Seleção e comparação de modelos

Prof. Caio Azevedo

(grande parte do material apresentado foi extraído do livro Modelos de regressão com apoio computacional do Prof. Gilberto A. Paula)

http : //*www.ime.usp.br*/∼*giapaula*/*texto*_2013.*pdf*

Introdução

- Vimos como verificar se um determinado modelo (normal-linear-homocedástico) se ajusta adequadamente aos dados.
- Uma outra questão de interesse surge quando se dispõe de diversos modelos (que se ajustam adequadamente aos dados) e respondem às perguntas de interesse, e queremos escolher um como o "mais apropriado".
- Há diversas técnicas disponíveis para este fim.
- Veremos técnicas baseadas em testes de hipótese e comparação de estatísticas de qualidade de ajuste.

Teste da razão de verossimilhanças

- Sejam M_1 e M_2 dois modelos, em que M_1 está encaixado em M_2, ou seja, o modelo M_1 é um caso particular de M_2.
- Por exemplo, M_1 é um modelo linear obtido de M_2, o qual é um modelo quadrático.
- Neste caso temos que

  H_0 : **o modelo** M_1 **é preferível ao modelo** M_2 **vs** H_1 : **o modelo** M_2 **é preferível ao modelo** M_1.

Teste da razão de verossimilhanças (cont.)

- Seja $\widehat{\boldsymbol{\theta}}_i$ o estimador de máxima verossimilhança obtido sob o modelo i e $\widetilde{\boldsymbol{\theta}}_i$ sua respectiva estimativa.
- Denote por $L_i(\widehat{\boldsymbol{\theta}}_i)$ e $l_i(\widehat{\boldsymbol{\theta}}_i)$ o máximo da verossimilhança e da log-verossimilhança do modelo i, respectivamente, avaliados nos respectivos estimadores de MV, enquanto que $L_i(\widetilde{\boldsymbol{\theta}}_i)$ e $l_i(\widetilde{\boldsymbol{\theta}}_i)$ são os respectivos máximos avaliados nas estimativas de MV.

Teste da razão de verossimilhanças (cont.)

- A estatística do TRV é dada por $\Delta = \frac{L_1(\widehat{\boldsymbol{\theta}}_1)}{L_2(\widehat{\boldsymbol{\theta}}_2)}$.
- Rejeita-se H_0 se $\Delta \leq \delta_c$, em que δ_c é um valor crítico adequado.
- Alternativamente, rejeitamos H_0 se

$$\Lambda = -2ln(\Delta) = -2\left(l_1(\widehat{\boldsymbol{\theta}}_1) - l_2(\widehat{\boldsymbol{\theta}}_2)\right) \geq \lambda_c,$$

em que $P(Q \geq \lambda_c) = \alpha$, $Q \approx \chi^2_{(\gamma)}$ e

$\gamma =$ número de parâmetros do modelo M_2 - número de parâmetros do modelo M_1.

- Nesse caso, $p - valor \approx P(Q \geq \lambda | H_0)$, em que λ é o valor observado da estatística Λ e $Q \sim \chi^2_{(\gamma)}$. Assim, rejeita-e H_0 se $p - valor \leq \alpha$.

Estatísticas de comparação de modelos

- O TRV é apropriado na comparação somente de modelos encaixados (o modelo com menor número de parâmetros é um caso particular do modelo com maior número de parâmetros).
- Além disso, ele não leva em consideração (diretamente) o número de parâmetros do modelo (somente na distribuição da estatística).
- Existem várias alternativas, em termos de estatísticas para comparar modelos, que "penalizam" a verossimilhança em relação ao número de parâmetros, tamanho da amostra entre outros fatores.
- Veremos o AIC e o BIC.

Estatísticas de comparação de modelos (cont.)

- O AIC e BIC, para o i-ésimo modelo, são dados, respectivamente, por:

$$\begin{aligned} AIC_i &= -2l_i(\widetilde{\theta}_i) + 2k \\ BIC_i &= -2l_i(\widetilde{\theta}_i) + k\ln(n) \end{aligned}$$

que $l_i(\widetilde{\theta}_i)$ denota a log-verossimilhança do i-ésimo modelo avaliada em alguma estimativa (p.e. máxima verossimilhança), k é o número de parâmetros e n é o número de observações.

- Portanto, o modelo que apresentar os menores valores, será o modelo "melhor ajustado" aos dados.

Métodos de seleção "dinâmico" ou automatizados

- Existem métodos que selecionam modelos, fixados alguns critérios, de modo "dinâmico" (automatizado).
- Veremos os métodos "forward", "backward" e "stepwise".
- Tais métodos são particularmente úteis quanto se dispões de muitas covariáveis.
- Sem perda de generalidade, vamos considerar um determinado modelo (p.e., normal linear homocedástico) tal que o preditor linear é dado por

$$\eta_{ij} = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}$$

Método "forward"

- Primeiramente, ajustamos um modelo com somente o intercepto, ou seja $\eta_{ij} = \beta_0$. Ajustamos então, para cada variável explicativa, um modelo

$$\eta_{ij} = \beta_0 + \beta_j x_{ij}, j = 1, 2, ..., p - 1$$

- Testa-se $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, j=1,2,...,p-1 (usando-se algum teste como o TRV, teste $\mathbf{C}\boldsymbol{\beta}$, ou alguma estatística de comparação de modelos). Seja P o menor nível descritivo entre os $p - 1$ testes. Se $P \leq P_E$ a variável correspondente entra no modelo (caso contrário, o processo é interrompido).

Métodos "forward" (cont.)

- Vamor supor que a variável x_1 foi escolhida. Então, no passo seguinte, ajustamos os modelos

$$\eta_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_j x_{ij}, j = 2, ..., p-1$$

- Testa-se $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, j=2,...,p-1 (usando-se algum teste como TRV, teste $\mathbf{C}\boldsymbol{\beta}$, ou alguma estatística de comparação de modelos). Seja P o menor nível descritivo entre os $p-2$ testes. Se $P \leq P_E$ a variável correspondente entra no modelo. Repetimos o procedimento até que ocorra $P > P_E$.

Método "backward"

- Primeiramente, ajustamos o seguinte modelo:

$$\eta_{ij} = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}$$

- Testa-se $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, j=1,2,...,p-1 (usando-se algum teste como o TRV, teste $\mathbf{C}\boldsymbol{\beta}$, ou alguma estatística de comparação de modelos). Seja P o maior nível descritivo entre os $p-1$ testes. Se $P > P_S$ a variável correspondente sai do modelo (caso contrário, o processo é interrompido).

Método "backward" (cont.)

- Vamos supor que x_1 tenha saído do modelo. Então ajustamos o seguinte modelo

$$\eta_{ij} = \beta_0 + \sum_{j=2}^{p-1} \beta_j x_{ij}$$

- Testa-se $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, j=2,...,p-1 (usando-se algum teste como TRV, teste $\mathbf{C}\boldsymbol{\beta}$, ou alguma estatística de comparação de modelos). Seja P o maior nível descritivo entre os $p-2$ testes. Se $P > P_S$ a variável correspondente sai do modelo. Repetimos o procedimento até que ocorra $P \leq P_S$.

