-16	***
A	0.002	0.001	5.232	1.86e-07	***
H1	-0.300	0.023	-12.982	< 2e-16	***
loc2	0.191	0.031	6.251	5.00e-10	***
loc3	0.264	0.033	8.022	1.78e-15	***

Análise de resíduos – *worm plot*

`wp(r2)`



Modelo aditivo generalizado (GAM)

(Hastie e Tibshirani, 1990)

Modelo aditivo generalizado (GAM)

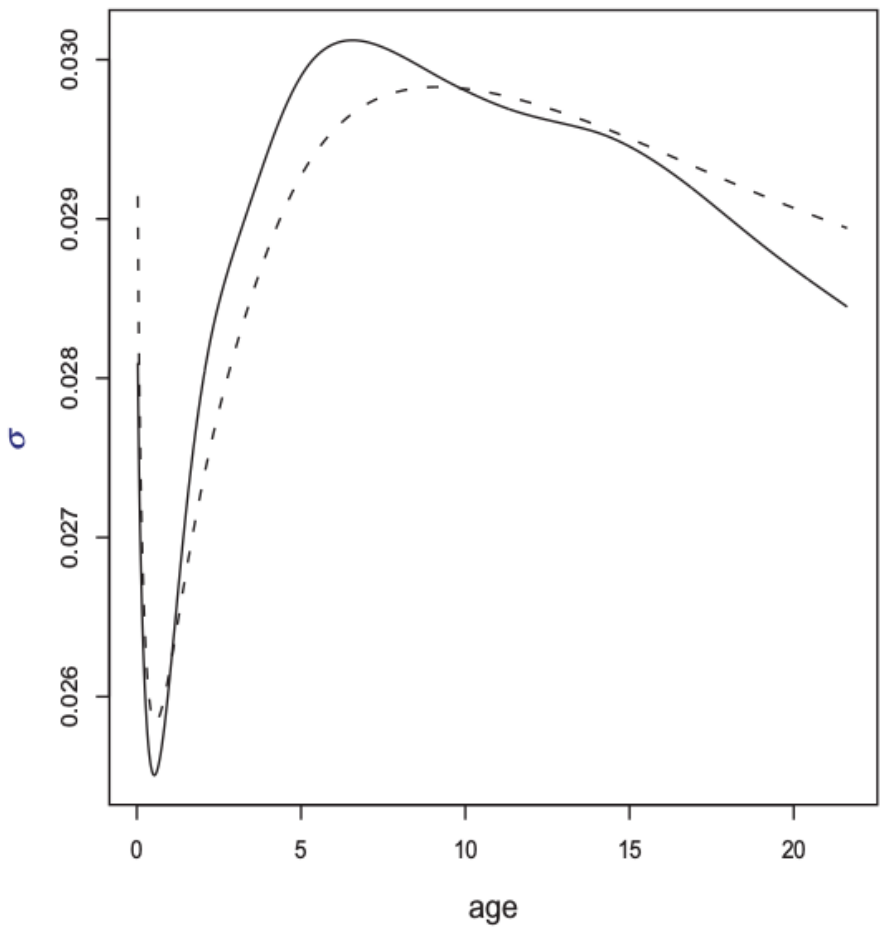
(Hastie e Tibshirani, 1990)

Definição

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{J} h_j(\boldsymbol{x}_j), \text{ em que } \boldsymbol{Y} \sim ExpFamily(\boldsymbol{\mu}, \phi)$$

em que $h_j(\boldsymbol{x}_j)$ são funções de suavização não-paramétricas

Comportamento de uma função de suavização



Fonte: R.A. Rigby and D.M. Stasinopoulos. (2006). Using the Box-Cox *t* distribution in GAMLSS to model skewness and kurtosis. Statistical Modelling, 6, 209–229.

Exemplo: dados sobre aluguel

R: GAM (distribuição gamma)

```
r3 <- gamlss(R ~ pb(Fl) + pb(A) + H + loc, family=GA,
             data=rent)
summary(r3)
```

Exemplo: dados sobre aluguel

R: GAM (distribuição gamma)

```
r3 <- gamlss(R ~ pb(Fl) + pb(A) + H + loc, family=GA,
             data=rent)
summary(r3)
```

