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)

https://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 (em que todos 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 $\Delta = \dfrac{L_1(\widehat{\boldsymbol{\theta}}_1)}{L_2(\widehat{\boldsymbol{\theta}}_2)}$.
- 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.

Estatísticas de comparação de modelos

- Veremos as aseguintes medidas:
  - AIC ("Akaike Information Criteria" - Critério de Informação de Akaike).
  - BIC ("Bayesian Information Criteria" - Critério de informação Bayesiano, também conhecido como SIC).
  - AIC_c ("Corrected Akaike Information Criteria" - critério de Informação de Akaike corrigido).
  - SABIC ("Sample Adjusted BIC" - BIC ajustado pelo tamanho da amostra).
  - HQNIC ("HannanQuinn information criterion" - Critério de informação de de Hanna-Quinn).

- As estatísticas mencionadas anteriormente, para o i-ésimo modelo, são dadas, respectivamente, por:

$$
\begin{aligned}
AIC_i &= -2l_i(\widetilde{\theta}_i) + 2k \;;\; BIC_i = -2l_i(\widetilde{\theta}_i) + k\ln(n) \\
AIC_{c_i} &= AIC_i + \frac{2k(k+1)}{n-k-1} \;;\; SABIC_i = -2l_i(\widetilde{\theta}_i) + k\ln\left(\frac{n+2}{24}\right) \\
HQCIC_i &= -2l_i(\widetilde{\theta}_i) + 2k\ln\left(\ln(n)\right)
\end{aligned}
$$

  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 $\boldsymbol{C\beta}$, 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 $\boldsymbol{C\beta}$, 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 $\boldsymbol{C\beta}$, 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 $\boldsymbol{C}\boldsymbol{\beta}$, 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.

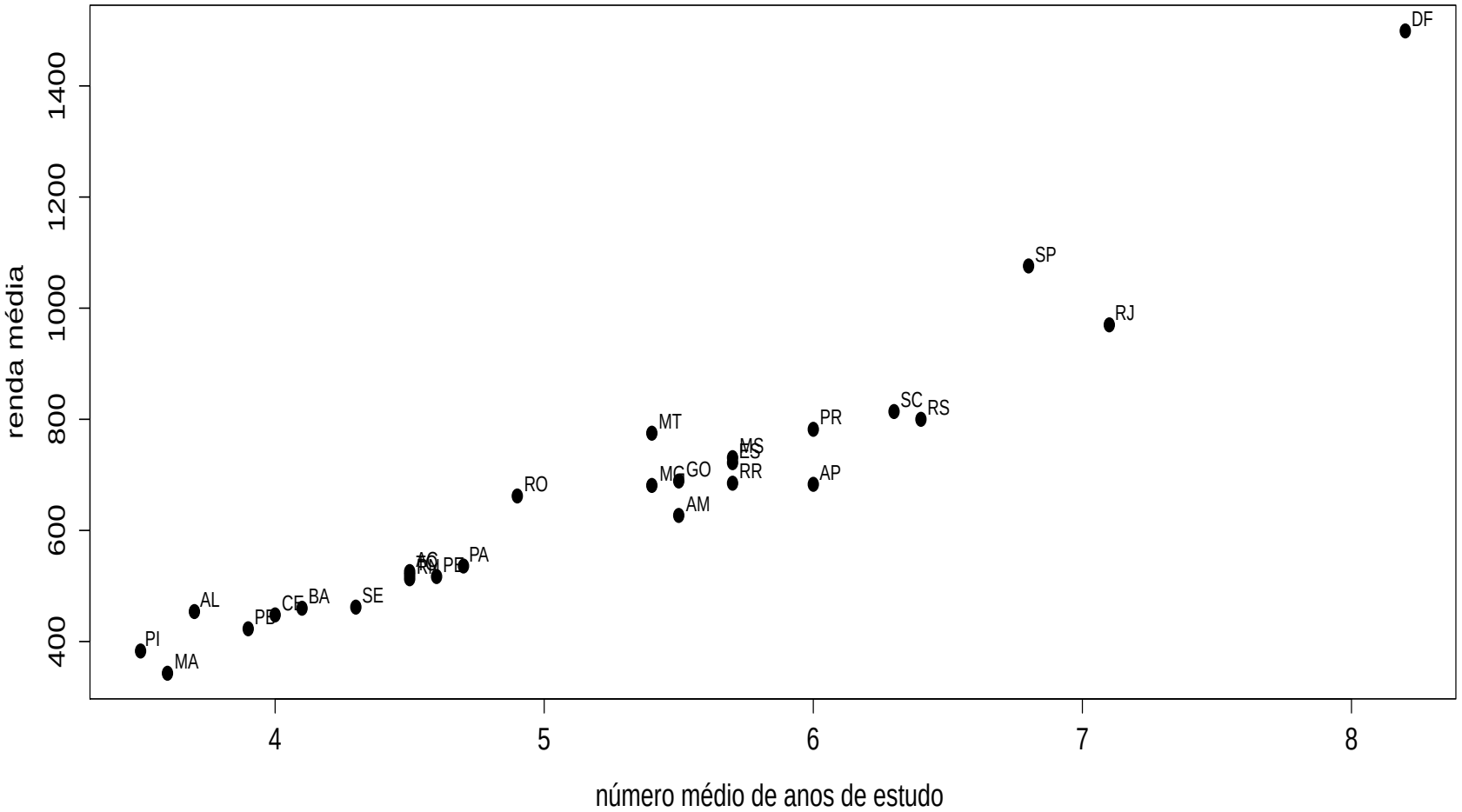
Métodos anteriores usando AIC/BIC

- Para qualquer um dos métodos anteriores, se usarmos alguma estatística de comparação de modelos, 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.