Método "stepwise"

- É uma mistura dos dois procedimentos anteriores.
- Iniciamos o processo com o modelo $\eta_{ij} = \beta_0$. Após duas variáveis terem sido incluídas no modelo, verificamos se a primeira sai ou não do modelo.
- O processo continua até que nenhuma variável seja incluída ou retirada do modelo.
- Geralmente adotamos $0,15 \leq P_E, P_S \leq 0,25$. Outra possibilidade é usar $P_E = P_S = 0,20$.
- Pode-se também começar pelo modelo completo e verificar se, após a exclusão de duas variáveis, se a primeira volta ou não ao modelo.

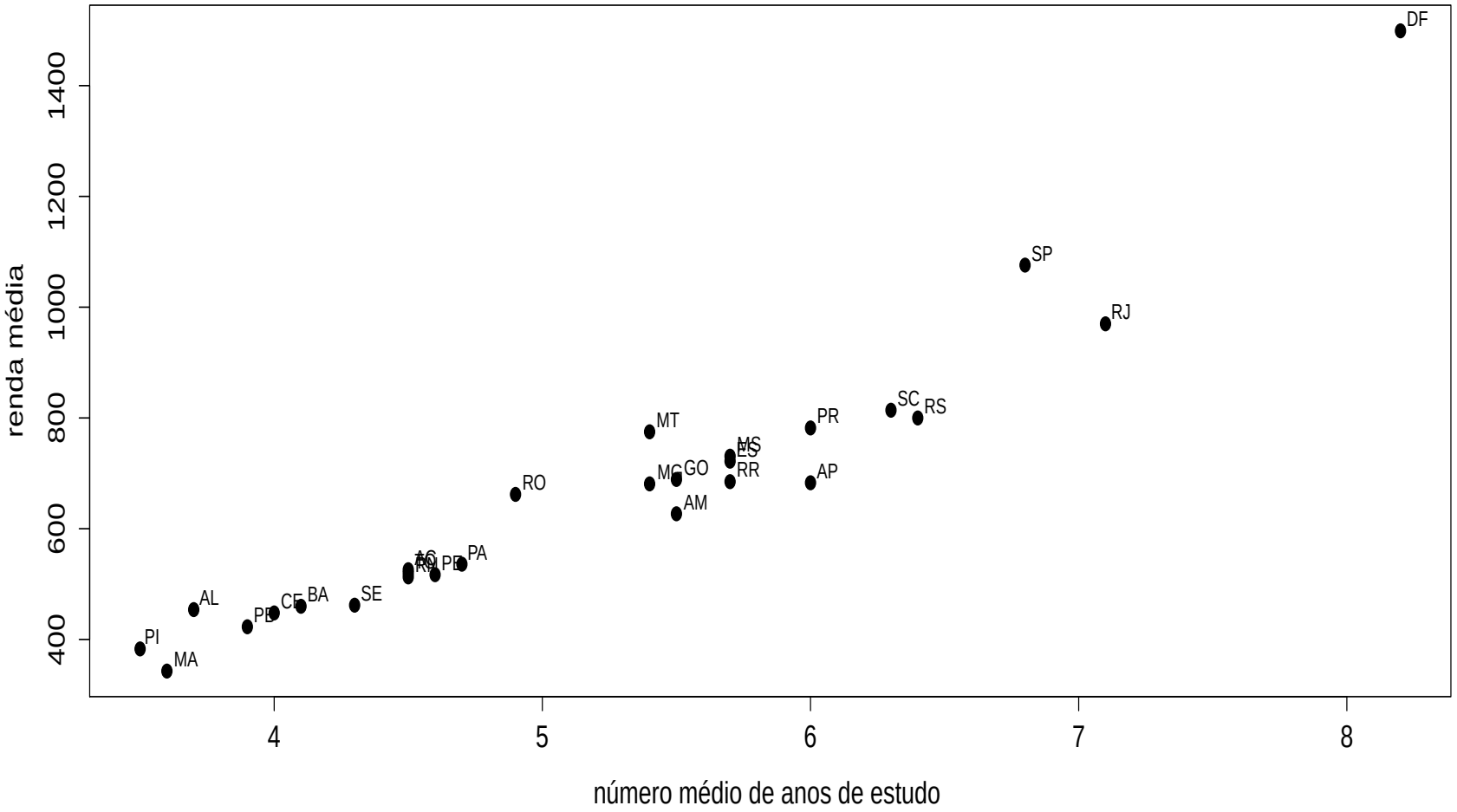
Métodos anteriores usando AIC/BIC

- Para qualquer um dos métodos anteriores, se usarmos alguma estatística de comparação de modelos (como AIC ou BIC), procedemos da seguinte forma
  - Sempre escolhemos o modelo (retirar/incluir a variável) que apresentar o menor valor da estatística.
  - O processo é interrompido quando as estatísticas para todos os modelos possíveis aumentarem em relação ao modelo corrente.
- Observação: as estatísticas AIC e BIC também servem para comparar modelos que difiram em termos da função de ligação e distribuição da variável resposta, entre outras características.

Exemplo 7: censo IBGE 2000

- O conjunto de dados em questão foi extraído do censo do IBGE de 2000 e apresenta para cada unidade da federação o número médio de anos de estudo e a renda média mensal (em reais) do chefe ou chefes do domicílio.
- Um dos objetivos é estudar o relacionamento da renda média mensal em função do número médio de anos de estudo.

Dispersão entre anos de escolaridade e renda



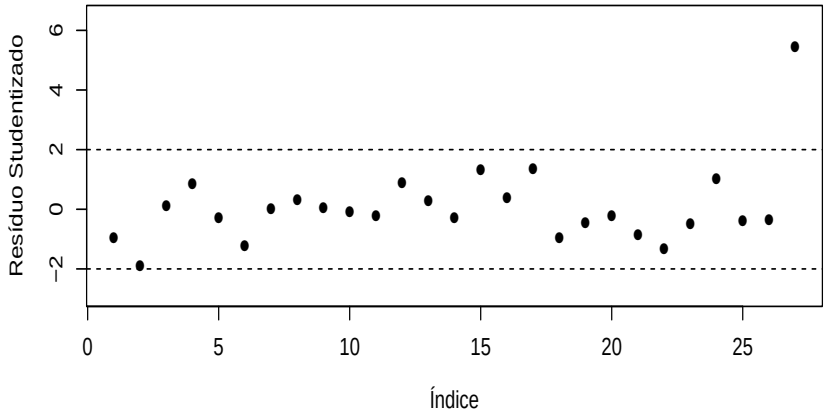
Cont.