Saída

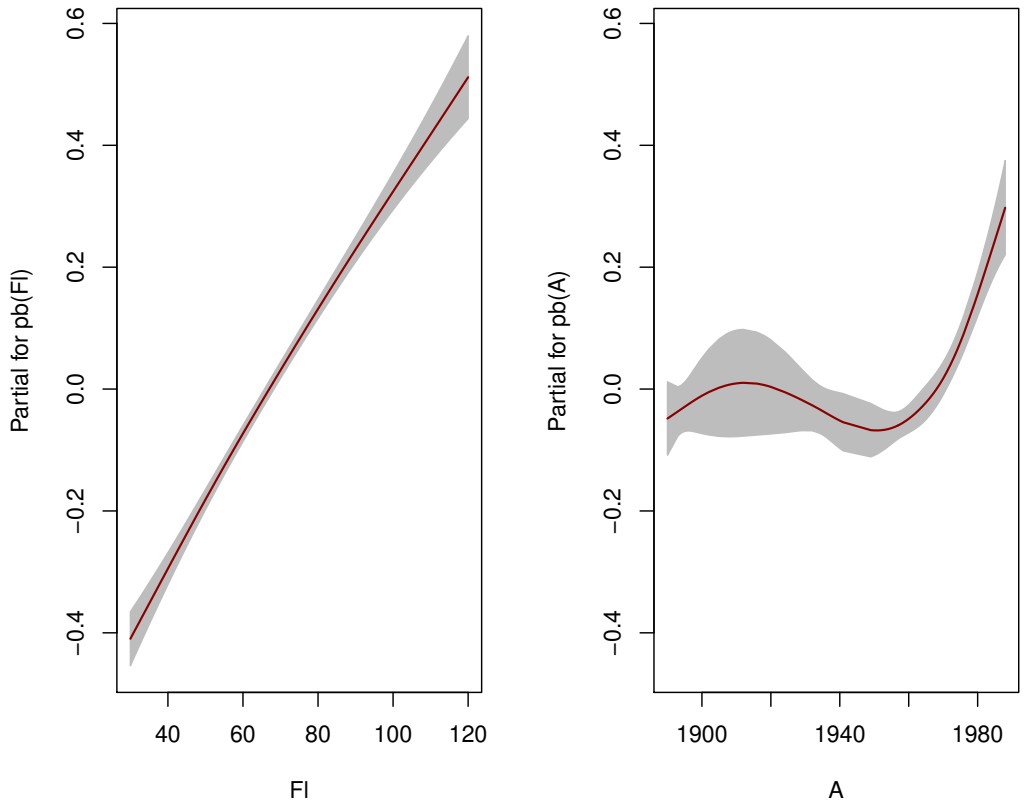
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.085	0.569	5.420	6.69e-08 ***
pb(Fl)	0.010	0.001	25.574	< 2e-16 ***
pb(A)	0.001	0.001	4.862	1.26e-06 ***
H1	-0.301	0.023	-13.318	< 2e-16 ***
loc2	0.189	0.030	6.304	3.57e-10 ***
loc3	0.272	0.032	8.425	< 2e-16 ***

NOTE: Additive smoothing terms exist in the formulas:

i) Std. Error for smoothers are for the linear effect only.

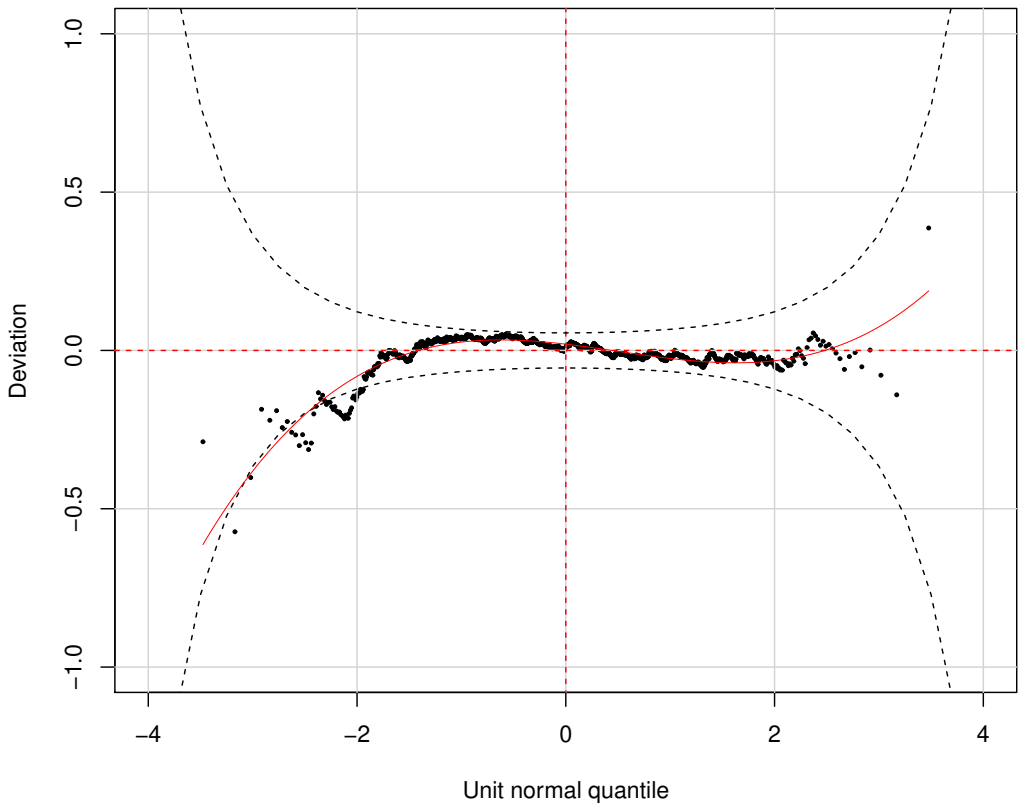
Comportamento das funções de suavização

```
term.plot(r3, pages=2)
```



Análise de resíduos – *worm plot*

`wp(r3)`



Modelos aditivos generalizados para locação, escala e forma (GAMLSS)