Exemplo 8: 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.

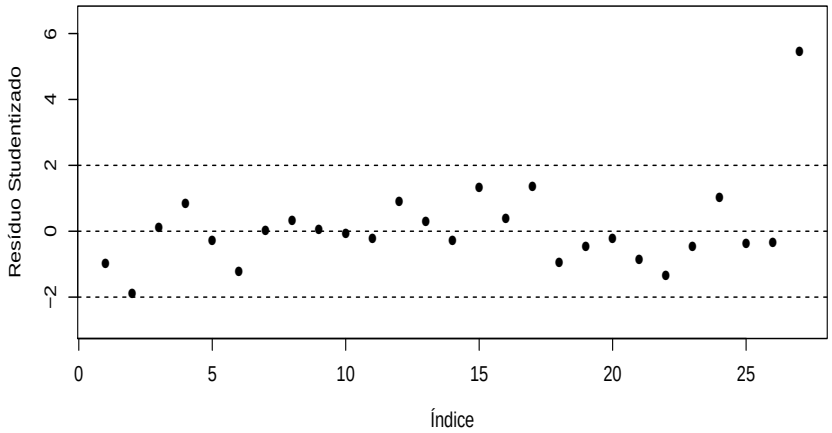
- Modelo 1: $Y_j = \beta_0 + \beta_1 x_j + \xi_j$
- Modelo 2: $Y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \xi_j$

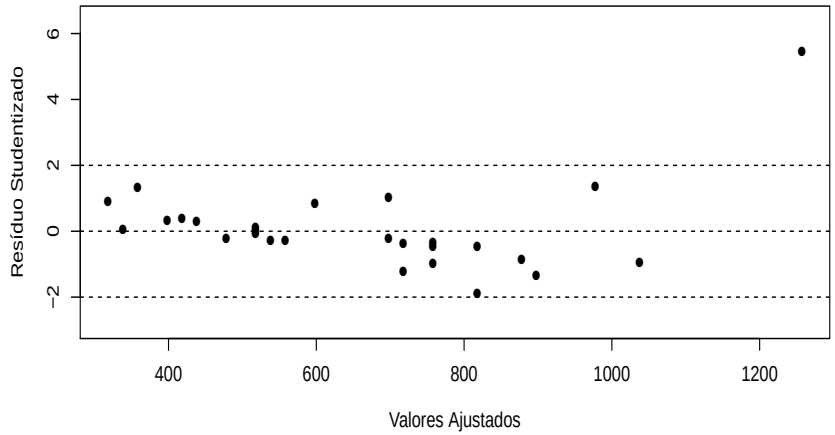
  em que

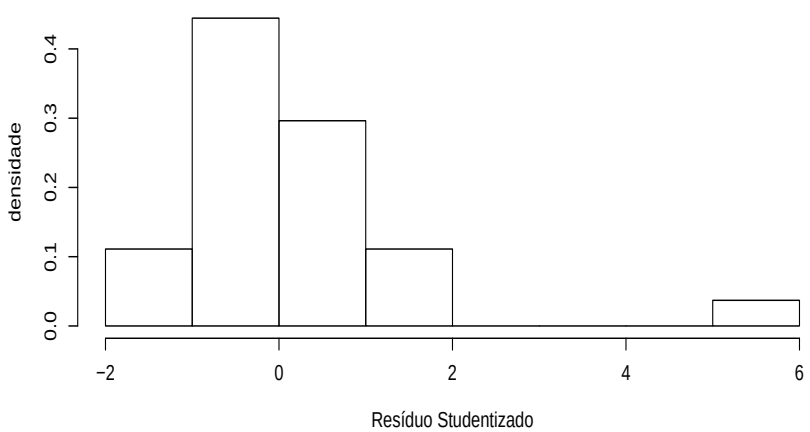
$$\xi_j \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

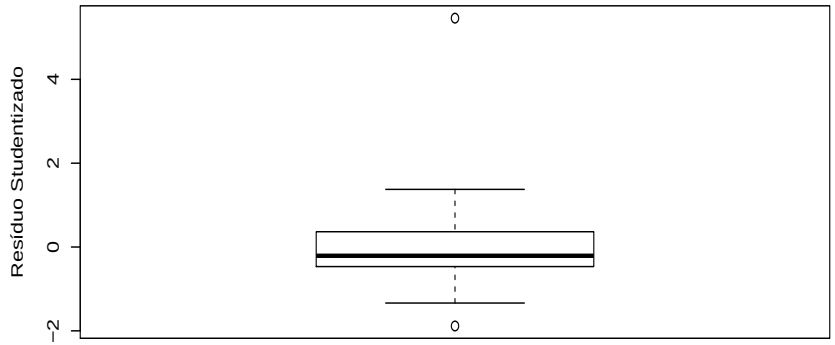
- Exercício: interpretar os parâmetros do modelo 2 (e/ou funções deles) determinando qual é o número médio de anos de idade que leva (dentro do intervalo observado) a renda mínima e/ou máxima bem como seria a respectiva renda (estimativa pontual e intervalar).

