CE043 - GAMLSS

Por que GAMLSS?

Taconeli, C.A.; Silva, J.P

01 de agosto, 2020

Conteúdo

Introdução

GAMLSS

GAMLSS (*Generalized additive models for location, scale and shape*) configura uma família de modelos de regressão (semi) paramétricos, contemplando uma grande variedade de distribuições, e em que qualquer parâmetro da distribuição pode ser modelado em função de covariáveis.

Introdução

GAMLSS

GAMLSS (*Generalized additive models for location, scale and shape*) configura uma família de modelos de regressão (semi) paramétricos, contemplando uma grande variedade de distribuições, e em que qualquer parâmetro da distribuição pode ser modelado em função de covariáveis.

- Modelar dados com dispersão não constante, diferentes níveis de assimetria e curtose, relações não lineares...

Introdução

GAMLSS

GAMLSS (*Generalized additive models for location, scale and shape*) configura uma família de modelos de regressão (semi) paramétricos, contemplando uma grande variedade de distribuições, e em que qualquer parâmetro da distribuição pode ser modelado em função de covariáveis.

- Modelar dados com dispersão não constante, diferentes níveis de assimetria e curtose, relações não lineares...

- O pacote `gamlss` e pacotes complementares permitem ajustar modelos da classe GAMLSS para diferentes tipos e estruturas de dados.

Metodologia da aula

- Motivação baseada em dados de preços de aluguel de imóveis em Munique, 1980 (data set `rent`, pacote `gamlss`).

Metodologia da aula

- Motivação baseada em dados de preços de aluguel de imóveis em Munique, 1980 (data set `rent`, pacote `gamlss`).

- Vamos explorar, de maneira preliminar, recursos computacionais implementados na biblioteca `gamlss` do R.

Metodologia da aula

- Motivação baseada em dados de preços de aluguel de imóveis em Munique, 1980 (data set `rent`, pacote `gamlss`).
- Vamos explorar, de maneira preliminar, recursos computacionais implementados na biblioteca `gamlss` do R.
- Vamos ajustar uma sequência de modelos com nível crescente de complexidade, partindo de uma regressão linear.

Metodologia da aula

- Motivação baseada em dados de preços de aluguel de imóveis em Munique, 1980 (data set `rent`, pacote `gamlss`).
- Vamos explorar, de maneira preliminar, recursos computacionais implementados na biblioteca `gamlss` do R.
- Vamos ajustar uma sequência de modelos com nível crescente de complexidade, partindo de uma regressão linear.
- Os scripts em R estão disponíveis na página da disciplina.

Exemplo - Modelagem dos preços de aluguel de imóveis

Dados sobre preços de aluguel em Munique, 1980

- Dados de 1969 imóveis disponíveis para locação em nove variáveis, das quais cinco serão usadas na análise:

Dados sobre preços de aluguel em Munique, 1980

- Dados de 1969 imóveis disponíveis para locação em nove variáveis, das quais cinco serão usadas na análise:
    - R: valor de aluguel (em Marcos alemães) descontado o custo utilitário do imóvel;

Dados sobre preços de aluguel em Munique, 1980

- Dados de 1969 imóveis disponíveis para locação em nove variáveis, das quais cinco serão usadas na análise:
  - R: valor de aluguel (em Marcos alemães) descontado o custo utilitário do imóvel;
  - Fl: Área construída, em metros quadrados;

Dados sobre preços de aluguel em Munique, 1980

- Dados de 1969 imóveis disponíveis para locação em nove variáveis, das quais cinco serão usadas na análise:
  - R: valor de aluguel (em Marcos alemães) descontado o custo utilitário do imóvel;
  - Fl: Área construída, em metros quadrados;
  - A: Ano de construção;

Dados sobre preços de aluguel em Munique, 1980

- Dados de 1969 imóveis disponíveis para locação em nove variáveis, das quais cinco serão usadas na análise:
  - R: valor de aluguel (em Marcos alemães) descontado o custo utilitário do imóvel;
  - Fl: Área construída, em metros quadrados;
  - A: Ano de construção;
  - H: Fator com dois níveis, (0) para imóveis com aquecimento central e (1) para os que não tem;

Dados sobre preços de aluguel em Munique, 1980

- Dados de 1969 imóveis disponíveis para locação em nove variáveis, das quais cinco serão usadas na análise:
  - R: valor de aluguel (em Marcos alemães) descontado o custo utilitário do imóvel;
  - Fl: Área construída, em metros quadrados;
  - A: Ano de construção;
  - H: Fator com dois níveis, (0) para imóveis com aquecimento central e (1) para os que não tem;
  - loc: Fator com três níveis, (1) se a localização do imóvel é classificada como abaixo da média, (2) na média e (3) acima da média.

Análise exploratória

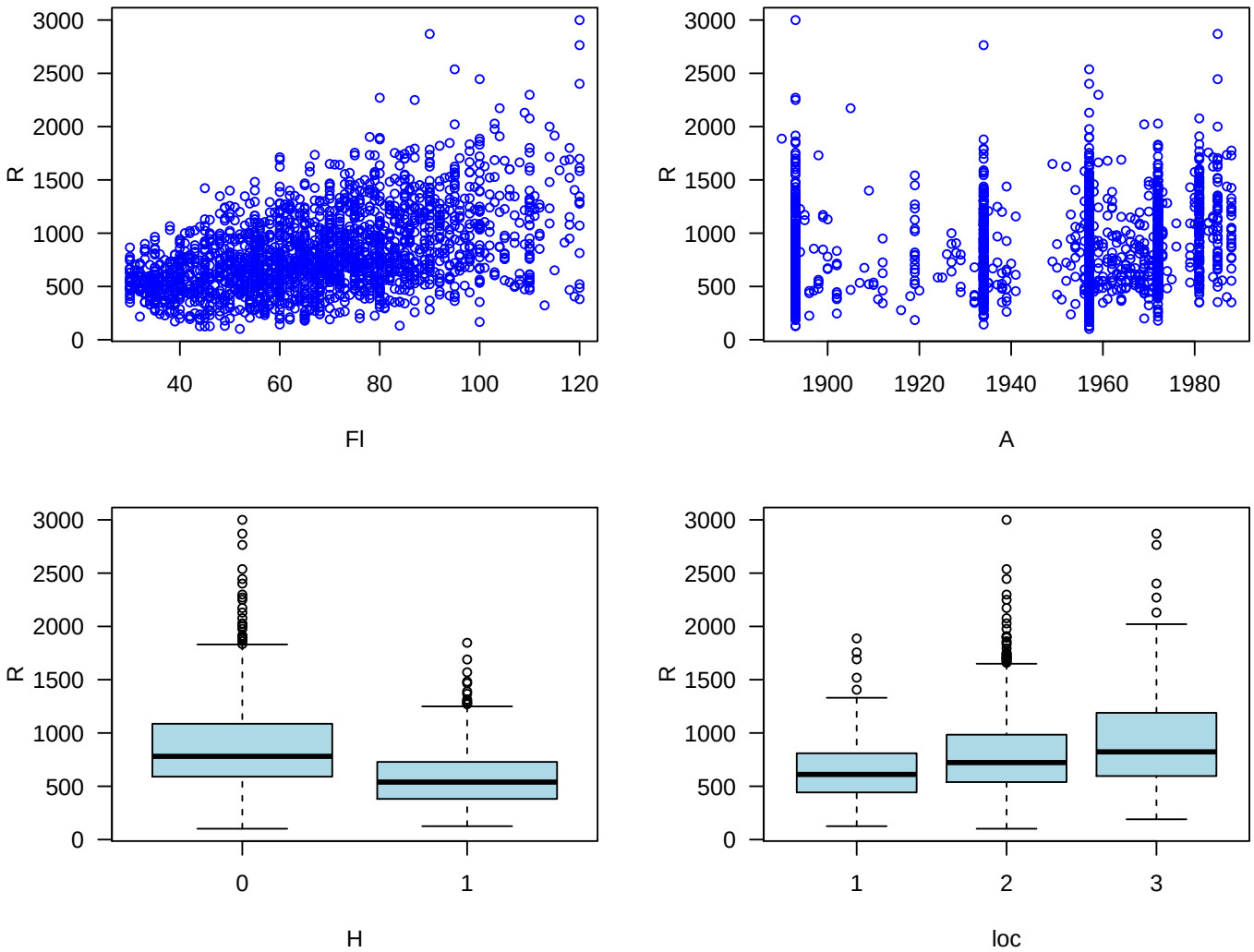

Figura 1: Gráficos para o valor de aluguel *vs* variáveis explicativas.