(Rigby e Stasinopoulos, 2005)

Modelos aditivos generalizados para locação, escala e forma (GAMLSS)

(Rigby e Stasinopoulos, 2005)

Definição

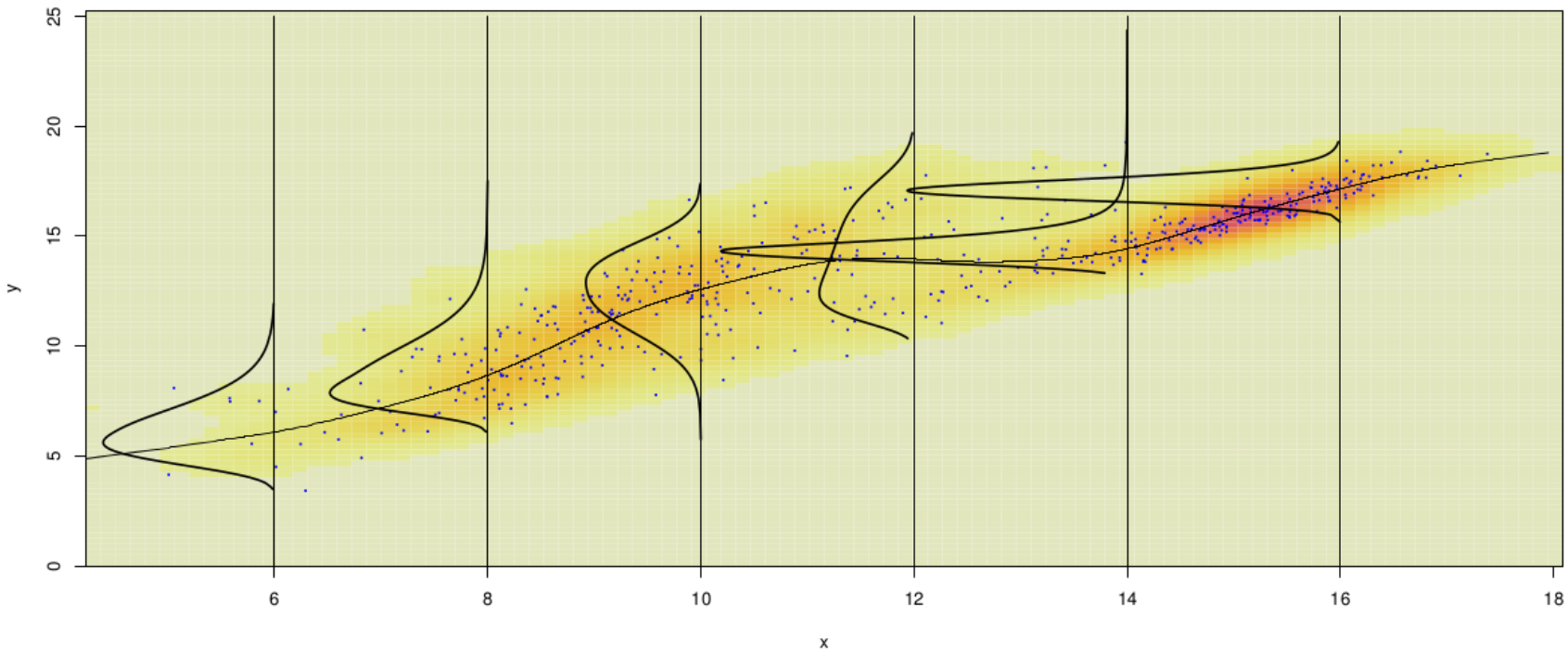
$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\boldsymbol{x}_{jk}),\ k = 1, \ldots, p$$

em que $\boldsymbol{Y} \sim \mathcal{D}(\theta_k)$ e $\mathcal{D}$ denota qualquer distribuição (não necessariamente da família exponencial) com p parâmetros

Referência

D.M. Stasinopoulos, R.A. Rigby, G.Z. Heller, V. Voudouris and F. De Bastiani. (2017). Flexible Regression and Smoothing: using GAMLSS in R, CRC Press.

Suposições do modelo



GAMLSS

- Modelos de regressão semi-paramétricos
  - paramétrico: envolve uma distribuição para a variável resposta (que não necessariamente pertence à família exponencial)
  - semi: pode envolver partes lineares (paramétrica) e/ou termos de suavização não-paramétricos
- Pacote `gamlss` no `R` (Stasinopoulos e Rigby, 2007)

Distribuições

Distribuições

Tipos de distribuição

Existem, atualmente, mais de 100 distribuições implementadas junto ao pacote `gamlss.dist`, entre elas:

- Contínuas
- Discretas
- Mistas

Importante

Parametrizações diferentes das habituais!

Referência

R.A. Rigby, D.M. Stasinopoulos, G.Z. Heller and F. De Bastiani. (2019). Distributions for Modelling Location, Scale and Shape: using GAMLSS in R, CRC Press.