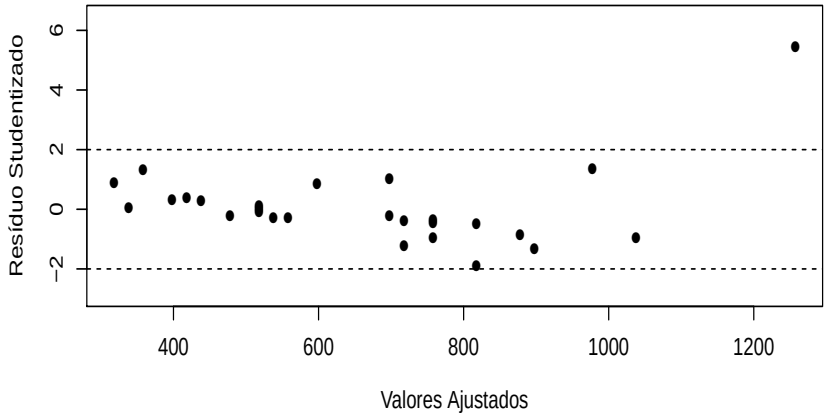
- Modelo 1: $Y_j = \beta_0 + \beta_1 x_j + \xi_j$
- Modelo 2: $Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \xi_j$

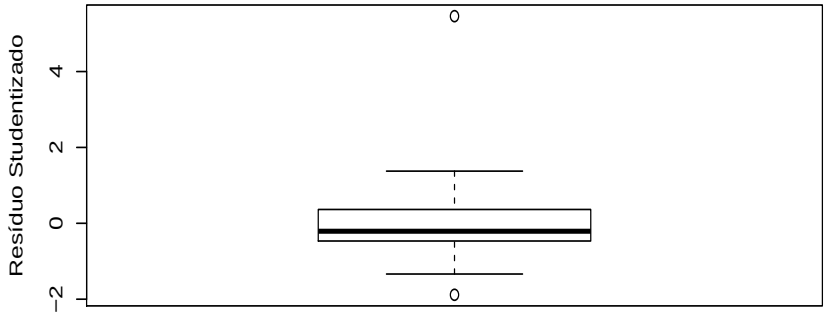
em que

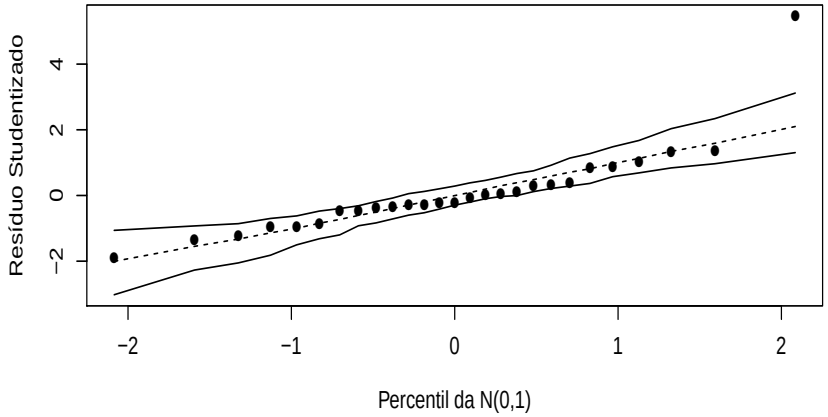
$$\xi_j \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