Alguns achados

- Não linearidade da relação entre a variável resposta e alguma(s) variáveis explicativas;

Alguns achados

- Não linearidade da relação entre a variável resposta e alguma(s) variáveis explicativas;

- A variância dos valores de aluguel não é constante para diferentes valores das variáveis explicativas;

Alguns achados

- Não linearidade da relação entre a variável resposta e alguma(s) variáveis explicativas;
- A variância dos valores de aluguel não é constante para diferentes valores das variáveis explicativas;
- Distribuição dos valores de aluguel é assimétrica, e o nível de assimetria aparentemente varia conforme os valores das variáveis explicativas.

Modelo linear

Modelo de regressão linear

- Para o modelo de regressão linear assumimos que a variável resposta, condicional aos valores das covariáveis, tem distribuição normal.

Modelo de regressão linear

- Para o modelo de regressão linear assumimos que a variável resposta, condicional aos valores das covariáveis, tem distribuição normal.

- A média da distribuição pode ser expressa como uma combinação linear de covariáveis e parâmetros associados.

Modelo de regressão linear

- Para o modelo de regressão linear assumimos que a variável resposta, condicional aos valores das covariáveis, tem distribuição normal.

- A média da distribuição pode ser expressa como uma combinação linear de covariáveis e parâmetros associados.

- A variância, no entanto, é assumida constante, para quaisquer valores das covariáveis sob estudo.

Modelo de regressão linear

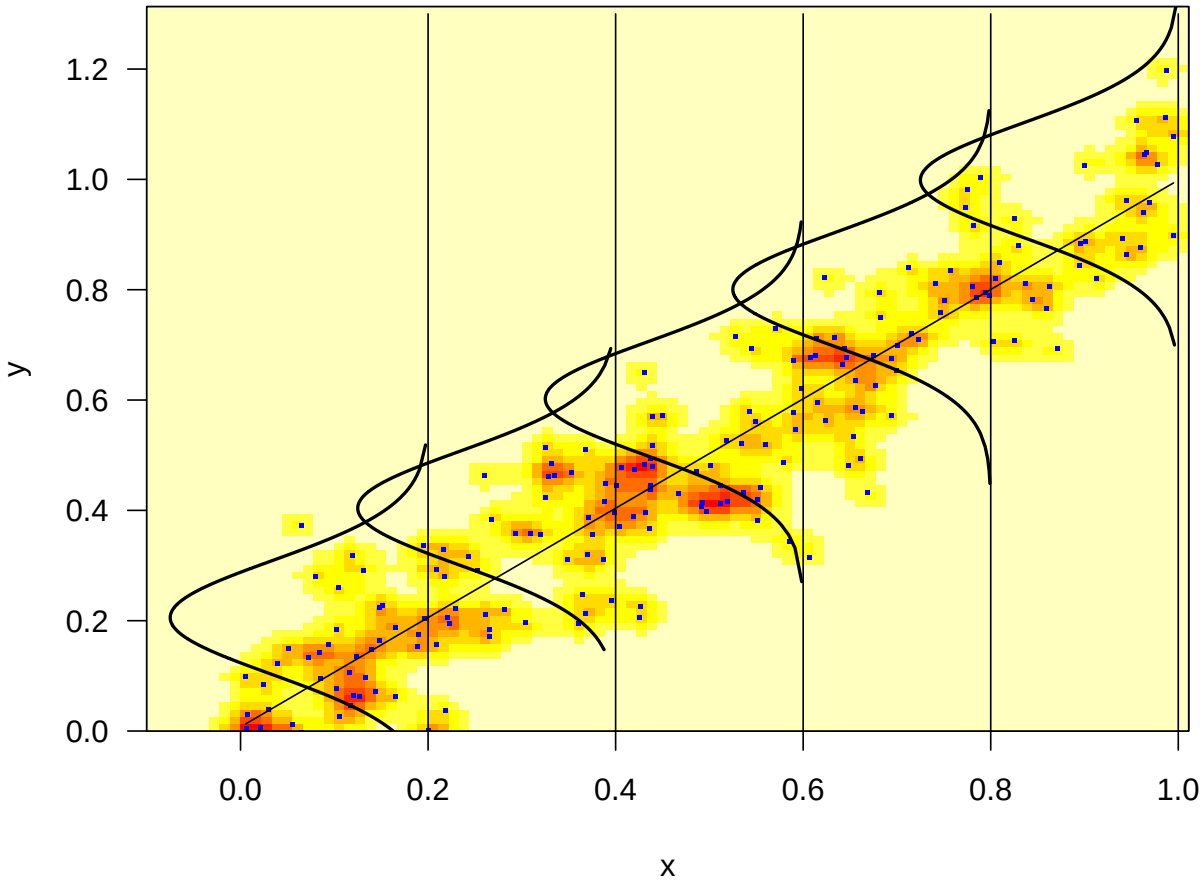

Figura 2: Ilustração de modelo de regressão linear com uma covariável x.

Modelo de regressão linear

- Considere y a variável resposta e $x_1, x_2, ..., x_r$ um conjunto de r covariáveis avaliados em uma amostra de tamanho n.

Modelo de regressão linear

- Considere y a variável resposta e $x_1, x_2, ..., x_r$ um conjunto de r covariáveis avaliados em uma amostra de tamanho n.

- O modelo de regressão linear fica definido por:

Modelo de regressão linear

- Considere y a variável resposta e $x_1, x_2, ..., x_r$ um conjunto de r covariáveis avaliados em uma amostra de tamanho n.

- O modelo de regressão linear fica definido por:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_r x_{ir} + \epsilon_i,$$

com $\epsilon_i \overset{ind}{\sim} N(0, \sigma^2)$, para $i = 1, 2, ..., n$.

Modelo de regressão linear

- O modelo de regressão linear pode ser representado na forma matricial por:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

onde $\boldsymbol{y} = (y_1, ..., y_n)'$ é o vetor de respostas, $\boldsymbol{X}$ é a matriz do modelo $n \times p$ $(p = r + 1)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_r)'$ é o vetor de parâmetros de regressão e $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)'$ o vetor de erros.

Modelo de regressão linear

- O modelo de regressão linear pode ser representado na forma matricial por:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

onde $\boldsymbol{y} = (y_1, ..., y_n)'$ é o vetor de respostas, $\boldsymbol{X}$ é a matriz do modelo $n \times p$ $(p = r + 1)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_r)'$ é o vetor de parâmetros de regressão e $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)'$ o vetor de erros.

- Uma forma equivalente (e mais flexível) de representar o modelo de regressão linear é a seguinte:

$$y|\boldsymbol{x} \overset{ind}{\sim} N(\mu_{\boldsymbol{x}}, \sigma^2),$$

onde $\mu_{\boldsymbol{x}} = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r$.

Modelo de regressão linear

- O método de mínimos quadrados é usado na estimação dos parâmetros do modelo ($\beta's$).

Modelo de regressão linear

- O método de mínimos quadrados é usado na estimação dos parâmetros do modelo ($\beta's$).

- O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é o vetor $\hat{\boldsymbol{\beta}}$ que minimiza a soma de quadrados dos erros, dada por:

Modelo de regressão linear

- O método de mínimos quadrados é usado na estimação dos parâmetros do modelo ($\beta's$).

- O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é o vetor $\hat{\boldsymbol{\beta}}$ que minimiza a soma de quadrados dos erros, dada por:

$$SQE(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Modelo de regressão linear

- O método de mínimos quadrados é usado na estimação dos parâmetros do modelo ($\beta's$).

- O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é o vetor $\hat{\boldsymbol{\beta}}$ que minimiza a soma de quadrados dos erros, dada por:

$$SQE(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

- O vetor $\hat{\boldsymbol{\beta}}$ pode ser obtido de forma analítica, resultando em:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}.$$

Modelo de regressão linear

- Sob a especificação do modelo de regressão linear, o estimador de mínimos quadrados é também o estimador de máxima verossimilhança de $\boldsymbol{\beta}$.

Modelo de regressão linear

- Sob a especificação do modelo de regressão linear, o estimador de mínimos quadrados é também o estimador de máxima verossimilhança de $\boldsymbol{\beta}$.