Alguns pacotes no R

`gamlss`	o pacote original
`gamlss.dist`	todas as distribuições disponíveis
`gamlss.data`	conjuntos de dados
`gamlss.add`	para outros termos aditivos
`gamlss.cens`	variáveis respostas com censura
`gamlss.demo`	demos para distribuições e suavizações
`gamlss.nl`	modelos não-lineares
`gamlss.tr`	distribuições truncadas
`gamlss.mx`	distribuições de mistura e efeitos aleatórios
`gamlss.spatial`	modelos espaciais
`gamlss.inf`	distribuições inflacionadas
`gamlss.util`	outros

Seleção de modelos

Seleção de modelos

Os GAMLSS podem ser representados por $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \mathcal{L}\}$, em que

- $\mathcal{D}$ especifica a distribuição da variável resposta;
- $\mathcal{G}$ especifica o conjunto de funções de ligação para os parâmetros da distribuição μ, σ, ν e τ; e
- $\mathcal{T}$ especifica os termos que aparecem nos preditores para μ, σ, ν e τ.
- $\mathcal{L}$ especifica os parâmetros de suavização associados a cada um dos parâmetros da distribuição da variável resposta

Estratégia de seleção

Para o ajuste de um GAMLSS, todos os componentes citados no slide anterior devem ser especificados

Estratégia de seleção

Para o ajuste de um GAMLSS, todos os componentes citados no slide anterior devem ser especificados

Componente $\mathcal{D}$: seleção da distribuição

Estratégia de seleção

Para o ajuste de um GAMLSS, todos os componentes citados no slide anterior devem ser especificados

Componente $\mathcal{D}$: seleção da distribuição

1. Ajuste:

Estratégia de seleção

Para o ajuste de um GAMLSS, todos os componentes citados no slide anterior devem ser especificados

Componente $\mathcal{D}$: seleção da distribuição

1. Ajuste:
   - qual o suporte da variável resposta?

Estratégia de seleção

Para o ajuste de um GAMLSS, todos os componentes citados no slide anterior devem ser especificados

Componente $\mathcal{D}$: seleção da distribuição

1. Ajuste:
   - qual o suporte da variável resposta?
   - Critério de informação de Akaike generalizado (GAIC), dado por

$$GAIC(\kappa) = -2\,\hat{l} + \kappa\, df$$

em que df são os graus de liberdade. O modelo com menor GAIC é selecionado

Estratégia de seleção

Para o ajuste de um GAMLSS, todos os componentes citados no slide anterior devem ser especificados

Componente $\mathcal{D}$: seleção da distribuição

1. Ajuste:
   - qual o suporte da variável resposta?
   - Critério de informação de Akaike generalizado (GAIC), dado por

$$GAIC(\kappa) = -2\,\hat{l} + \kappa\, df$$

   em que df são os graus de liberdade. O modelo com menor GAIC é selecionado

2. Diagnóstico: utilização de gráficos de resíduos

Componente $\mathcal{G}$: seleção das funções de ligação

Componente $\mathcal{G}$: seleção das funções de ligação

Em geral (por *default*), nos GAMLSS, utilizamos as funções de ligação baseados no espaço dos parâmetros. Por exemplo, para a distribuição normal, temos:

Componente $\mathcal{G}$: seleção das funções de ligação

Em geral (por *default*), nos GAMLSS, utilizamos as funções de ligação baseados no espaço dos parâmetros. Por exemplo, para a distribuição normal, temos:

- $-\infty < \mu < \infty$, logo utilizamos a função identidade
- $\sigma > 0$, logo utilizamos a função logaritmica

Componente $\mathcal{G}$: seleção das funções de ligação

Em geral (por *default*), nos GAMLSS, utilizamos as funções de ligação baseados no espaço dos parâmetros. Por exemplo, para a distribuição normal, temos:

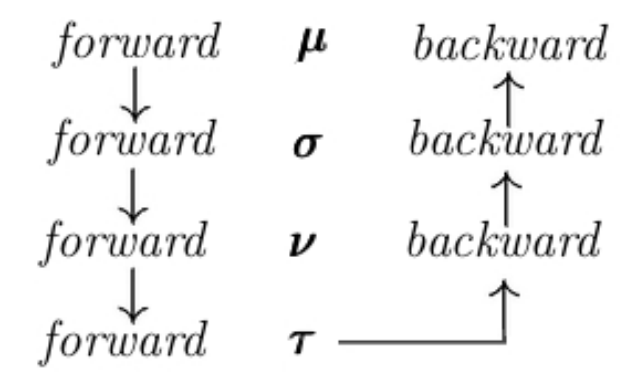
- $-\infty < \mu < \infty$, logo utilizamos a função identidade
- $\sigma > 0$, logo utilizamos a função logaritmica

A utilização de diferentes funções de ligação também pode ser comparada por meio do GAIC e de gráficos de resíduos

Componente $\mathcal{T}$: seleção das covariáveis do modelo

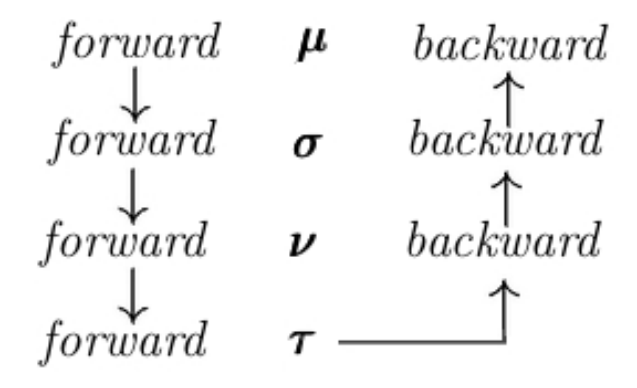
Componente $\mathcal{T}$: seleção das covariáveis do modelo