Modelo 1: gráficos de resíduos

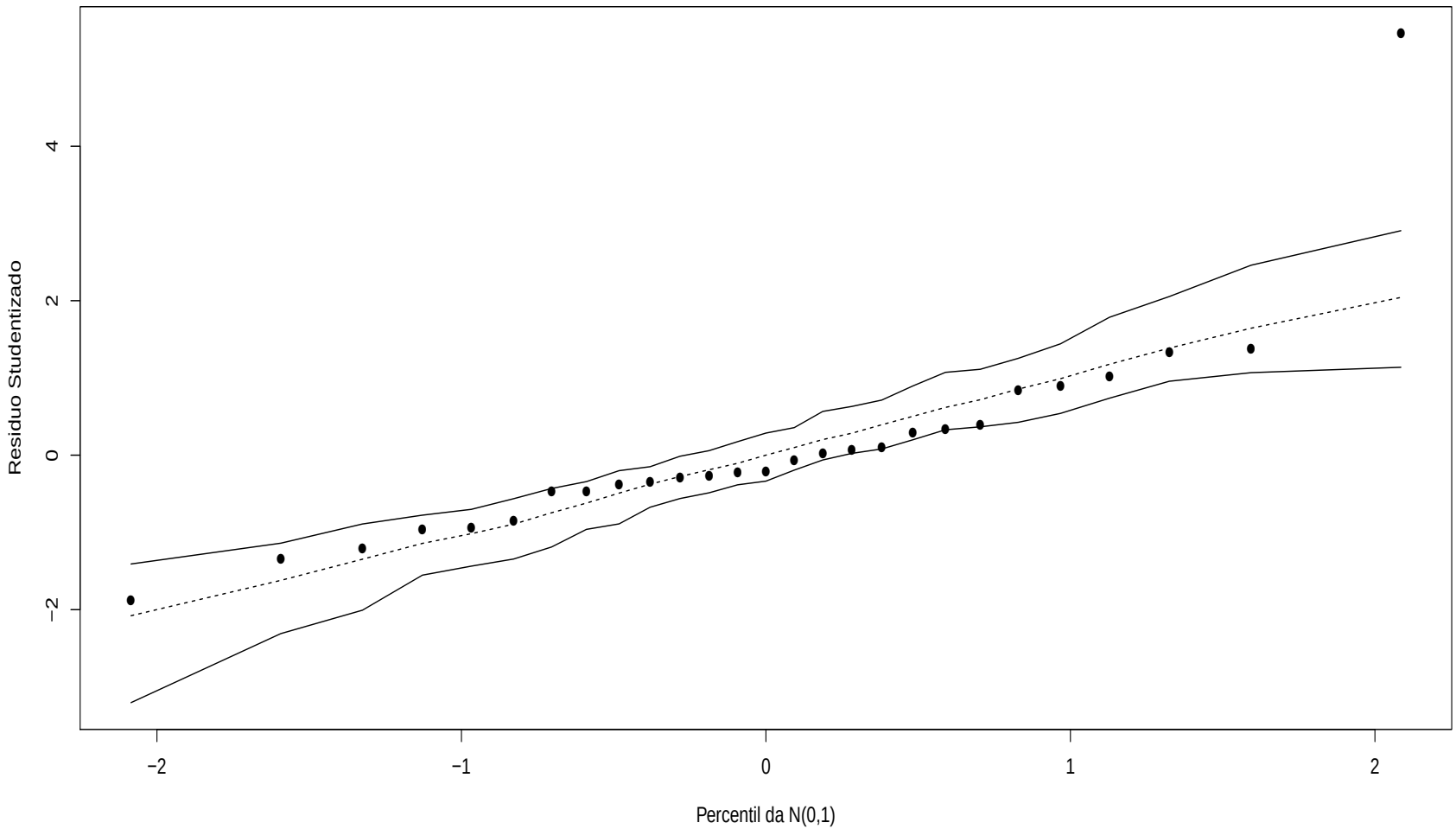








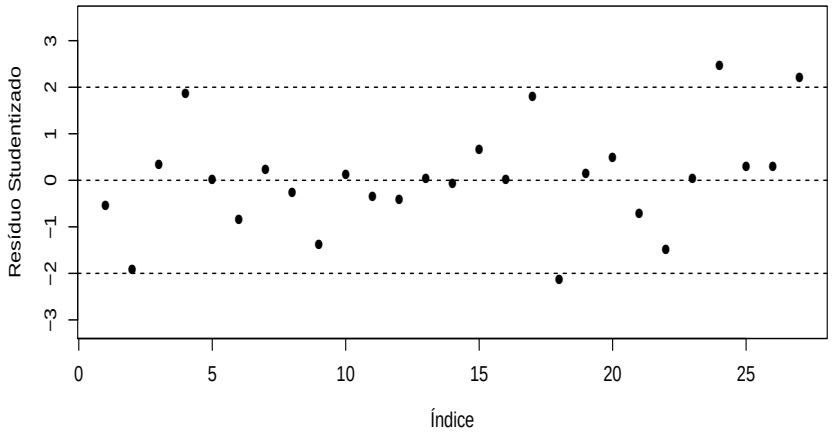
Modelo 1: gráfico de envelopes para os resíduos