- O estimador de máxima verossimilhança de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{SQ_{Res}}{n} = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}$$

Modelo de regressão linear

- Sob a especificação do modelo de regressão linear, o estimador de mínimos quadrados é também o estimador de máxima verossimilhança de $\boldsymbol{\beta}$.

- O estimador de máxima verossimilhança de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{SQ_{Res}}{n} = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}$$

Modelo de regressão linear

- Voltando à análise dos dados de aluguéis de imóveis, o seguinte modelo de regressão linear é proposto:

$$y|\boldsymbol{x} \sim Normal(\mu_{\boldsymbol{x}}, \sigma^2),$$

Modelo de regressão linear

- Voltando à análise dos dados de aluguéis de imóveis, o seguinte modelo de regressão linear é proposto:

$$y|\boldsymbol{x} \sim Normal(\mu_{\boldsymbol{x}}, \sigma^2),$$

em que

$$\begin{aligned}\mu_{\boldsymbol{x}} =& \beta_0 + \beta_1 \times Fl + \beta_2 \times A + \beta_3 \times I(H = 1) \\ &+ \beta_4 \times I(loc = 2) + \beta_5 \times I(loc = 3),\end{aligned}$$

sendo $I(\cdot)$ é a função indicadora, tal que, por exemplo, $I(H = 1) = 1$ para os imóveis sem aquecimento central e $I(H = 1) = 0$ para os imóveis com aquecimento central.

Modelo de regressão linear

- O modelo de regressão linear ajustado tem a seguinte expressão:

$$\hat{\mu}_{\boldsymbol{x}} = -2775.04 + 8.84 \times Fl + 1.48 \times A - 204.76 \times I(H = 1) + 134.05 \times I(loc = 2) + 209.58 \times I(loc = 3).$$

Modelo de regressão linear

- O modelo de regressão linear ajustado tem a seguinte expressão:

$$\hat{\mu}_{\boldsymbol{x}} = -2775.04 + 8.84 \times Fl + 1.48 \times A - 204.76 \times I(H = 1) + 134.05 \times I(loc = 2) + 209.58 \times I(loc = 3).$$

- Além disso:

$$log(\hat{\sigma}) = 5.73165,$$

tal que $\hat{\sigma} = 308.48$.

Modelo linear generalizado

Modelo linear generalizado

- Modelos lineares generalizados configuram extensões dos modelos de regressão linear, tendo como diferenciais:

Modelo linear generalizado

- Modelos lineares generalizados configuram extensões dos modelos de regressão linear, tendo como diferenciais:
    - Permitem modelar respostas com distribuição pertencente à família exponencial;

Modelo linear generalizado

- Modelos lineares generalizados configuram extensões dos modelos de regressão linear, tendo como diferenciais:
  - Permitem modelar respostas com distribuição pertencente à família exponencial;
  - A relação entre a média de y e as covariáveis é linearizada por meio de uma função de ligação monotônica $g(\cdot)$;

Modelo linear generalizado

- Modelos lineares generalizados configuram extensões dos modelos de regressão linear, tendo como diferenciais:
  - Permitem modelar respostas com distribuição pertencente à família exponencial;
  - A relação entre a média de y e as covariáveis é linearizada por meio de uma função de ligação monotônica $g(\cdot)$;
  - A estimação dos parâmetros do modelo se dá por um algoritmo de mínimos quadrados ponderados iterativamente.

Modelo linear generalizado

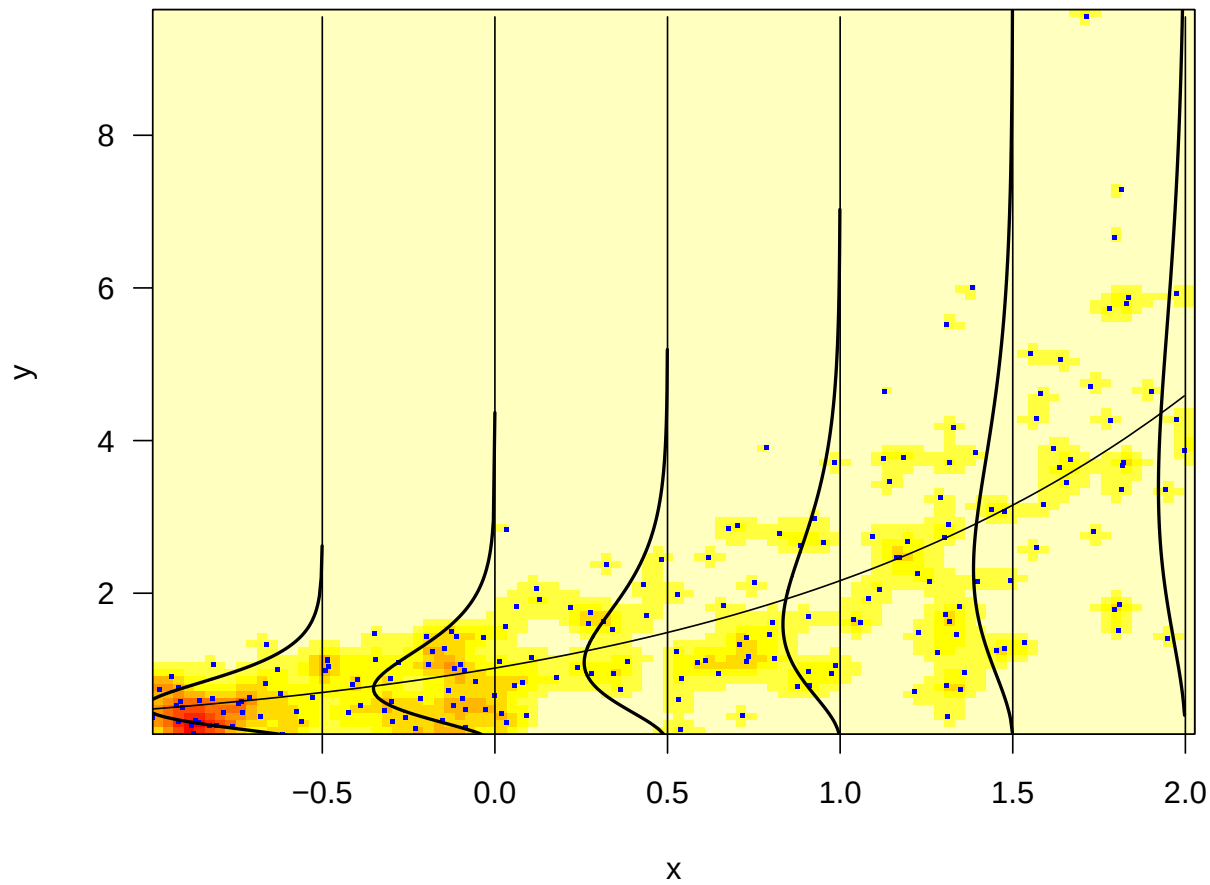

Figura 3: Ilustração de modelo de regressão com resposta gama e função de ligação logarítmica.

Modelo linear generalizado

- Uma variável aleatória y tem distribuição pertencente à família exponencial se sua função (densidade) de probabilidade puder ser expressa na seguinte forma:

$$f(y;\mu,\phi) = exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\},$$

em que θ e ϕ são parâmetros canônico e de dispersão, respectivamente.

Modelo linear generalizado

- Uma variável aleatória y tem distribuição pertencente à família exponencial se sua função (densidade) de probabilidade puder ser expressa na seguinte forma:

$$f(y;\mu,\phi) = exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\},$$

em que θ e ϕ são parâmetros canônico e de dispersão, respectivamente.

- A média e a variância de y são dadas, repsectivamente, por $E(y) = b'(\theta)$ e $Var(y) = \phi V(\mu)$, onde $V(\mu) = b''[\theta(\mu)]$ é a chamada *função de variância*.