Procedimento *stepwise* baseado no GAIC (Stasinopoulos et al., 2017):



Componente $\mathcal{T}$: seleção das covariáveis do modelo

Procedimento *stepwise* baseado no GAIC (Stasinopoulos et al., 2017):



Após a realização dos passos supracitados, seleciona-se o modelo final. Observe que, por meio deste procedimento, conjuntos de variáveis explicativas distintas podem ser selecionadas para cada um dos parâmetros da distribuição da variável resposta

Componente $\mathcal{L}$: seleção dos parâmetros de suavização do modelo

Componente $\mathcal{L}$: seleção dos parâmetros de suavização do modelo

- Métodos globais (fora dos algoritmos do GAMLSS) e métodos locais (dentro dos algoritmos do GAMLSS)
  - Máxima verossimilhança
  - Critério de informação de Akaike generalizado
  - Validação cruzada

Exemplo: dados sobre aluguel

Exemplo: dados sobre aluguel

R: GAMLSS

```
f1 <- fitDist(rent$R~1, type ='realplus')
f1$fits[1:5]
```

Exemplo: dados sobre aluguel

R: GAMLSS

```
f1 <- fitDist(rent$R~1, type ='realplus')
f1$fits[1:5]
```

Saída

```
  BCCGo      GG         BCPEo      GIG        BCTo
28613.69  28614.01   28615.43   28615.55   28615.57
```

Exemplo: dados sobre aluguel

R: GAMLSS

```
f1 <- fitDist(rent$R~1, type ='realplus')
f1$fits[1:5]
```

Saída

```
  BCCGo      GG         BCPEo      GIG        BCTo
28613.69  28614.01   28615.43   28615.55   28615.57
```

Distribuição BCCGo

- $\mu > 0$: Mediana
- $\sigma > 0$: Coeficiente de variação
- $-\infty < \nu < \infty$: Assimetria

Exemplo: dados sobre aluguel

R: GAMLSS

```
f1 <- fitDist(rent$R~1, type ='realplus')
f1$fits[1:5]
```

Saída

```
  BCCGo      GG         BCPEo      GIG        BCTo
28613.69  28614.01   28615.43   28615.55   28615.57
```

Distribuição BCCGo

- $\mu > 0$: Mediana
- $\sigma > 0$: Coeficiente de variação
- $-\infty < \nu < \infty$: Assimetria

Seleção do modelo baseado na distribuição BCCGo

```
r4 <- gamlss(R~1, data=rent, family=BCCGo)
r4 <- stepGAICAll.A(r4, scope=list(lower=~1,
                    upper=~pb(Fl)+pb(A)+H+loc))
```

summary(r4)

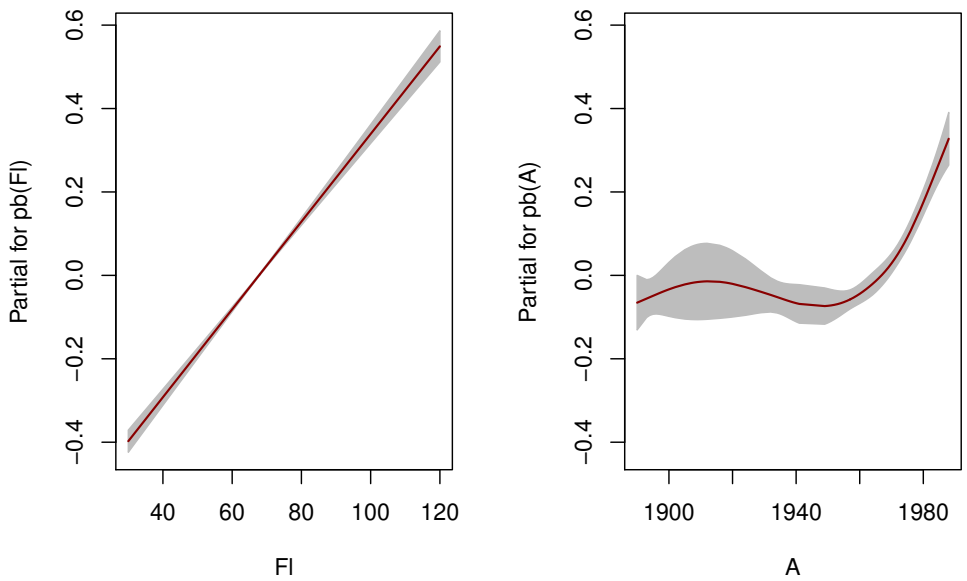
```
Mu Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)      2.005        0.579     3.460   0.000553   ***
pb(Fl)           0.011        0.001    29.478    < 2e-16   ***
H1              -0.321        0.027   -11.975    < 2e-16   ***
pb(A)            0.002        0.001     6.569   6.49e-11   ***
loc2             0.199        0.032     6.226   5.83e-10   ***
loc3             0.283        0.034     8.438    < 2e-16   ***


Sigma Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)      6.532        0.869     7.519   8.41e-14   ***
pb(A)           -0.004        0.001    -8.846    < 2e-16   ***
pb(Fl)           0.002        0.001     1.906     0.0568   .
loc2            -0.100        0.061    -1.648     0.0996   .
loc3            -0.151        0.065    -2.315     0.0207   *
H1               0.083        0.043     1.939     0.0527   .


 Nu Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
 (Intercept)    -3.748        2.946    -1.272     0.2034
 H1             -0.282        0.119    -2.371     0.0178   *
 pb(A)           0.002        0.002     1.445     0.1485
```