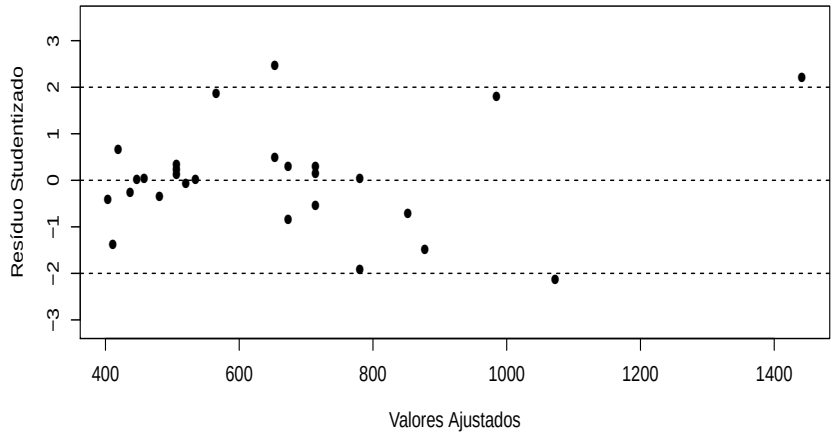


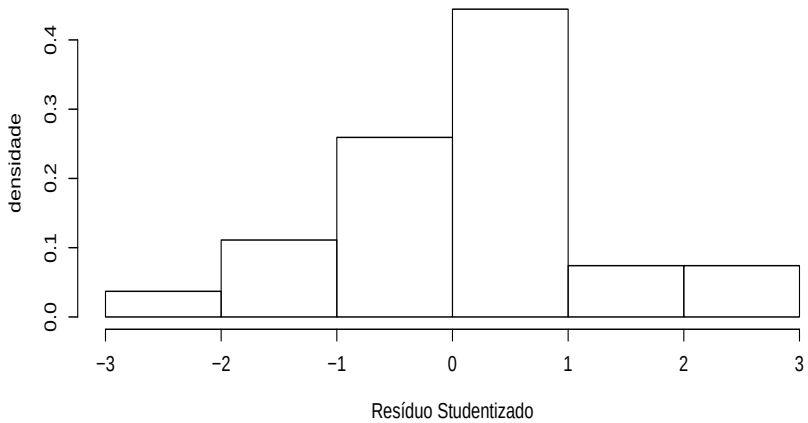
Comentários

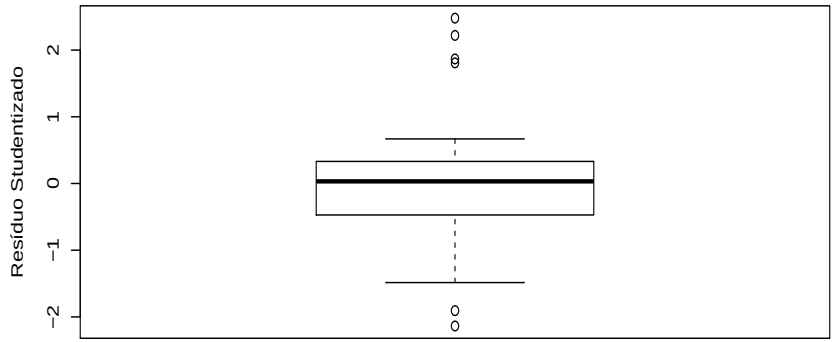
- Presença de heterocedasticidade nos resíduos.
- Provável ausência de correlação entre os resíduos.
- Possível ausência de normalidade dos resíduos.
- O ponto que aparece destacado é devido ao fato de que o modelo linear não capta bem a relação entre a renda e os anos de escolaridade.