Modelo linear generalizado

- Dentre as principais distribuições contempladas pela teoria de MLG estão:

Modelo linear generalizado

- Dentre as principais distribuições contempladas pela teoria de MLG estão:
    - Normal: $V(\mu) = 1$;

Modelo linear generalizado

- Dentre as principais distribuições contempladas pela teoria de MLG estão:
  - Normal: $V(\mu) = 1$;
  - Binomial: $V(\mu) = \mu(1 - \mu)$;

Modelo linear generalizado

- Dentre as principais distribuições contempladas pela teoria de MLG estão:
  - Normal: $V(\mu) = 1$;
  - Binomial: $V(\mu) = \mu(1 - \mu)$;
  - Poisson: $V(\mu) = \mu$;

Modelo linear generalizado

- Dentre as principais distribuições contempladas pela teoria de MLG estão:
  - Normal: $V(\mu) = 1$;
  - Binomial: $V(\mu) = \mu(1 - \mu)$;
  - Poisson: $V(\mu) = \mu$;
  - Gama: $V(\mu) = \mu^2$;

Modelo linear generalizado

- Dentre as principais distribuições contempladas pela teoria de MLG estão:
  - Normal: $V(\mu) = 1$;
  - Binomial: $V(\mu) = \mu(1 - \mu)$;
  - Poisson: $V(\mu) = \mu$;
  - Gama: $V(\mu) = \mu^2$;
  - Normal inversa: $V(\mu) = \mu^3$.

Modelo linear generalizado

- Um modelo linear generalizado pode ser representado, genericamente, da seguinte forma:

$$y|\boldsymbol{x} \overset{ind}{\sim} f(\mu_{\boldsymbol{x}}, \phi),$$

em que $f(\cdot)$ denota uma particular distribuição da família exponencial, ϕ é o parâmetro de dispersão e

$$g(\mu_{\boldsymbol{x}}) = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r.$$

Modelo linear generalizado

- Na aplicação dos dados sobre os valores de aluguel de imóveis, vamos considerar a distribuição gama, que é uma alternativa para a modelagem de dados contínuos assimétricos.

Modelo linear generalizado

- Na aplicação dos dados sobre os valores de aluguel de imóveis, vamos considerar a distribuição gama, que é uma alternativa para a modelagem de dados contínuos assimétricos.

- Uma variável aleatória com distribuição gama de média μ e parâmetro de dispersão ϕ tem a função densidade de probabilidade dada por:

$$f(y;\mu,\phi) = \frac{y^{\frac{1}{\phi}-1} exp\left(-\frac{y}{\phi\mu}\right)}{(\phi\mu)^{1/\phi}\,\Gamma(1/\phi)}, \quad y>0, \mu>0, \phi>0.$$

Modelo linear generalizado

- Na aplicação dos dados sobre os valores de aluguel de imóveis, vamos considerar a distribuição gama, que é uma alternativa para a modelagem de dados contínuos assimétricos.

- Uma variável aleatória com distribuição gama de média μ e parâmetro de dispersão ϕ tem a função densidade de probabilidade dada por:

$$f(y;\mu,\phi) = \frac{y^{\frac{1}{\phi}-1} exp\left(-\frac{y}{\phi\mu}\right)}{(\phi\mu)^{1/\phi}\,\Gamma(1/\phi)}, \quad y>0, \mu>0, \phi>0.$$

- No pacote `gamlss` a distribuição gama é especificada pela média μ e por um parâmetro de escala $\sigma = \sqrt{\phi}$.

Modelo linear generalizado

- Como $\mu > 0$, vamos usar a função de ligação logarítmica na análise dos valores dos imóveis, de forma que o modelo a ser ajustado fica especificado por:

Modelo linear generalizado

- Como $\mu > 0$, vamos usar a função de ligação logarítmica na análise dos valores dos imóveis, de forma que o modelo a ser ajustado fica especificado por:

$$y|\boldsymbol{x} \overset{ind}{\sim} \text{gama}(\mu_{\boldsymbol{x}}, \sigma), \quad y > 0, \mu > 0, \sigma > 0,$$

Modelo linear generalizado

- Como $\mu > 0$, vamos usar a função de ligação logarítmica na análise dos valores dos imóveis, de forma que o modelo a ser ajustado fica especificado por:

$$y|\boldsymbol{x} \overset{ind}{\sim} \text{gama}(\mu_{\boldsymbol{x}}, \sigma), \quad y > 0, \mu > 0, \sigma > 0,$$

com

$$\begin{aligned}\log(\mu_{\boldsymbol{x}}) =& \beta_0 + \beta_1 \times Fl + \beta_2 \times A + \beta_3 \times I(H = 1) \\ &+ \beta_4 \times I(loc = 2) + \beta_5 \times I(loc = 3).\end{aligned}$$

Modelo linear generalizado

- O ajuste do modelo linear generalizado com resposta gama resulta em $y|\boldsymbol{x} \sim \text{gama}(\hat{\mu}_{\boldsymbol{x}}, \hat{\sigma})$, tal que:

$$\begin{aligned}\log(\hat{\mu}_{\boldsymbol{x}}) =& 2.8649 + 0.0106 \times Fl + 0.0015 \times A - 0.3001 \times I(H = 1) \\ &+ 0.1907 \times I(loc = 2) + 0.2641 \times I(loc = 3),\end{aligned}$$

Modelo linear generalizado

- O ajuste do modelo linear generalizado com resposta gama resulta em $y|\boldsymbol{x} \sim \text{gama}(\hat{\mu}_{\boldsymbol{x}}, \hat{\sigma})$, tal que:

$$\begin{aligned}\log(\hat{\mu}_{\boldsymbol{x}}) =& 2.8649 + 0.0106 \times Fl + 0.0015 \times A - 0.3001 \times I(H = 1) \\ &+ 0.1907 \times I(loc = 2) + 0.2641 \times I(loc = 3),\end{aligned}$$

com

$$log(\hat{\sigma}) = -0.9822,$$

e, consequentemente, $\hat{\sigma} = 0.3745$.

Modelo generalizado aditivo

Modelo generalizado aditivo

- Modelos aditivos são mais flexíveis do que modelos totalmente paramétricos, permitindo lidar com relações não lineares entre covariáveis e a resposta.

Modelo generalizado aditivo

- Modelos aditivos são mais flexíveis do que modelos totalmente paramétricos, permitindo lidar com relações não lineares entre covariáveis e a resposta.

- Os efeitos das covariáveis são inseridos ao preditor por meio de funções suaves, que podem ser interpretados usando gráficos apropriados.

Modelo generalizado aditivo

- Modelos aditivos são mais flexíveis do que modelos totalmente paramétricos, permitindo lidar com relações não lineares entre covariáveis e a resposta.

- Os efeitos das covariáveis são inseridos ao preditor por meio de funções suaves, que podem ser interpretados usando gráficos apropriados.

- O pacote `gamlss` oferece diferentes alternativas de funções suaves a serem usadas em modelos aditivos.

Modelo generalizado aditivo

- Modelos aditivos são mais flexíveis do que modelos totalmente paramétricos, permitindo lidar com relações não lineares entre covariáveis e a resposta.

- Os efeitos das covariáveis são inseridos ao preditor por meio de funções suaves, que podem ser interpretados usando gráficos apropriados.

- O pacote `gamlss` oferece diferentes alternativas de funções suaves a serem usadas em modelos aditivos.

- Na aplicação referente aos preços de aluguel vamos considerar a inclusão de suavizadores para os efeitos de área e de ano de construção do imóvel.