Modelo 1: gráficos de resíduos

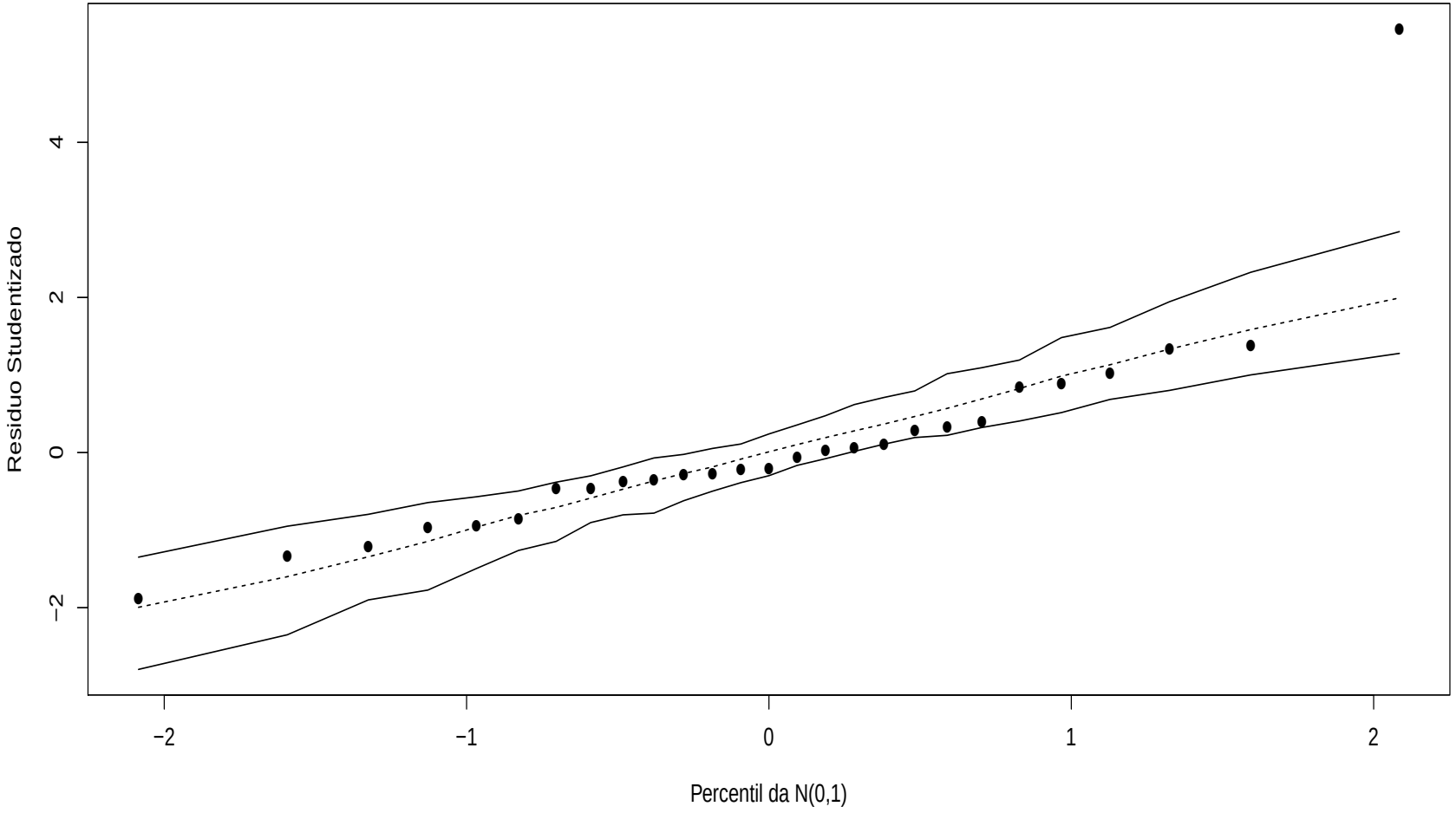








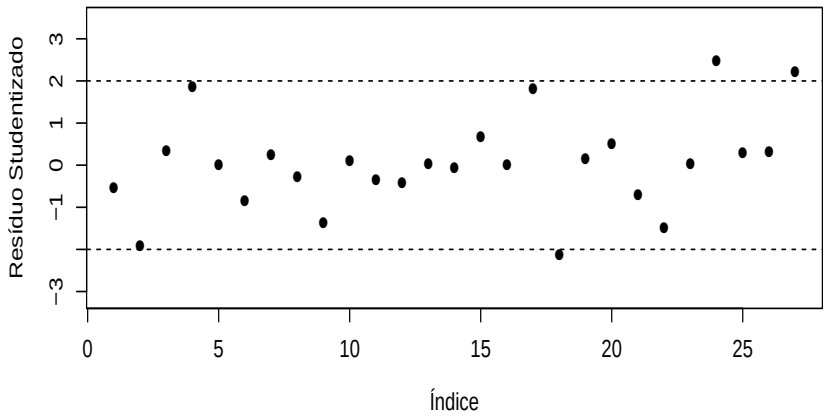
Modelo 1: gráfico de envelopes para os resíduos

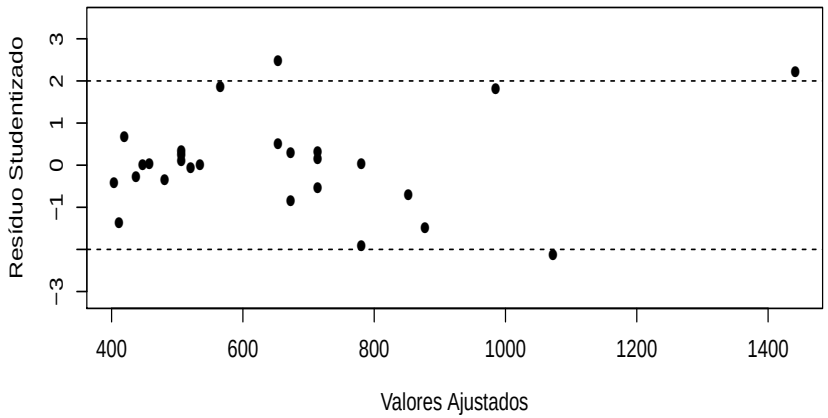


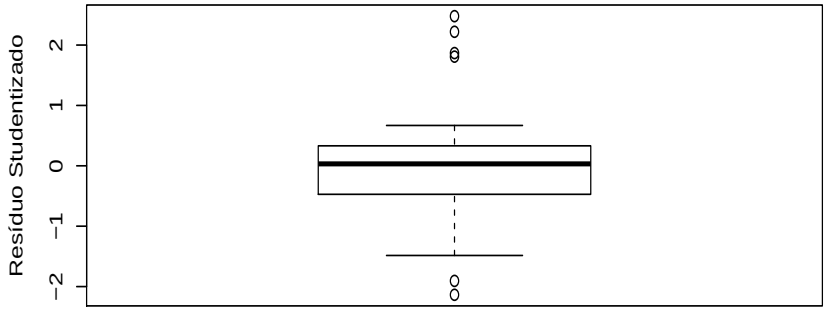
Comentários

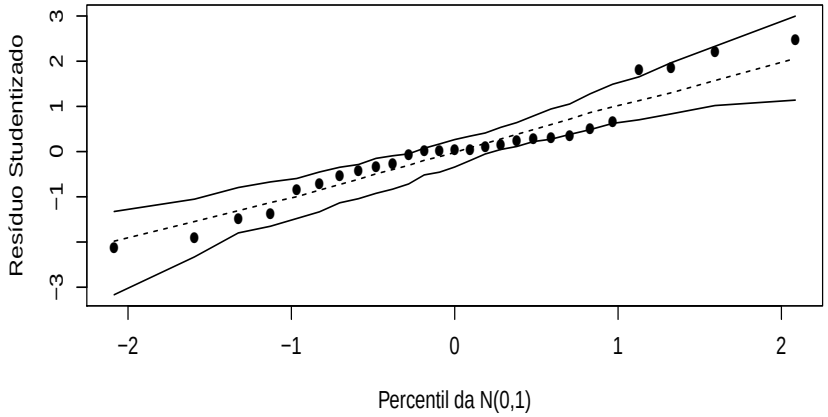
- Aparentemente os resíduos apresentam uma distribuição com caudas mais pesadas do que as da normal.
- Presença de heterocedasticidade nos resíduos.
- Provável ausência de correlação entre os resíduos.
- O ponto que aparece destacado é devido ao fato de que o modelo linear não capta bem a relação entre a renda e os anos de escolaridade.