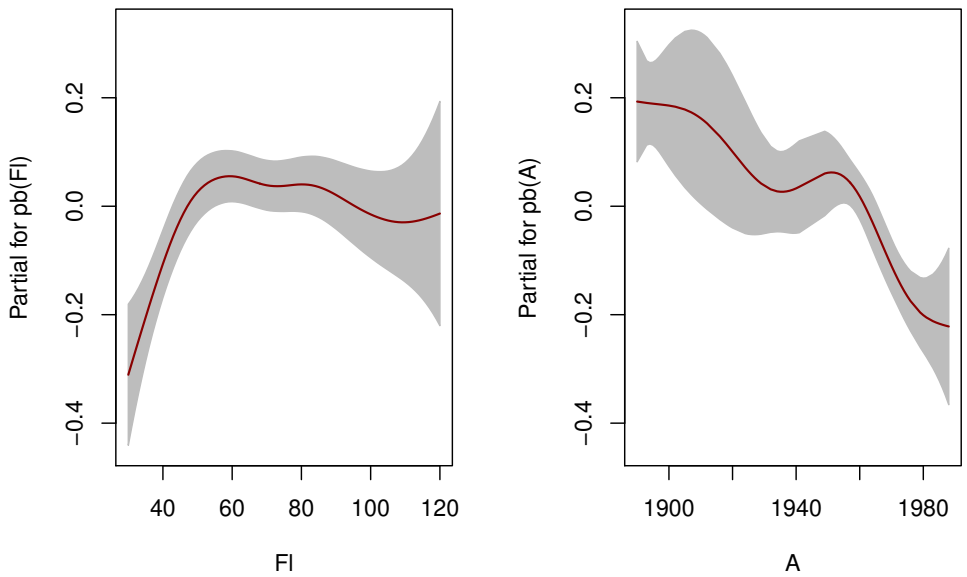
Comportamento das funções de suavização para μ

```
term.plot(r4, pages=2)
```



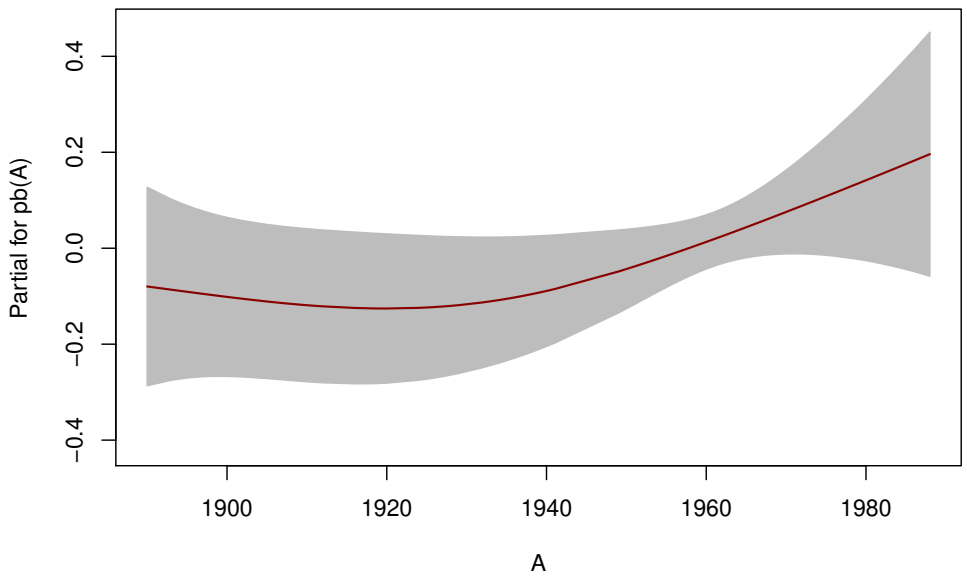
Comportamento das funções de suavização para σ

```
term.plot(r4, what='sigma', pages=2)
```