Modelo 2: gráficos de resíduos

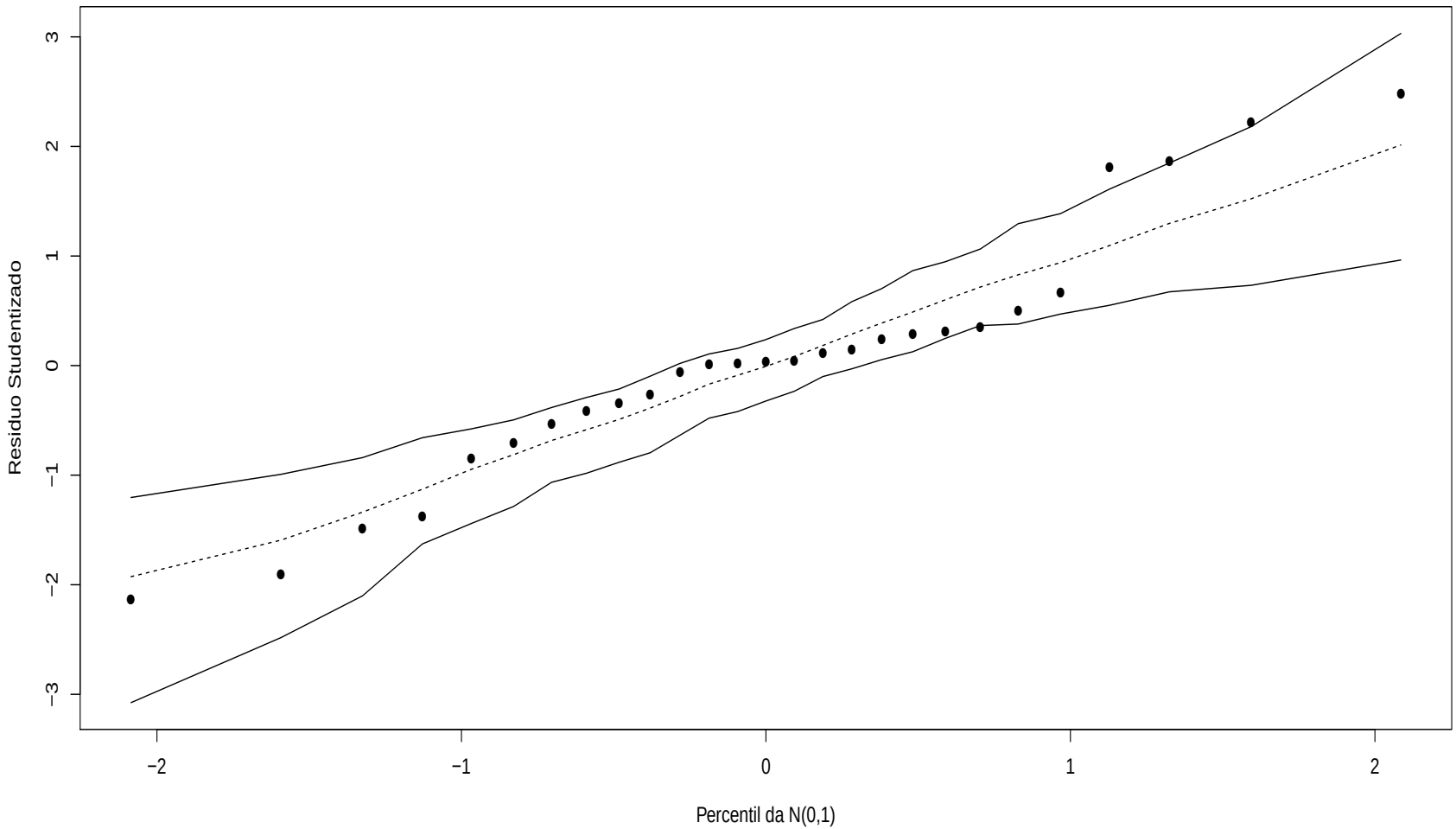








Modelo 2: gráfico de envelopes para os resíduos



Comentários

- Presença de heterocedasticidade nos resíduos.
- Provável ausência de correlação entre os resíduos.
- Possível ausência de normalidade dos 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

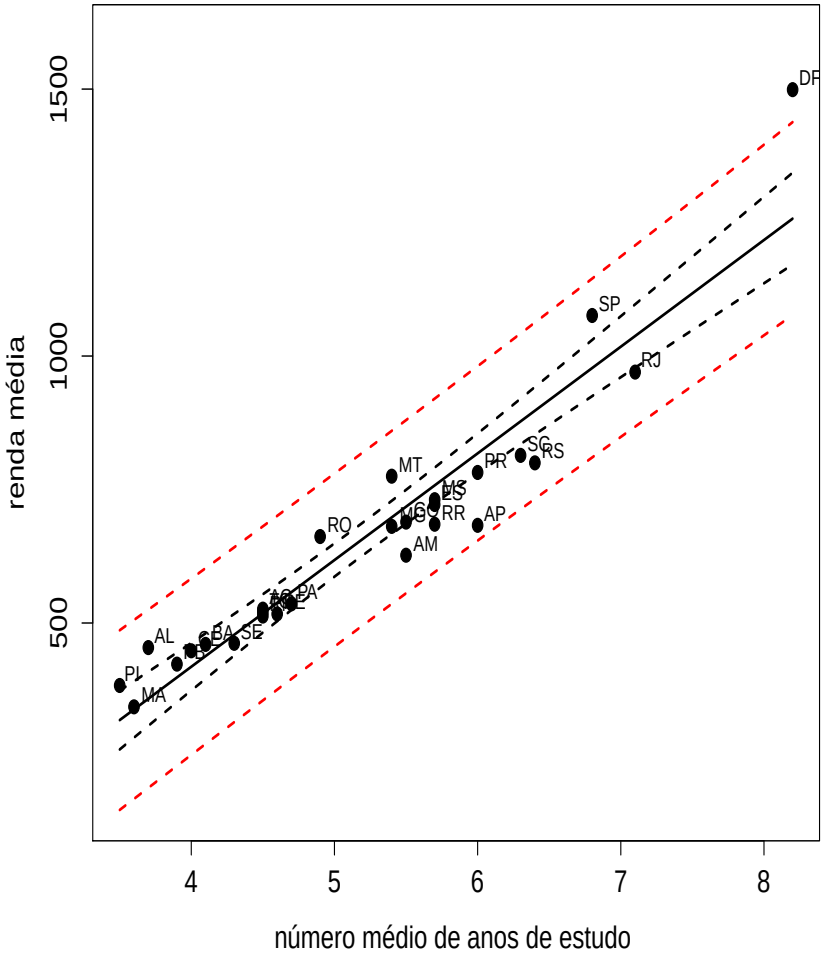
Cont.