Modelo generalizado aditivo

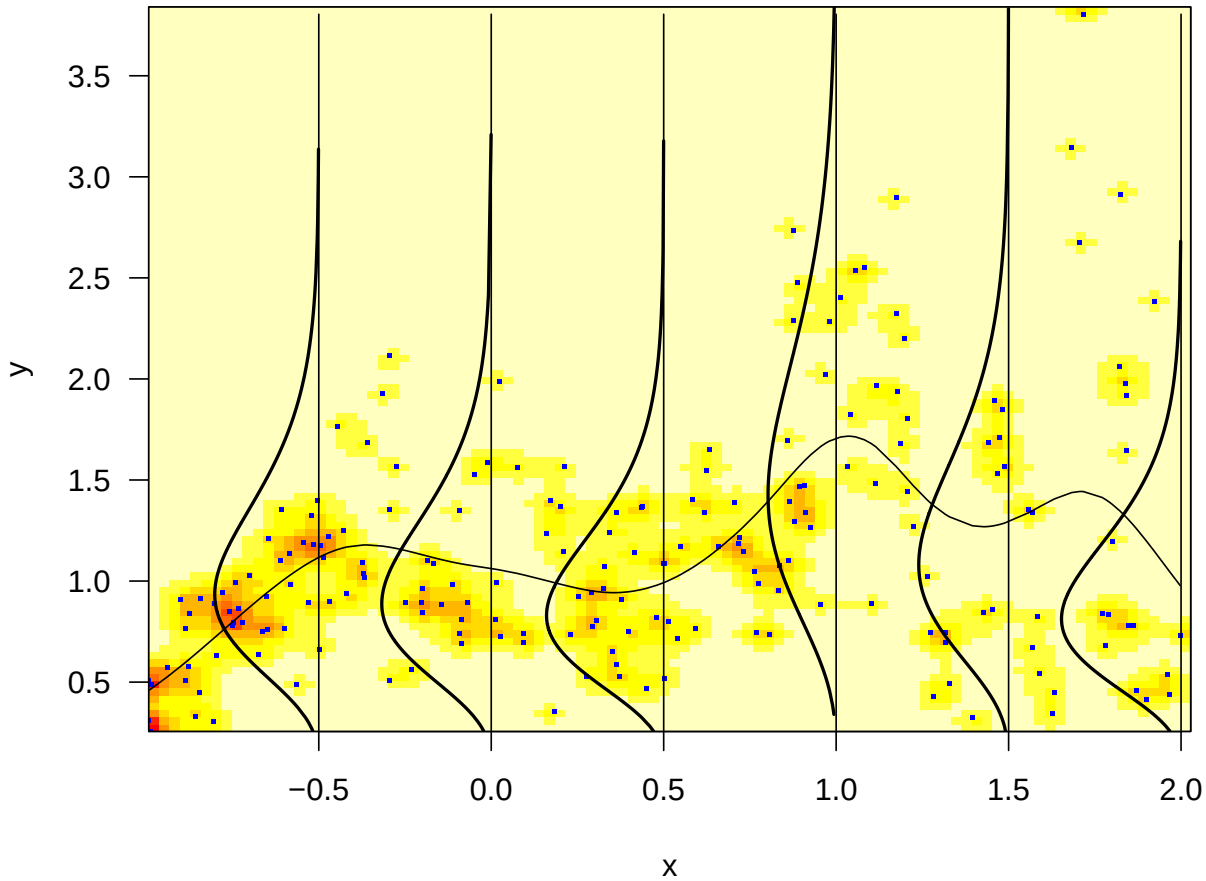

Figura 4: Modelo de regressão com resposta gama e suavizador para a covariável.

Modelo generalizado aditivo

- Modelos generalizados aditivos configuram extensões dos MLGs em que o efeito de ao menos uma das covariáveis é incorporado ao preditor linear através de uma função suavizadora (não paramétrica).

Modelo generalizado aditivo

- Modelos generalizados aditivos configuram extensões dos MLGs em que o efeito de ao menos uma das covariáveis é incorporado ao preditor linear através de uma função suavizadora (não paramétrica).

- Um modelo generalizado aditivo pode ser representado, de forma geral, por:

$$y|\boldsymbol{x} \overset{ind}{\sim} f(\mu_{\boldsymbol{x}}, \phi),$$

onde

$$g(\mu_{\boldsymbol{x}}) = \beta_0 + \beta_1 x_1 + ... + \beta_j x_j + ... + s_{j+1}(x_{j+1}) + ... + s_r(x_r),$$

em que s_k é uma função suave não paramétrica aplicada à covariável x_k, $k = j+1, ..., r$.

Modelo generalizado aditivo

- O modelo generalizado aditivo a ser ajustado aos dados de preços de aluguel, considerando resposta gama, é especificado da seguinte forma:

$$y|\boldsymbol{x} \overset{ind}{\sim} \text{gama}(\mu_{\boldsymbol{x}}, \sigma), \quad y > 0, \mu > 0, \sigma > 0,$$

Modelo generalizado aditivo

- O modelo generalizado aditivo a ser ajustado aos dados de preços de aluguel, considerando resposta gama, é especificado da seguinte forma:

$$y|\boldsymbol{x} \overset{ind}{\sim} \text{gama}(\mu_{\boldsymbol{x}}, \sigma), \quad y > 0, \mu > 0, \sigma > 0,$$

com

$$\begin{aligned}\log(\mu_{\boldsymbol{x}}) =& \beta_0 + s_1(Fl) + s_2(A) + \beta_1 \times I(H = 1) \\ &+ \beta_2 \times I(loc = 2) + \beta_3 \times I(loc = 3),\end{aligned}$$

em que s_1 e s_2 são suavizadores não paramétricos aplicados às variáveis Fl e A, respectivamente.

Modelo generalizado aditivo

- O modelo ajustado fica dado por:

$$\begin{aligned}\log(\hat{\mu}_{\boldsymbol{x}}) =& 3.0851 + s_1(Fl) + s_2(A) - 0.3008 \times I(H = 1) \\ &+ 0.1887 \times I(loc = 2) + 0.2720 \times I(loc = 3),\end{aligned}$$

Modelo generalizado aditivo

- O modelo ajustado fica dado por:

$$\begin{aligned}\log(\hat{\mu}_{\boldsymbol{x}}) =& 3.0851 + s_1(Fl) + s_2(A) - 0.3008 \times I(H = 1) \\ &+ 0.1887 \times I(loc = 2) + 0.2720 \times I(loc = 3),\end{aligned}$$

com

$$log(\hat{\sigma}) = -1.0019,$$

tal que $\hat{\sigma} = 0.33672$.

Modelo generalizado aditivo duplo

Modelo generalizado aditivo duplo

- Até o momento consideramos apenas modelos em que a média (parâmetro de locação) da distribuição varia conforme os valores das covariáveis.

Modelo generalizado aditivo duplo

- Até o momento consideramos apenas modelos em que a média (parâmetro de locação) da distribuição varia conforme os valores das covariáveis.

- Modelos mais gerais permitem incluir covariáveis também na modelagem de outros parâmetros da distribuição (por exemplo, para o parâmetro de escala).

Modelo generalizado aditivo duplo

- Até o momento consideramos apenas modelos em que a média (parâmetro de locação) da distribuição varia conforme os valores das covariáveis.

- Modelos mais gerais permitem incluir covariáveis também na modelagem de outros parâmetros da distribuição (por exemplo, para o parâmetro de escala).

- Para a distribuição gama, por exemplo, temos que $Var(y) = \sigma^2\mu^2$, ou seja, $\sigma = \sqrt{Var(y)}/\mu$ é o coeficiente de variação de y.

Modelo generalizado aditivo duplo

- Até o momento consideramos apenas modelos em que a média (parâmetro de locação) da distribuição varia conforme os valores das covariáveis.

- Modelos mais gerais permitem incluir covariáveis também na modelagem de outros parâmetros da distribuição (por exemplo, para o parâmetro de escala).

- Para a distribuição gama, por exemplo, temos que $Var(y) = \sigma^2\mu^2$, ou seja, $\sigma = \sqrt{Var(y)}/\mu$ é o coeficiente de variação de y.

- Neste caso, podemos modelar também σ em função de covariáveis.