Modelo 2: gráficos de resíduos

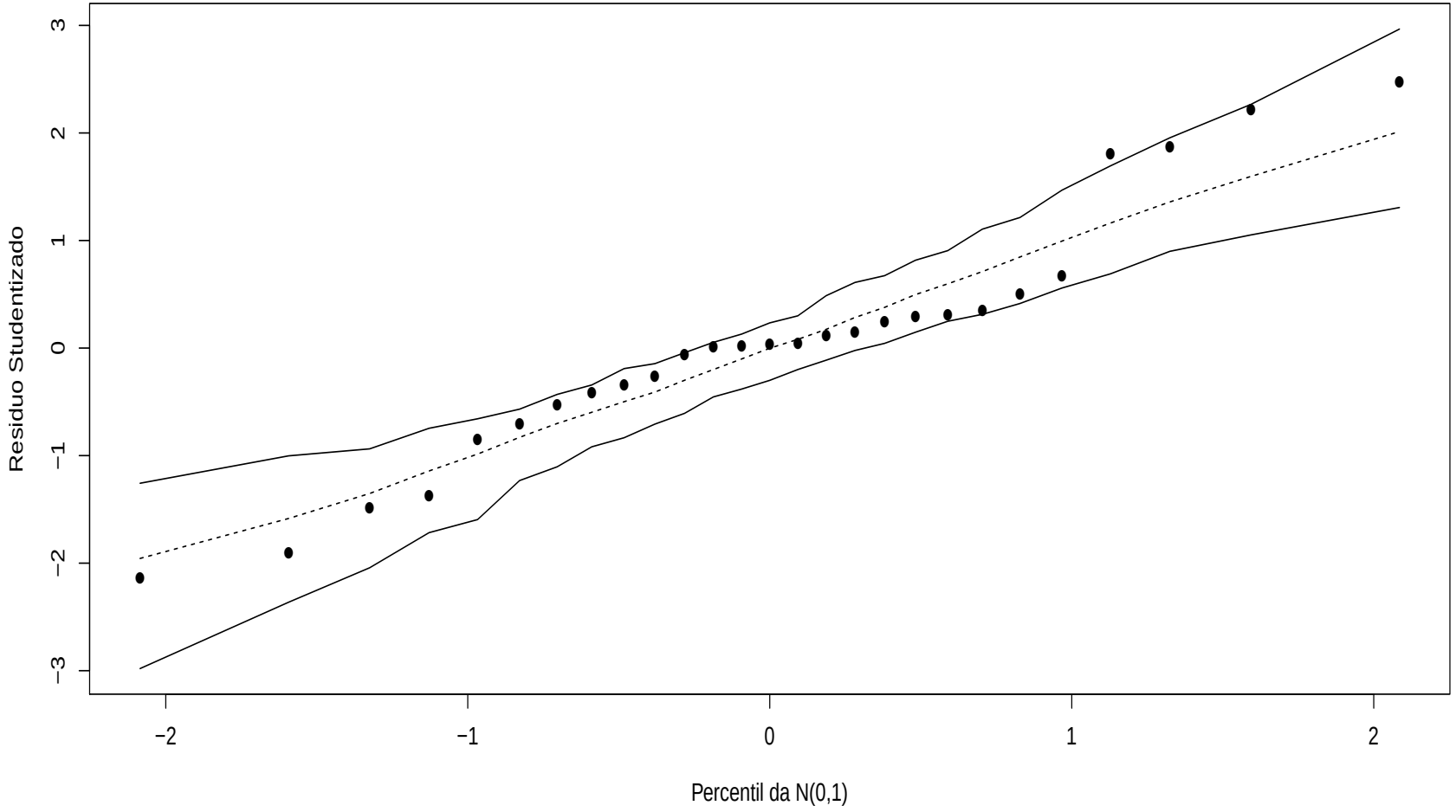








Modelo 2: gráfico de envelopes para os resíduos



Comentários

- Aparentemente os resíduos apresentam uma distribuição com caudas mais pesadas do que as da normal.
- Presença de heterocedasticidade nos resíduos.
- Provável ausência de correlação entre os resíduos.

Cont.

- Modelo 1

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	-381,28	69,40	[-524,23 ; -238,34]	-5,49	0,0001
β_1	199,83	13,03	[172,99 ; 226,66]	15,34	0,0001

- Modelo 2

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	546,98	196,80	[140,80 ; 953,16]	2,78	0,0104
β_1	-152,62	72,86	[-303,00 ; -2,24]	-2,09	0,0469
β_2	31,92	6,54	[18,41 ; 45,42]	4,88	0,0001

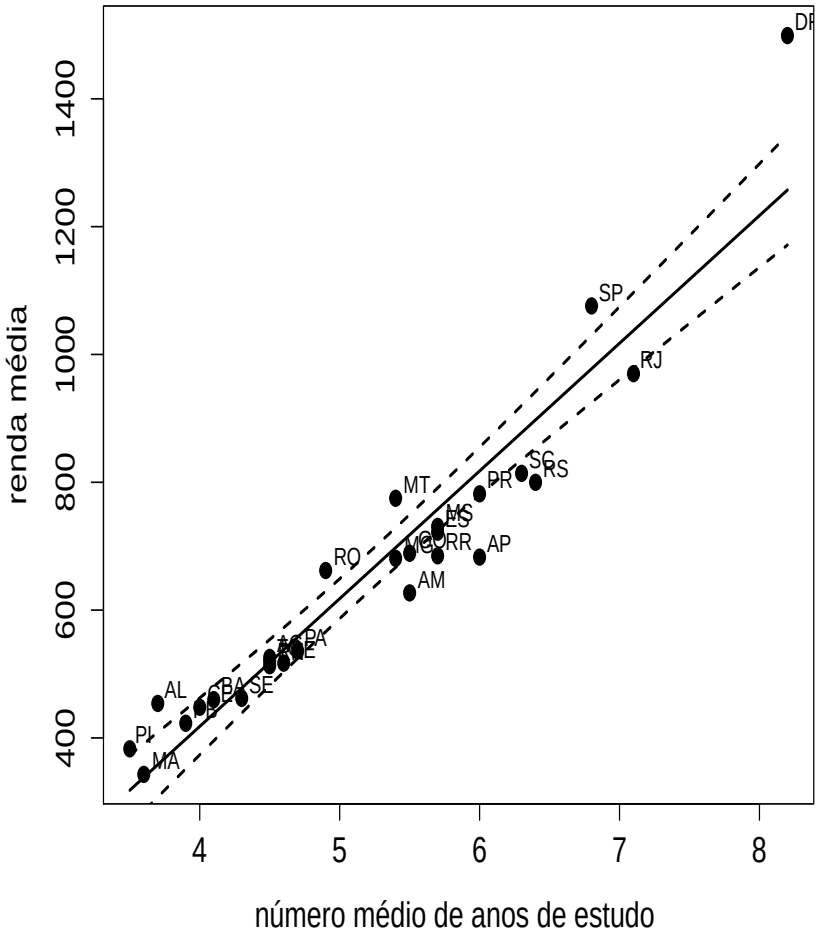
Cont.

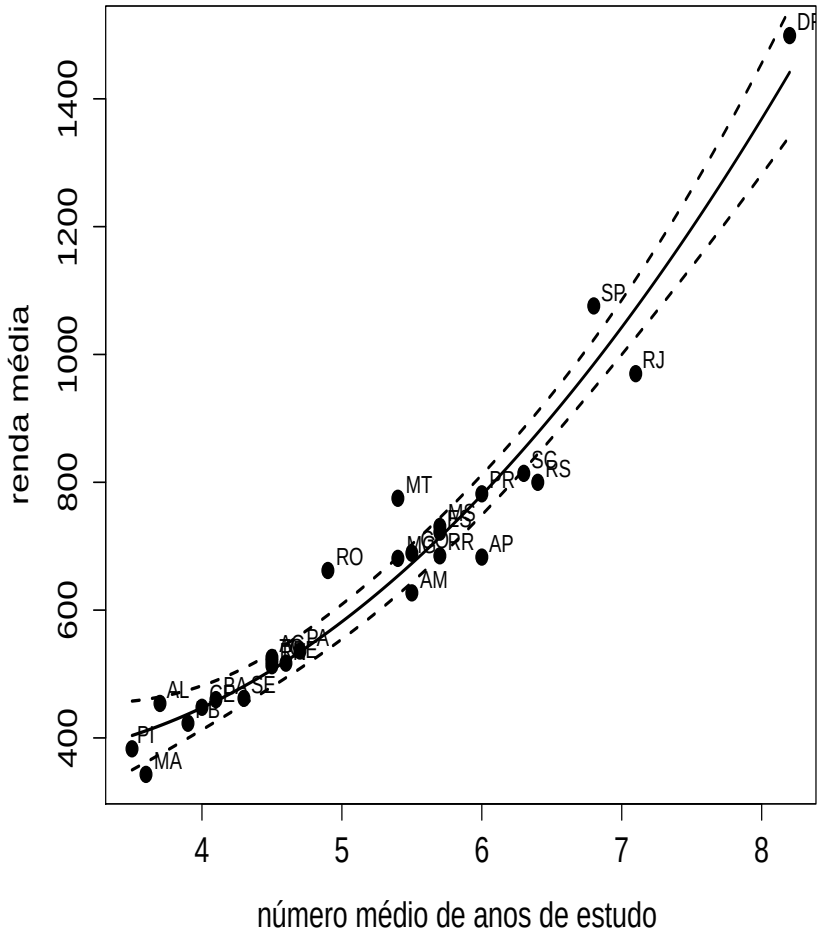
- Estatísticas de comparação dos modelos

Estatística	Modelo 1	Modelo 2
AIC	315,26	298,66
BIC	319,15	303,85
log-verossim.	-154,63	-145,33

- TRV, estatísticas e pvalor entre parênteses (H_0 modelo 1 vs H_1 : modelo 2): 18,80 ($< 0{,}0001$).