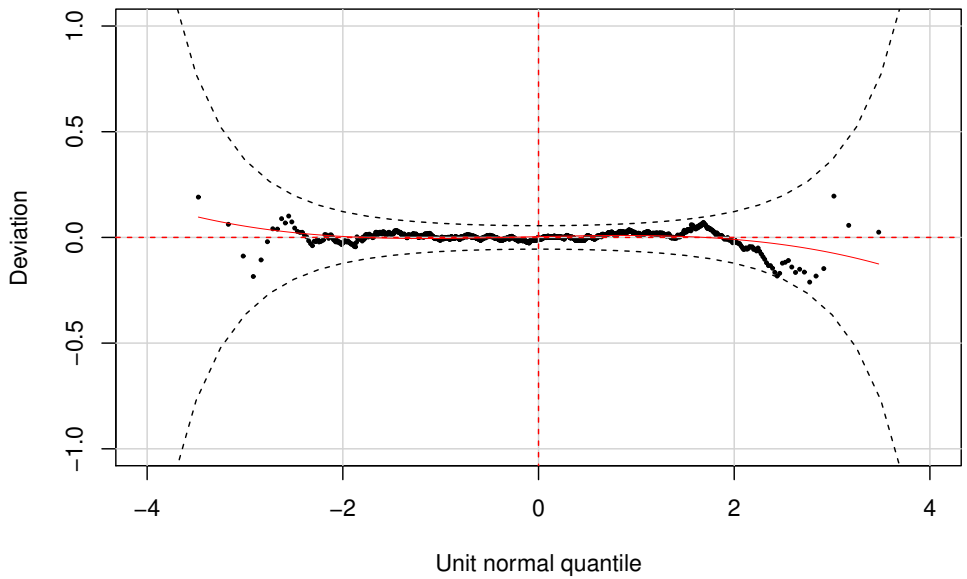
Comportamento das funções de suavização para ν

```
term.plot(r4, what='nu')
```



Análise de resíduos – *worm plot*

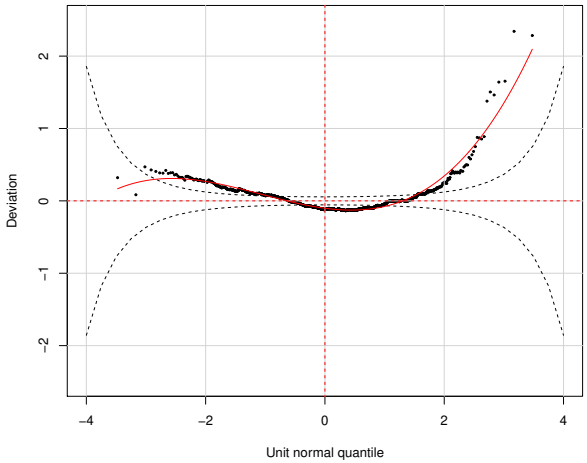
`wp(r4)`

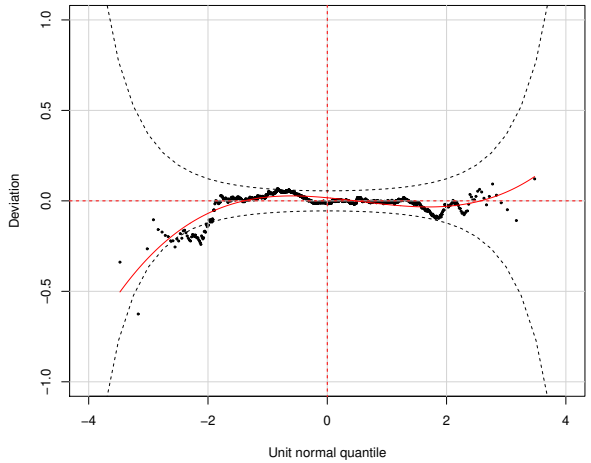


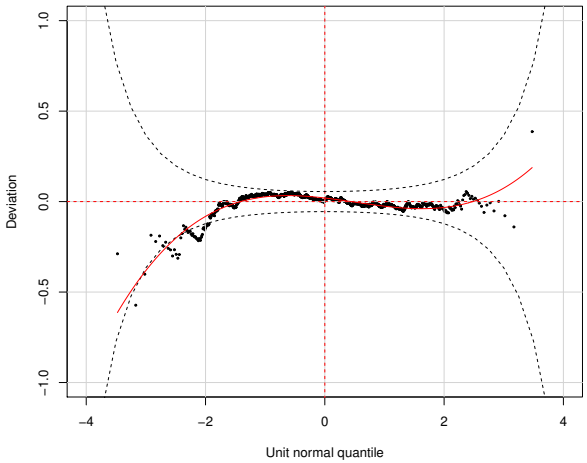
Comparação entre os modelos ajustados

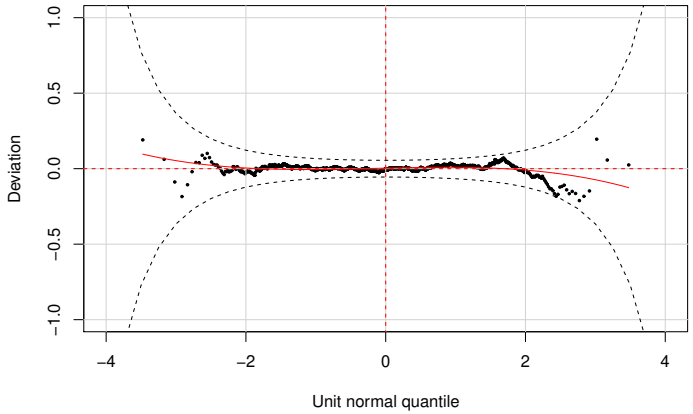
```
AIC(r1,r2,r3,r4)
```

	df	AIC
r4	25.53	27603.38
r3	11.22	27705.65
r2	7.00	27778.59
r1	7.00	28173.00


Deviation
Unit normal quantile


Deviation
Unit normal quantile


Deviation
Unit normal quantile


Deviation
Unit normal quantile

References

P.K. Dunn and G.K. Smyth. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**, 236–245.