Modelo generalizado aditivo duplo

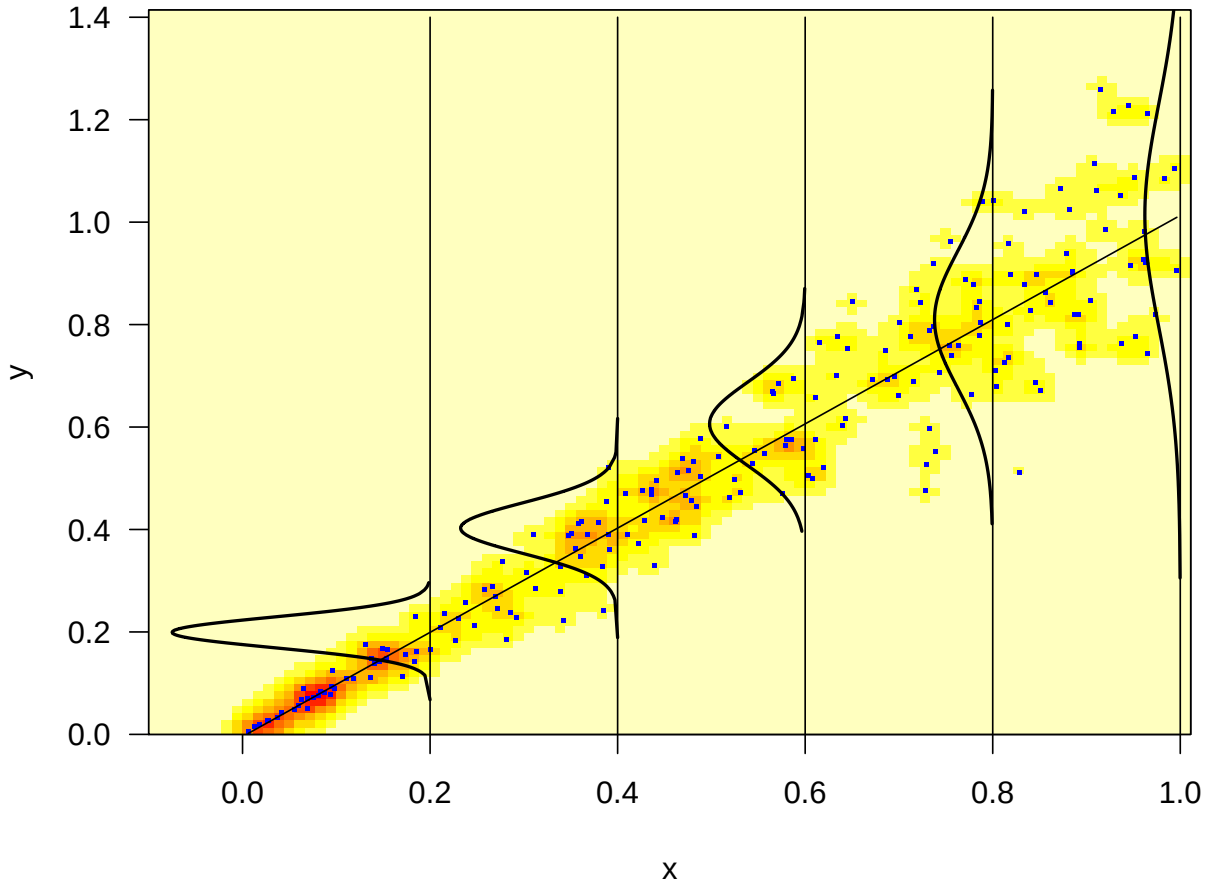

Figura 5: Distribuição normal com média e variância dependentes de uma covariável x.

Modelo generalizado aditivo duplo

- Uma formulação geral para o modelo, neste caso, seria:

$$y|\boldsymbol{x} \overset{ind}{\sim} D(\mu_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}}),$$

onde

$$g_1(\mu_{\boldsymbol{x}}) = \beta_{10} + \beta_{11}x_1 + ... + \beta_{1j_1}x_{j_1} + ... + s_{j_1+1}(x_{j_1+1}) + ... + s_{r_1}(x_{r_1})$$
$$g_2(\sigma_{\boldsymbol{x}}) = \beta_{20} + \beta_{21}x_1 + ... + \beta_{2j_2}x_{j_2} + ... + s_{j_2+1}(x_{j_2+1}) + ... + s_{r_2}(x_{r_2}),$$

Modelo generalizado aditivo duplo

- Uma formulação geral para o modelo, neste caso, seria:

$$y|\boldsymbol{x} \overset{ind}{\sim} D(\mu_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}}),$$

onde

$$g_1(\mu_{\boldsymbol{x}}) = \beta_{10} + \beta_{11}x_1 + ... + \beta_{1j_1}x_{j_1} + ... + s_{j_1+1}(x_{j_1+1}) + ... + s_{r_1}(x_{r_1})$$
$$g_2(\sigma_{\boldsymbol{x}}) = \beta_{20} + \beta_{21}x_1 + ... + \beta_{2j_2}x_{j_2} + ... + s_{j_2+1}(x_{j_2+1}) + ... + s_{r_2}(x_{r_2}),$$

em que D representa alguma distribuição de probabilidades tal que dois de seus parâmetros (μ e σ) são funções de covariáveis.

Modelo generalizado aditivo duplo

- Vamos retomar a análise dos dados de preços de aluguel com o ajuste de modelos com resposta gama e normal inversa, inserindo covariáveis também na modelagem do parâmetro de escala.

Modelo generalizado aditivo duplo

- Vamos retomar a análise dos dados de preços de aluguel com o ajuste de modelos com resposta gama e normal inversa, inserindo covariáveis também na modelagem do parâmetro de escala.

- Em ambos os casos vamos considerar função de ligação logarítmica tanto para μ quanto para σ, produzindo os seguintes modelos:

Modelo generalizado aditivo duplo

- Vamos retomar a análise dos dados de preços de aluguel com o ajuste de modelos com resposta gama e normal inversa, inserindo covariáveis também na modelagem do parâmetro de escala.

- Em ambos os casos vamos considerar função de ligação logarítmica tanto para μ quanto para σ, produzindo os seguintes modelos:

$$\begin{aligned}\log(\mu_{\boldsymbol{x}}) =& \beta_{10} + s_{11}(Fl) + s_{12}(A) + \beta_{11} \times I(H = 1) \\ &+ \beta_{12} \times I(loc = 2) + \beta_{13} \times I(loc = 3),\end{aligned}$$

e

$$\begin{aligned}\log(\sigma_{\boldsymbol{x}}) =& \beta_{20} + s_{21}(Fl) + s_{22}(A) + \beta_{21} \times I(H = 1) \\ &+ \beta_{22} \times I(loc = 2) + \beta_{23} \times I(loc = 3).\end{aligned}$$

Modelo generalizado aditivo duplo

- O modelo ajustado com resposta gama (que produziu menor AIC do que com resposta normal inversa) é o seguinte:

$$\begin{aligned}\log(\hat{\mu}_{\boldsymbol{x}}) =& 2.8844 + s_{11}(Fl) + s_{12}(A) - 0.2918 \times I(H = 1) \\ &+ 0.1938 \times I(loc = 2) + 0.2734 \times I(loc = 3),\end{aligned}$$

com

$$\begin{aligned}log(\hat{\sigma}_{\boldsymbol{x}}) =& 5.9225 + s_{21}(Fl) + s_{22}(A) + 0.0659 \times I(H = 1) \\ &- 0.1166 \times I(loc = 2) - 0.1702 \times I(loc = 3).\end{aligned}$$

Modelo generalizado aditivo para locação, escala e forma

Modelo generalizado aditivo para locação, escala e forma

- Os modelos considerados anteriormente baseiam-se na inclusão de covariáveis nos preditores dos parâmetros de locação e escala (apenas no caso do generalizado aditivo duplo) da distribução.

Modelo generalizado aditivo para locação, escala e forma

- Os modelos considerados anteriormente baseiam-se na inclusão de covariáveis nos preditores dos parâmetros de locação e escala (apenas no caso do generalizado aditivo duplo) da distribução.

- Em GAMLSS, distribuições com até quatro parâmetros podem ter cada um deles modelados em função de covariáveis.

Modelo generalizado aditivo para locação, escala e forma

- Os modelos considerados anteriormente baseiam-se na inclusão de covariáveis nos preditores dos parâmetros de locação e escala (apenas no caso do generalizado aditivo duplo) da distribução.

- Em GAMLSS, distribuições com até quatro parâmetros podem ter cada um deles modelados em função de covariáveis.

- Desta forma, propriedades como assimetria, curtose, excesso ou escassez de zeros podem ser explicadas em função de covariáveis.