Modelos ajustados





Exemplo 2

- Vamos considerar o mesmo conjunto de dados.
- Além do modelo quadrático (anteriormente apresentado), vamos considerar o seguinte modelo (doravante, Modelo 3)

$$\begin{aligned} Y_i &\overset{ind}{\sim} Gama(\mu_i, \phi) \\ E(Y_i) &= \mu_i = e^{\beta_0+\beta_1 x_i} \\ V(Y_i) &= \mu_i^2 \phi^{-1} \end{aligned}$$

- Note que não existe estrutura hierárquica entre os modelos. Portanto, o TRV não pode ser utilizado.

Cont.

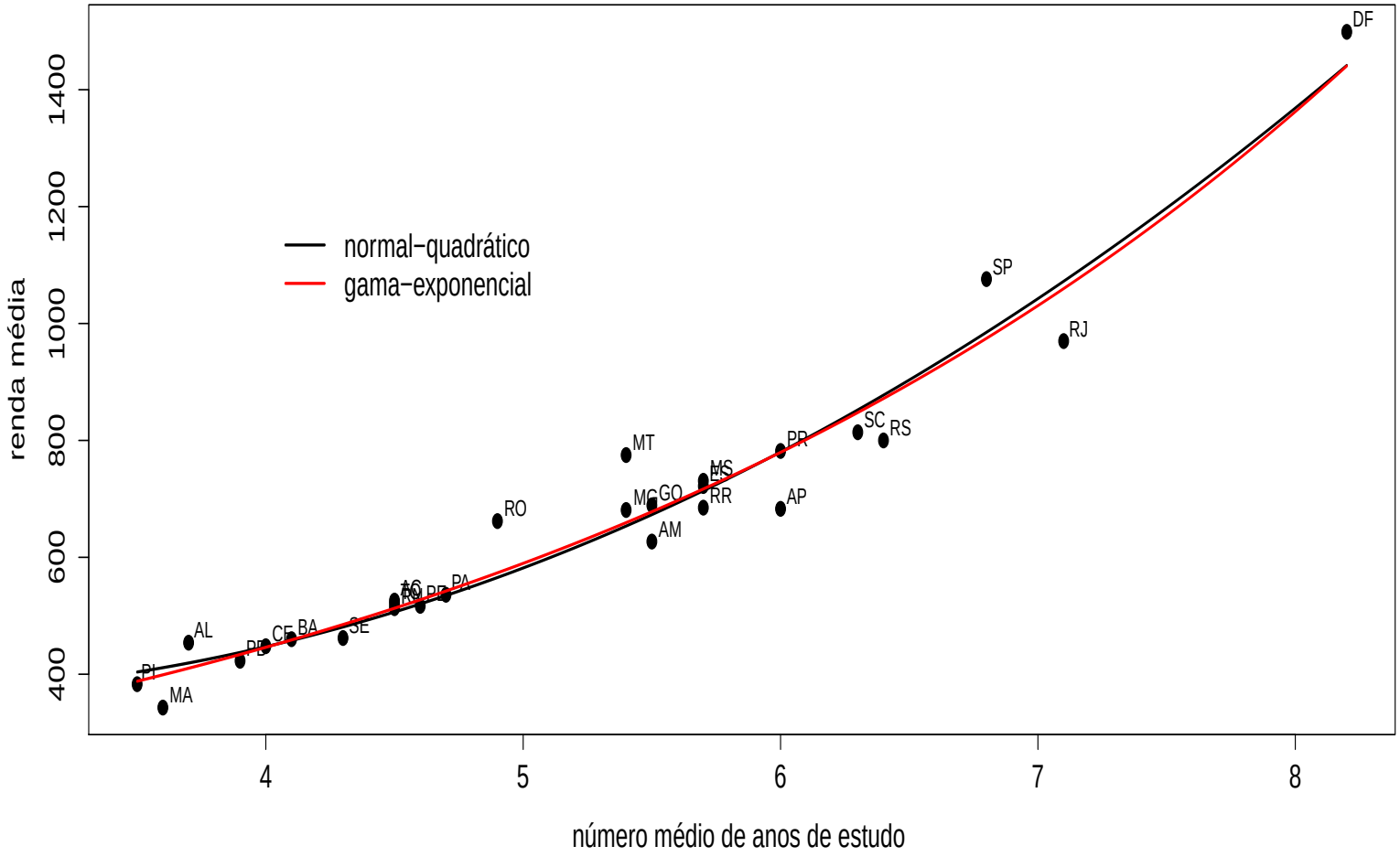
- Resumo do ajuste

Parâmetro	Estimativa	EP	Estat. t	p-valor
β_0	4,98	0,07	73,36	<0,0001
β_1	0,28	0,01	21,89	<0,0001

- Estatísticas de comparação dos modelos

Estatística	Modelo 2	Modelo 3
AIC	298,66	288,13
BIC	303,85	292,02
log-verossim.	-145,33	-141,07