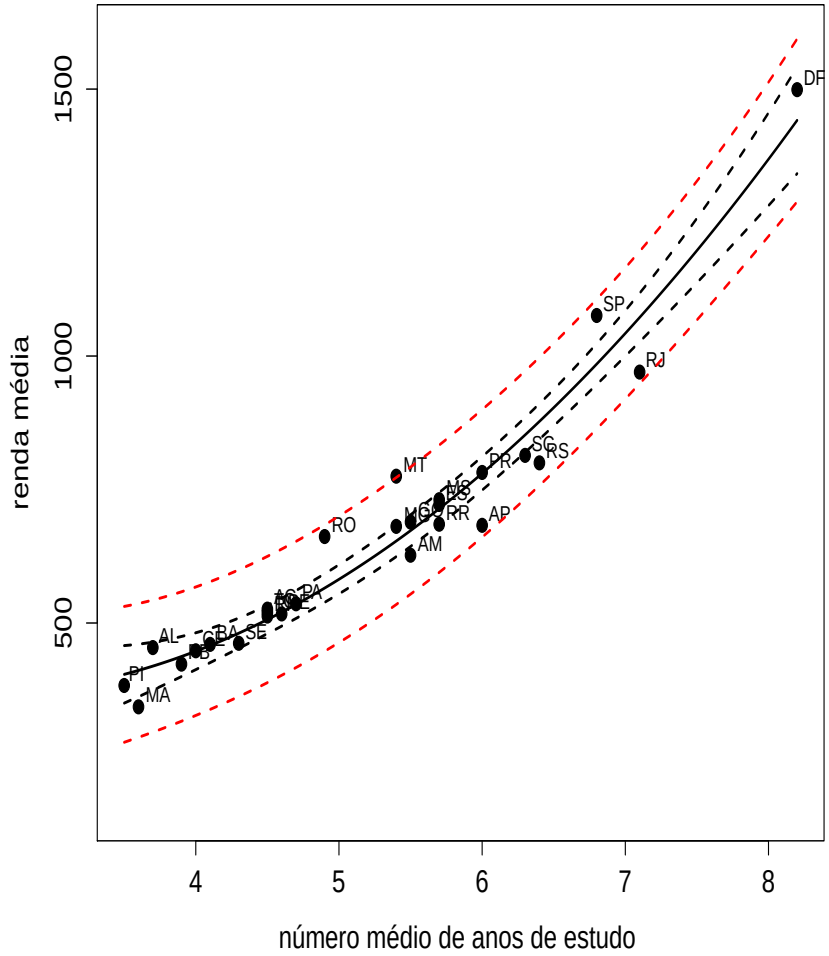
- Estatísticas de comparação dos modelos

Estatística	Modelo linear	Modelo quadrático
AIC	315,26	298,66
BIC	319,15	303,85
AICc	315,76	299,71
SABIC	309,64	291,23
HQCIC	314,03	297,82
$-2\times$ logver.	309,26	290,66

- TRV (versão assintótica), 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)

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

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

Cont.

- Resumo do ajuste

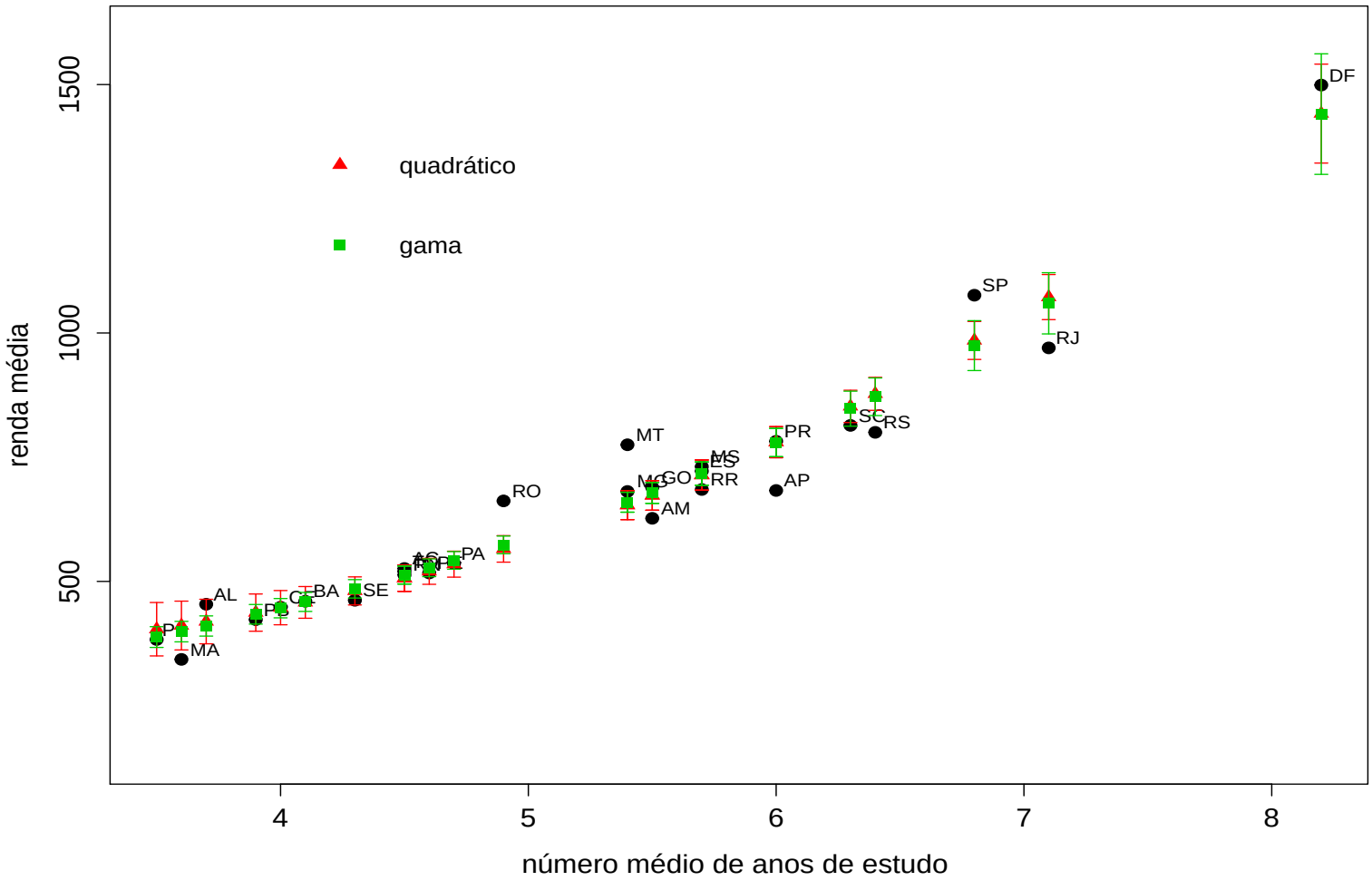
Parâmetro	Estimativa	EP	IC(95%)	t value	Pr(>|t|)
β_0	4,98	0,07	[4,84 ; 5,12]	73,36	$<$ 0,0001
β_1	0,28	0,01	[0,25 ; 0,31]	21,89	$<$ 0,0001

Cont.