Modelo generalizado aditivo para locação, escala e forma

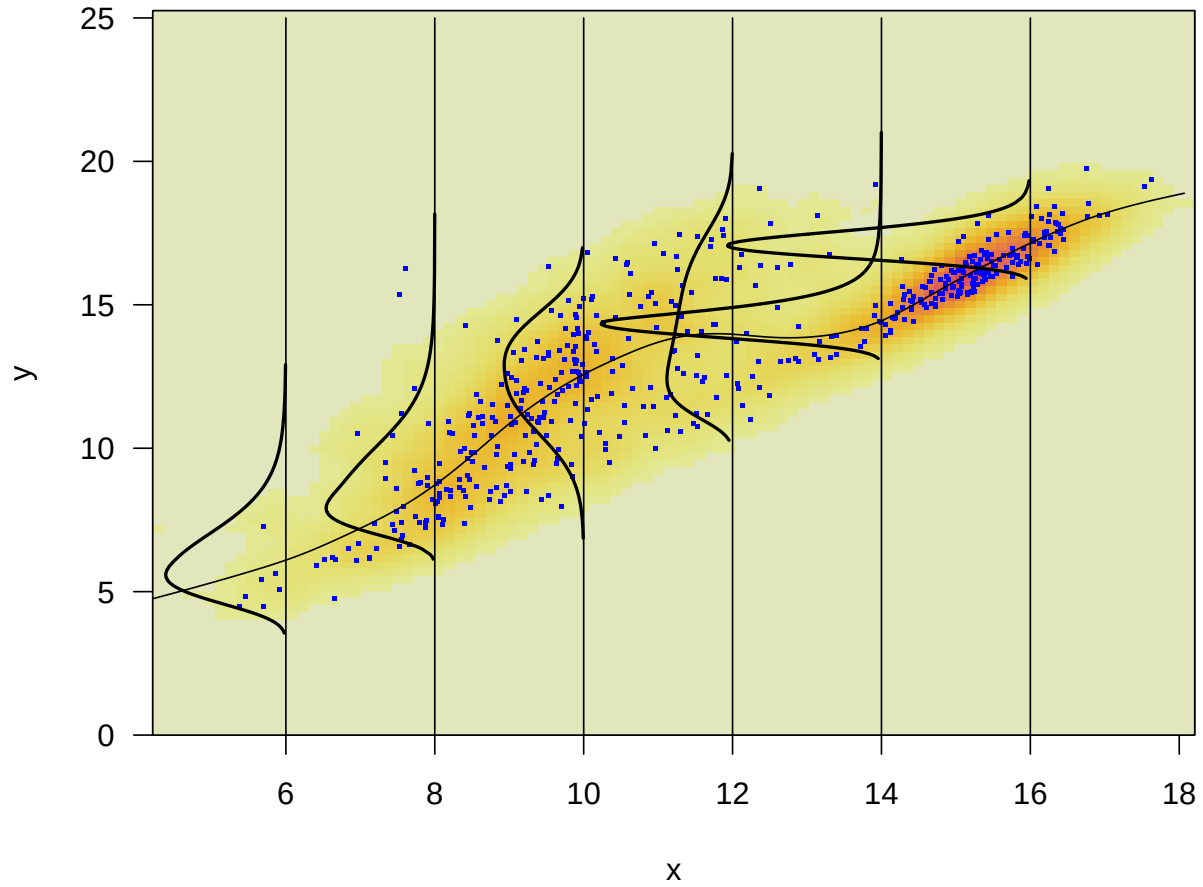

Figura 6: Distribuição com locação, dispersão e forma dependentes de x.

Modelo generalizado aditivo para locação, escala e forma

- Um modelo generalizado aditivo para locação, escala e forma é definido, de forma geral, por:

$$y|\boldsymbol{x} \overset{ind}{\sim} D(\mu_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}}, \nu_{\boldsymbol{x}}, \tau_{\boldsymbol{x}}),$$

onde

$$\begin{aligned}
g_1(\mu_{\boldsymbol{x}}) &= \beta_{10} + \beta_{11}x_1 + ... + \beta_{1j_1}x_{j_1} + ... + s_{j_1+1}(x_{j_1+1}) + ... + s_{r_1}(x_{r_1}) \\
g_2(\sigma_{\boldsymbol{x}}) &= \beta_{20} + \beta_{21}x_1 + ... + \beta_{2j_2}x_{j_2} + ... + s_{j_2+1}(x_{j_2+1}) + ... + s_{r_2}(x_{r_2}) \\
g_3(\nu_{\boldsymbol{x}}) &= \beta_{30} + \beta_{31}x_1 + ... + \beta_{3j_3}x_{j_3} + ... + s_{j_3+1}(x_{j_3+1}) + ... + s_{r_3}(x_{r_3}) \\
g_4(\tau_{\boldsymbol{x}}) &= \beta_{40} + \beta_{41}x_1 + ... + \beta_{4j_4}x_{j_4} + ... + s_{j_4+1}(x_{j_4+1}) + ... + s_{r_4}(x_{r_4})
\end{aligned},$$

em que $D(\mu, \sigma, \nu, \tau)$ é uma distribuição de quatro parâmetros e ν e τ são parâmetros de forma.

Modelo generalizado aditivo para locação, escala e forma

- O pacote `gamlss` dispõe de dezenas de distribuições implementadas. Além disso:

Modelo generalizado aditivo para locação, escala e forma

- O pacote `gamlss` dispõe de dezenas de distribuições implementadas. Além disso:
  - É possível ao usuário implementar novas distribuições;

Modelo generalizado aditivo para locação, escala e forma

- O pacote `gamlss` dispõe de dezenas de distribuições implementadas. Além disso:
  - É possível ao usuário implementar novas distribuições;
  - Versões truncadas ou censuradas podem ser definidas a partir das distribuições originais;

Modelo generalizado aditivo para locação, escala e forma

- O pacote `gamlss` dispõe de dezenas de distribuições implementadas. Além disso:
  - É possível ao usuário implementar novas distribuições;
  - Versões truncadas ou censuradas podem ser definidas a partir das distribuições originais;
  - Pode-se criar novas distribuições a partir de misturas das distribuições originais;

Modelo generalizado aditivo para locação, escala e forma

- O pacote `gamlss` dispõe de dezenas de distribuições implementadas. Além disso:
  - É possível ao usuário implementar novas distribuições;
  - Versões truncadas ou censuradas podem ser definidas a partir das distribuições originais;
  - Pode-se criar novas distribuições a partir de misturas das distribuições originais;
  - Distribuições discretas podem ser criadas a partir de distribuições originalmente contínuas;

Modelo generalizado aditivo para locação, escala e forma

- O pacote `gamlss` dispõe de dezenas de distribuições implementadas. Além disso:
  - É possível ao usuário implementar novas distribuições;
  - Versões truncadas ou censuradas podem ser definidas a partir das distribuições originais;
  - Pode-se criar novas distribuições a partir de misturas das distribuições originais;
  - Distribuições discretas podem ser criadas a partir de distribuições originalmente contínuas;
  - Distribuições com suporte nos intervalos $(0, \infty)$ ou $(0, 1)$ podem ser geradas a partir de variáveis com suporte no intervalo $(-\infty, \infty)$.

Modelo generalizado aditivo para locação, escala e forma

- Termos aditivos podem ser adicionados ao modelo de diferentes formas, usando, por exemplo:

Modelo generalizado aditivo para locação, escala e forma

- Termos aditivos podem ser adicionados ao modelo de diferentes formas, usando, por exemplo:
    - Penalized B-splines (P-splines);
    - Monotone P-splines;
    - Cycle P-splines;
    - Varying coefficient P-splynes;
    - Cubic smoothing P-splynes;
    - Loess curve fitting;
    - Fractional polynomials;
    - Random effects;
    - Ridge regression;
    - Nonlinear parametric fits.

Modelo generalizado aditivo para locação, escala e forma

- A estimação dos parâmetros em gamlss baseia-se no método de máxima verossimilhança penalizada.

Modelo generalizado aditivo para locação, escala e forma

- A estimação dos parâmetros em gamlss baseia-se no método de máxima verossimilhança penalizada.

- Dois algoritmos estão implementados em `gamlss` para o ajuste dos modelos: CG (Cole e Green) e RS (Rigby e Stasinopoulos), que podem ainda ser usados de forma combinada.

Modelo generalizado aditivo para locação, escala e forma

- Dando sequência à análise dos dados de preços de aluguel, vamos considerar, como alternativa à distribuição gama, a distribuição Box-Cox Cole e Green (BCCGo).

Modelo generalizado aditivo para locação, escala e forma

- Dando sequência à análise dos dados de preços de aluguel, vamos considerar, como alternativa à distribuição gama, a distribuição Box-Cox Cole e Green (BCCGo).

- A distribuição BCCGo tem três parâmetros (μ, σ e τ), sendo τ um parâmetro de forma.

Modelo generalizado aditivo para locação, escala e forma

- Dando sequência à análise dos dados de preços de aluguel, vamos considerar, como alternativa à distribuição gama, a distribuição Box-Cox Cole e Green (BCCGo).

- A distribuição BCCGo tem três parâmetros (μ, σ e τ), sendo τ um parâmetro de forma.

- Cada um dos parâmetros pode ou não ser modelado em função de covariáveis. Além disso, diferentes conjuntos de covariáveis podem ser incluídos para cada parâmetro.

Modelo generalizado aditivo para locação, escala e forma

- Dando sequência à análise dos dados de preços de aluguel, vamos considerar, como alternativa à distribuição gama, a distribuição Box-Cox Cole e Green (BCCGo).

- A distribuição BCCGo tem três parâmetros (μ, σ e τ), sendo τ um parâmetro de forma.

- Cada um dos parâmetros pode ou não ser modelado em função de covariáveis. Além disso, diferentes conjuntos de covariáveis podem ser incluídos para cada parâmetro.