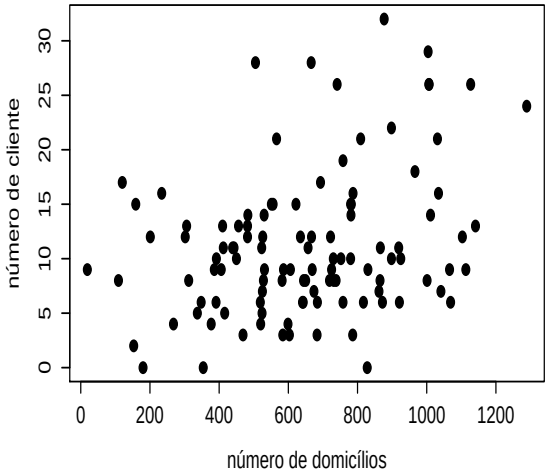
Modelos ajustados

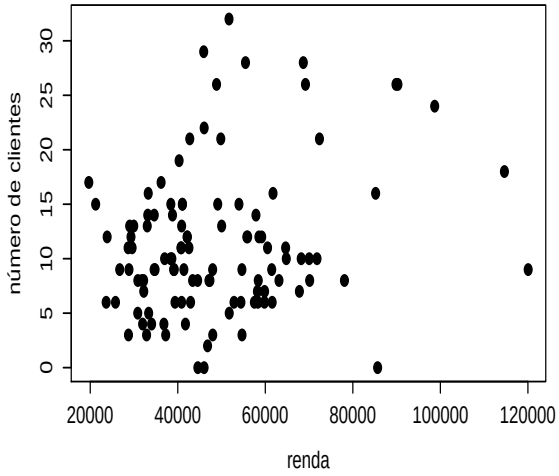


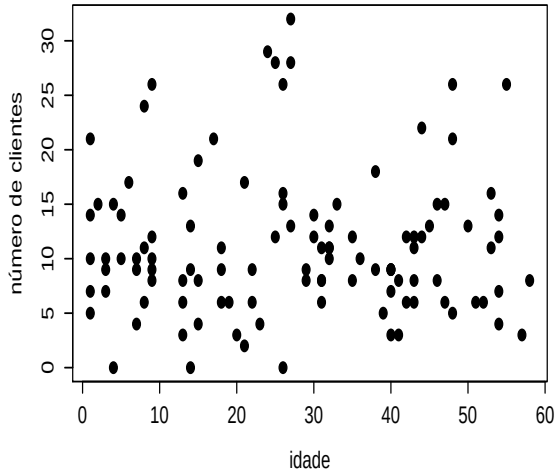
Exemplo 8: perfil dos clientes de uma loja

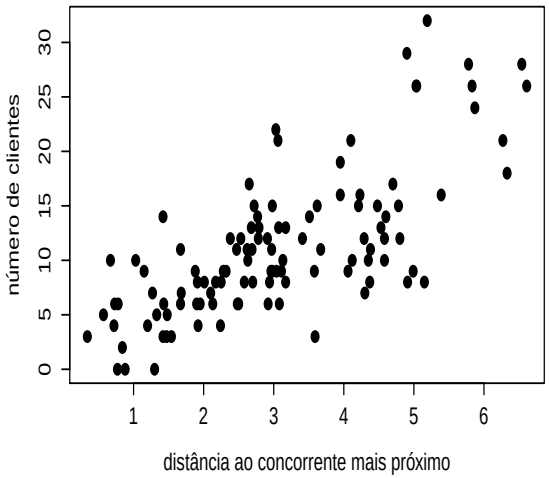
- Interesse: estudar o perfil dos clientes de uma determinada loja oriundos de 110 áreas de uma determinada cidade. Cada uma das 110 observações corresponde à uma área da cidade.
- Verificar como certas características (variáveis explicativas) afetam o número esperado de clientes em cada área (variável resposta).
- Variáveis explicativas: número de domicílios (em milhares) (x_1), renda média anual (em milhares de USD) (x_2), idade média dos domicílios (em anos)(x_3), distância ao concorrente mais próximo (em milhas) (x_4) e distância à loja (em milhas) (x_5).
- Variável resposta : número de clientes da referida loja (Y).

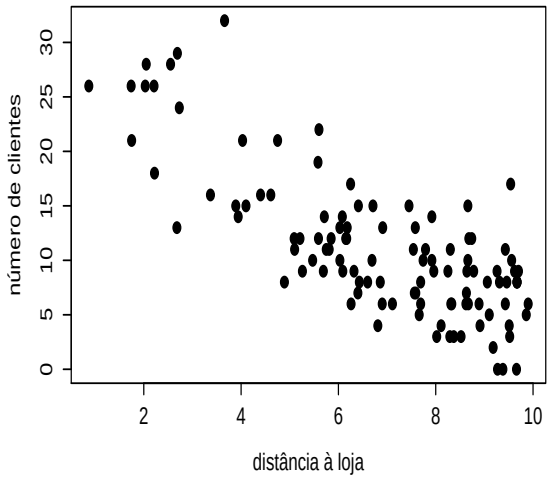
Gráficos de dispersão











Modelo (completo)

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(x_{1i} - \overline{x}_1) + \beta_2(x_{2i} - \overline{x}_2) + \beta_3(x_{3i} - \overline{x}_3) + \beta_4(x_{4i} - \overline{x}_4) \\ &+ \beta_5(x_{5i} - \overline{x}_5) + \xi, i = 1, 2, ..., 110 \end{aligned}$$

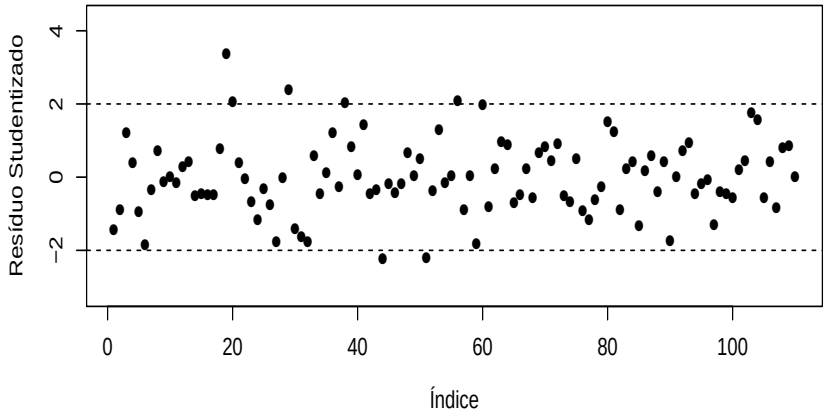
- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- x_{ji} : valor da variável explicativa j, $j = 1, 2, ..., 5$, associada à área i e $\overline{x}_j = \frac{1}{110}\sum_{i=1}^{110} x_{ji}$, $j = 1, 2, ..., 5$.
- β_j : incremento (positivo ou negativo) no valor esperado do número de clientes, para o aumento em uma unidade no valor da covariável j mantendo-se todas as outras fixas.