- Estatísticas de comparação dos modelos

Estatística	Modelo 3	Modelo 2
AIC	288,13	298,66
BIC	292,02	303,85
AICc	288,63	299,71
SABIC	282,51	291,23
HQCIC	286,90	297,82
$-2\times$ logver.	282,13	290,66

Modelos ajustados



Exemplo 3: efeito do fósforo na produção de milho

- Vamos ajustar dois modelos e compará-los.
- Modelo linear (reta) e modelo quadrático (parábola): ambos fazem parte da família de modelos de regressão linear.

Modelo linear 1: reta

$$Y_i = \beta_0 + \beta_1 x_i + \xi_i, i = 1, 2, ..., 20$$

- x_i : quantidade de fósforo ministrada a i-ésima parcela.
- β_0 : valor esperado (média) da produção de milho quando a quantidade de fósforo aplicada é igual à 0.
- β_1 : incremento no valor esperado da produção de milho quando a quantidade de fósforo aplicada aumenta em uma unidade.
- $\xi_i \overset{iid}{\sim} N(0, \sigma^2)$.

Modelo linear 2: parábola

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \xi_i, i = 1, 2, ..., 20$$

- x_i : quantidade de fósforo ministrada a i-ésima parcela.
- β_0 : valor esperado (média) da produção de milho quando a quantidade de fósforo aplicada é igual à 0.
- A interpretação isolada dos parâmetros β_1 e β_2 é complicada mas, podemos dizer que $\frac{-\beta_1}{2\beta_2}$ é a quantidade de fósforo que leva à produção máxima de milho.
- $\xi_i \overset{iid}{\sim} N(0, \sigma^2)$.

Cont.

- Modelo 1

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	5,823	0,562	[4,643 ; 7,002]	10,370	$<$ 0,0001
β_1	0,044	0,009	[0,025 ; 0,064]	4,834	0,0001

- Modelo 2

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	5,0182	0,6122	[3,7266 ; 6,3098]	8,1971	$<$ 0,0001
β_1	0,1087	0,0290	[0,0475 ; 0,1699]	3,7460	0,0016
β_2	-0,0006	0,0003	[-0,0012 ; -0,0001]	-2,3132	0,0335

Cont.

- Estatísticas de comparação dos modelos

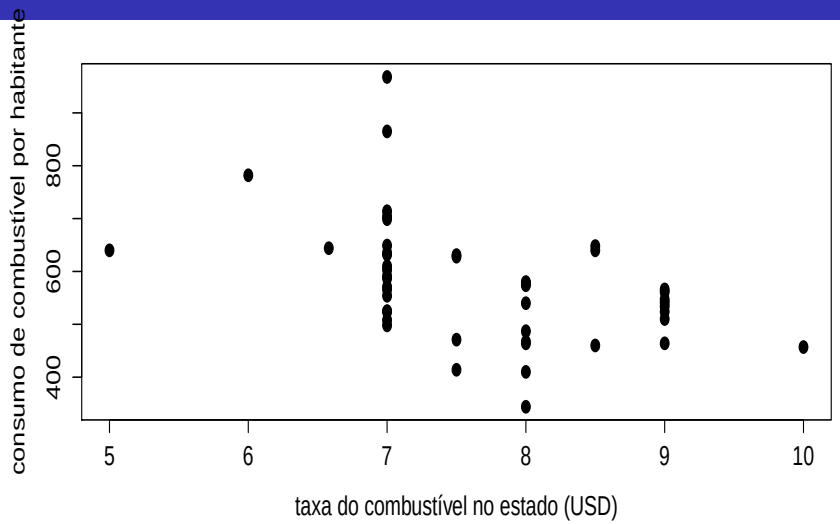
Estatística	Modelo 1	Modelo 2
AIC	75,50	72,03
BIC	78,49	76,01
AICc	76,21	73,53
SABIC	69,33	63,77
HQCIC	73,89	70,61
$-2\times$ lover.	69,50	64,03

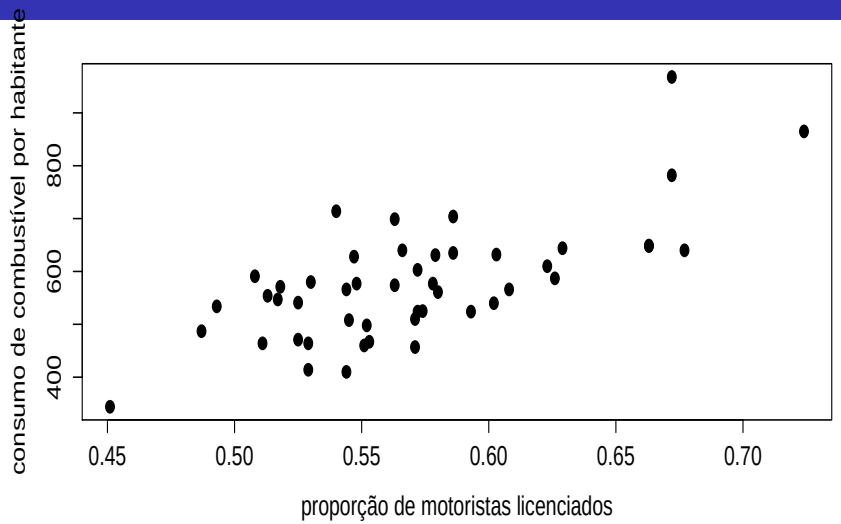
- TRV, estatísticas e pvalor entre parênteses (H_0 modelo 1 vs H_1 : modelo 2): 5,47 ($<$ 0,0193).