- Modelos com diferentes especificações (distribuições, termos aditivos, covariáveis...) podem ter seus ajustes comparados usando o critério de informação de Akaike (AIC) ou Akaike generalizado (GAIC), dentre outros.

Modelo generalizado aditivo para locação, escala e forma

- No modelo especificado para esta análise, todas as covariáveis foram incluídas na modelagem dos três parâmetros (μ, σ e ν).

Modelo generalizado aditivo para locação, escala e forma

- No modelo especificado para esta análise, todas as covariáveis foram incluídas na modelagem dos três parâmetros (μ, σ e ν).

- Como funções de ligação para cada parâmetro adotou-se o padrão do pacote `gamlss` para a BCCGo (log, log e identidade para μ, σ e ν, respectivamente).

Modelo generalizado aditivo para locação, escala e forma

- No modelo especificado para esta análise, todas as covariáveis foram incluídas na modelagem dos três parâmetros (μ, σ e ν).

- Como funções de ligação para cada parâmetro adotou-se o padrão do pacote `gamlss` para a BCCGo (log, log e identidade para μ, σ e ν, respectivamente).

- Novamente, funções suaves foram incorporadas para as variáveis Fl e A.

Modelo generalizado aditivo para locação, escala e forma

- O modelo proposto é o seguinte:

$$y|\boldsymbol{x} \overset{ind}{\sim} BCCGo(\mu_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}}, \nu_{\boldsymbol{x}}),$$

Modelo generalizado aditivo para locação, escala e forma

- O modelo proposto é o seguinte:

$$y|\boldsymbol{x} \overset{ind}{\sim} BCCGo(\mu_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}}, \nu_{\boldsymbol{x}}),$$

onde

$$\begin{aligned}\log(\mu_{\boldsymbol{x}}) =& \beta_{10} + s_{11}(Fl) + s_{12}(A) + \beta_{11} \times I(H = 1) \\ &+ \beta_{12} \times I(loc = 2) + \beta_{13} \times I(loc = 3),\end{aligned}$$

Modelo generalizado aditivo para locação, escala e forma

- O modelo proposto é o seguinte:

$$y|\boldsymbol{x} \overset{ind}{\sim} BCCGo(\mu_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}}, \nu_{\boldsymbol{x}}),$$

onde

$$\begin{aligned}\log(\mu_{\boldsymbol{x}}) =& \beta_{10} + s_{11}(Fl) + s_{12}(A) + \beta_{11} \times I(H = 1) \\ &+ \beta_{12} \times I(loc = 2) + \beta_{13} \times I(loc = 3),\end{aligned}$$

$$\begin{aligned}\log(\sigma_{\boldsymbol{x}}) =& \beta_{20} + s_{21}(Fl) + s_{22}(A) + \beta_{21} \times I(H = 1) \\ &+ \beta_{22} \times I(loc = 2) + \beta_{23} \times I(loc = 3),\end{aligned}$$

Modelo generalizado aditivo para locação, escala e forma

- O modelo proposto é o seguinte:

$$y|\boldsymbol{x} \overset{ind}{\sim} BCCGo(\mu_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}}, \nu_{\boldsymbol{x}}),$$

onde

$$\begin{aligned}\log(\mu_{\boldsymbol{x}}) =& \beta_{10} + s_{11}(Fl) + s_{12}(A) + \beta_{11} \times I(H = 1) \\ &+ \beta_{12} \times I(loc = 2) + \beta_{13} \times I(loc = 3),\end{aligned}$$

$$\begin{aligned}\log(\sigma_{\boldsymbol{x}}) =& \beta_{20} + s_{21}(Fl) + s_{22}(A) + \beta_{21} \times I(H = 1) \\ &+ \beta_{22} \times I(loc = 2) + \beta_{23} \times I(loc = 3),\end{aligned}$$

$$\begin{aligned}\nu_{\boldsymbol{x}} =& \beta_{30} + s_{31}(Fl) + s_{32}(A) + \beta_{31} \times I(H = 1) \\ &+ \beta_{32} \times I(loc = 2) + \beta_{33} \times I(loc = 3).\end{aligned}$$

Modelo generalizado aditivo para locação, escala e forma

- Como resultado temos o seguinte modelo ajustado:

$$\begin{aligned}\log(\hat{\mu}_{\boldsymbol{x}}) =& 2.0285 + s_{11}(Fl) + s_{12}(A) - 0.3213 \times I(H = 1) \\ &+ 0.1853 \times I(loc = 2) + 0.2742 \times I(loc = 3),\end{aligned}$$

Modelo generalizado aditivo para locação, escala e forma

- Como resultado temos o seguinte modelo ajustado:

$$\begin{aligned}\log(\hat{\mu}_{\boldsymbol{x}}) =& 2.0285 + s_{11}(Fl) + s_{12}(A) - 0.3213 \times I(H = 1) \\ &+ 0.1853 \times I(loc = 2) + 0.2742 \times I(loc = 3),\end{aligned}$$

$$\begin{aligned}\log(\hat{\sigma}_{\boldsymbol{x}}) =& 6.6534 + s_{21}(Fl) + s_{22}(A) + 0.0819 \times I(H = 1) \\ &- 0.0851 \times I(loc = 2) - 0.1410 \times I(loc = 3),\end{aligned}$$

Modelo generalizado aditivo para locação, escala e forma

- Como resultado temos o seguinte modelo ajustado:

$$\begin{aligned}\log(\hat{\mu}_{\boldsymbol{x}}) =& 2.0285 + s_{11}(Fl) + s_{12}(A) - 0.3213 \times I(H = 1) \\ &+ 0.1853 \times I(loc = 2) + 0.2742 \times I(loc = 3),\end{aligned}$$

$$\begin{aligned}\log(\hat{\sigma}_{\boldsymbol{x}}) =& 6.6534 + s_{21}(Fl) + s_{22}(A) + 0.0819 \times I(H = 1) \\ &- 0.0851 \times I(loc = 2) - 0.1410 \times I(loc = 3),\end{aligned}$$

$$\begin{aligned}\hat{\nu}_{\boldsymbol{x}} =& -3.1601 + s_{31}(Fl) + s_{32}(A) - 0.2866 \times I(H = 1) \\ &- 0.1711 \times I(loc = 2) - 0.0845 \times I(loc = 3).\end{aligned}$$

Resumindo

Por que GAMLSS?

- GAMLSS configura uma metodologia flexível para análise de regressão;

Por que GAMLSS?

- GAMLSS configura uma metodologia flexível para análise de regressão;
- Permite especificar diversas distribuições de probabilidades para a variável resposta;

Por que GAMLSS?

- GAMLSS configura uma metodologia flexível para análise de regressão;
- Permite especificar diversas distribuições de probabilidades para a variável resposta;
- Todos os parâmetros da distribuição podem ser modelados em função de covariáveis;

Por que GAMLSS?

- GAMLSS configura uma metodologia flexível para análise de regressão;
- Permite especificar diversas distribuições de probabilidades para a variável resposta;
- Todos os parâmetros da distribuição podem ser modelados em função de covariáveis;
- Diferentes tipos de termos aditivos podem ser incluídos nos preditores de cada parâmetro;

Por que GAMLSS?

- GAMLSS configura uma metodologia flexível para análise de regressão;
- Permite especificar diversas distribuições de probabilidades para a variável resposta;
- Todos os parâmetros da distribuição podem ser modelados em função de covariáveis;
- Diferentes tipos de termos aditivos podem ser incluídos nos preditores de cada parâmetro;
- GAMLSS estende diversas outras metodologias (como LM, GLM e GAM) permitindo modelar dados com super-dispersão, excesso de zeros, diferentes níveis de assimetria e curtose...

Próximos passos

Próximos passos

- Inferência em gamlss

Próximos passos

- Inferência em gamlss
    - Estimação por máxima verossimilhança penalizada;

Próximos passos

- Inferência em gamlss
  - Estimação por máxima verossimilhança penalizada;
  - Breve apresentação dos algoritmos de estimação;

Próximos passos

- Inferência em gamlss
  - Estimação por máxima verossimilhança penalizada;
  - Breve apresentação dos algoritmos de estimação;
  - Intervalos de confiança e testes de hipóteses;

Próximos passos

- Inferência em gamlss
  - Estimação por máxima verossimilhança penalizada;
  - Breve apresentação dos algoritmos de estimação;
  - Intervalos de confiança e testes de hipóteses;
  - Predição.