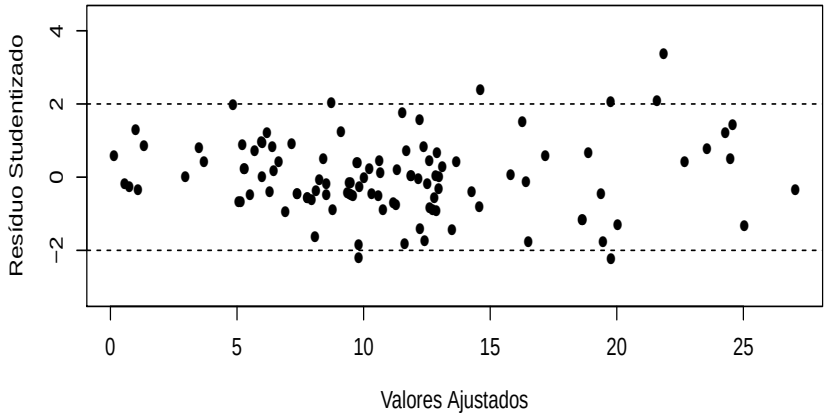
Estimativas dos parâmetros

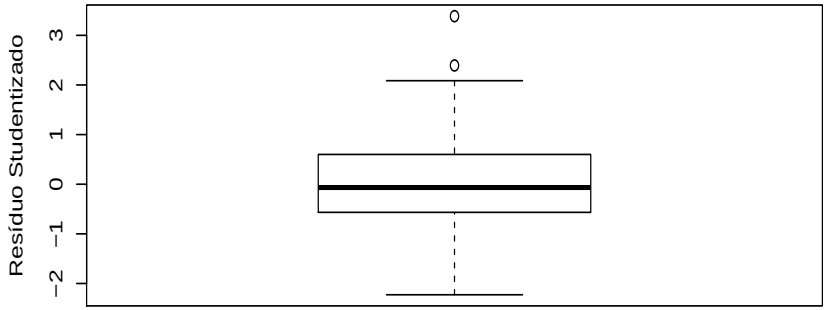
Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	11,2000	0,3063	[10,5928 ; 11,8072]	36,5641	$<$ 0,0001
β_1	0,0066	0,0015	[0,0036 ; 0,0096]	4,3386	$<$ 0,0001
β_2	-0,0001	$<$ 0,0001	[-0,0002 ; -0,0001]	-5,0885	$<$ 0,0001
β_3	-0,0359	0,0188	[-0,0732 ; 0,0013]	-1,9118	0,0587
β_4	1,9036	0,2579	[1,3925 ; 2,4147]	7,3824	$<$ 0,0001
β_5	-1,7104	0,1739	[-2,0550 ; -1,3657]	-9,8370	$<$ 0,0001

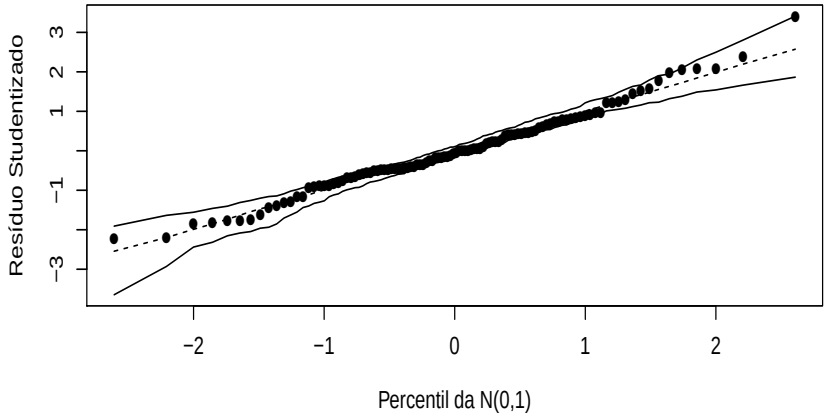
Todos os parâmetros são, aparentemente, significativos, à exceção, talvez, β_3 (idade média dos domicílios).

Modelo completo: gráficos de resíduos

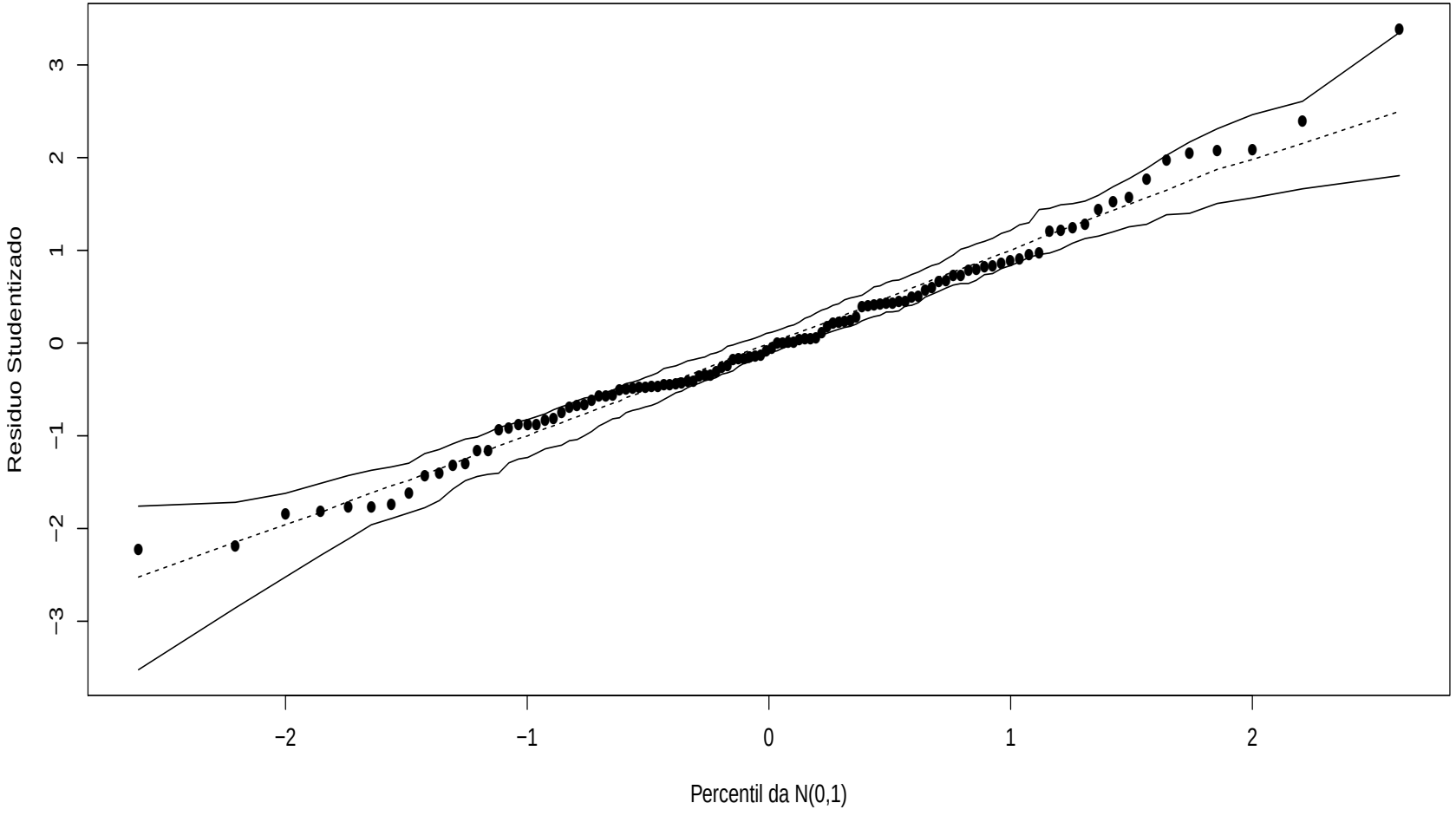








Modelo completo: gráficos de envelopes para os resíduos



Seleção de modelos

- A aplicação da metodologia stepwise, começando com o modelo só com o intercepto ou começando com o modelo completo, indicou, em ambos os casos, que todas as variáveis são significativas.
- Aparentemente, o modelo não se ajustou bem aos dados : presença de heterocedasticidade e aparente não normalidade dos resíduos (assimetria).
- Além disso, estamos impondo uma natureza contínua à variável resposta a qual é, como vimos, discreta.