P.H. Eilers and B.D. Marx. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

P.H. Eilers, B.D. Marx and M. Durbán. (2015). Twenty years of P-splines. *SORT* **39**, 149–186.

J.A. Nelder and R.W.M. Wedderburn. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A* **135**, 370–384.

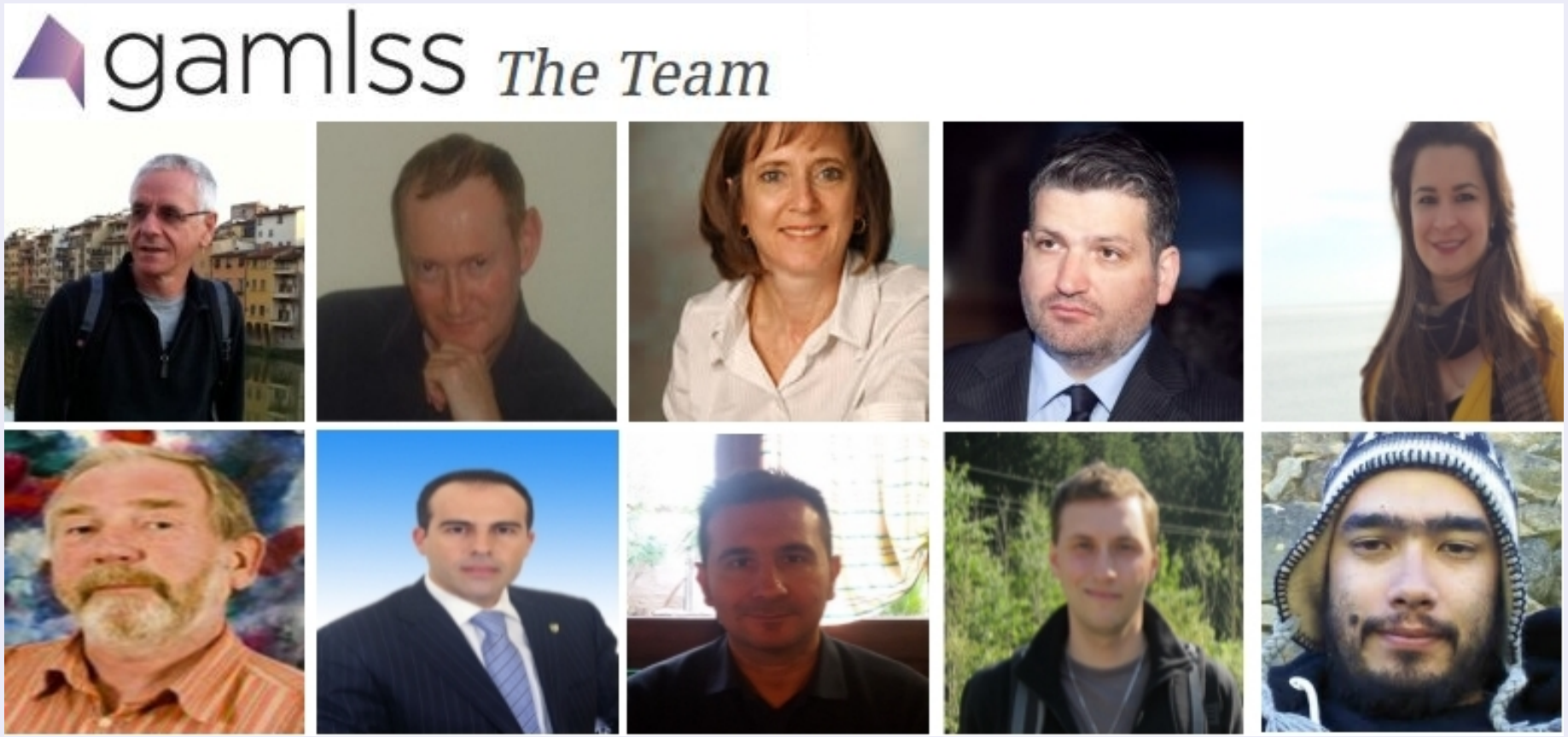
R.A. Rigby and D.M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C* **54**, 507–554.

R.A. Rigby, D.M. Stasinopoulos, G.Z. Heller and F. De Bastiani (2019). *Distributions for Modelling Location, Scale and Shape: using GAMLSS in R*, CRC Press.

D.M. Stasinopoulos and R.A. Rigby (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software* **23**, 1–10.

D.M. Stasinopoulos, R.A. Rigby, G.Z. Heller, V. Voudouris and F. De Bastiani (2017). *Flexible Regression and Smoothing: using GAMLSS in R*, CRC Press.

S. van Buuren and M. Fredriks. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine* **20**, 1259–1277.



www.gamlss.com

luiz.nakamura@ufsc.br