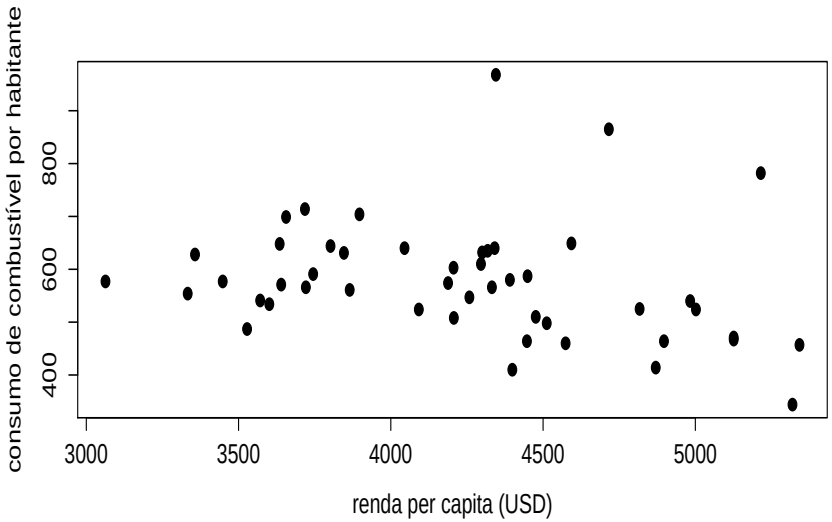
Exemplo 4: consumo de combustível

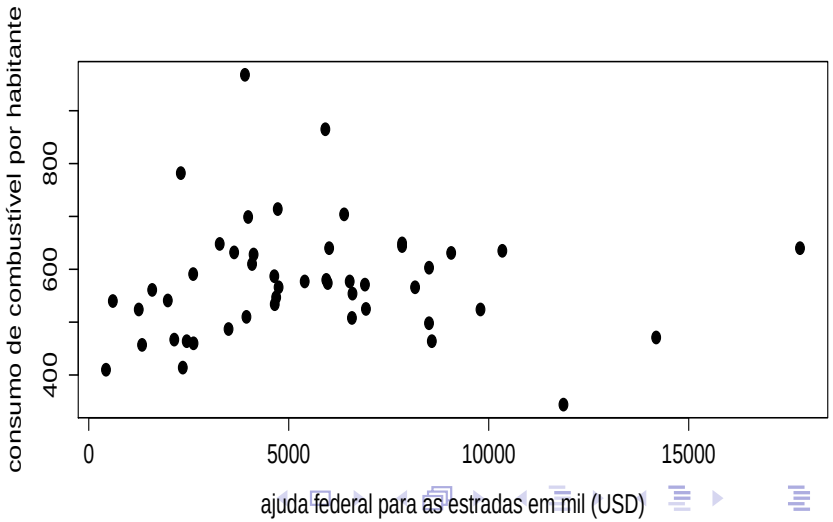
- Dos 48 estados (contíguos) dos Estados Unidos (em um certo ano) mediu-se:
  - Taxa do combustível no estado em USD - taxa (x_1).
  - Proporção de motoristas licenciados - licença (x_2).
  - Renda per capita em USD - renda (x_3).
  - Ajuda federal para as estradas em mil USD - estradas (x_4).
  - Consumo de combustível por habitante - consumo (Y).
- Objetivo: tentar explicar o consumo de combustível em função das outras variáveis. Fonte de consulta: Paula (2013). Fonte original (Gray, 1989).

Gráficos de dispersão









Exemplo 4: consumo de combustível

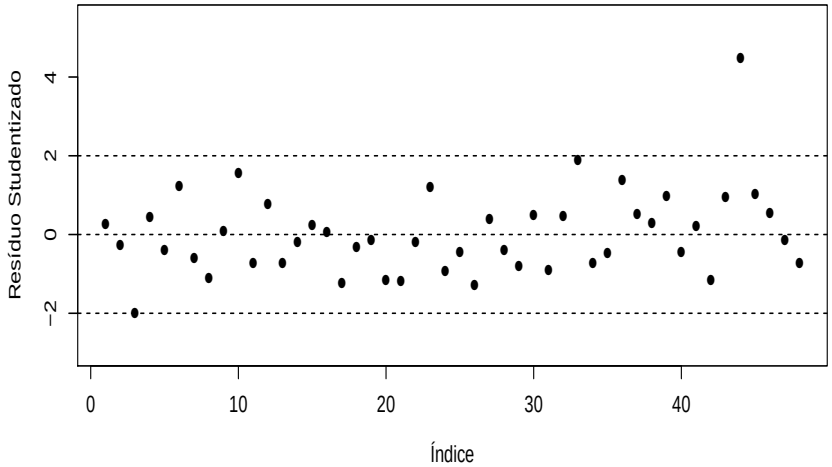
Modelo 2

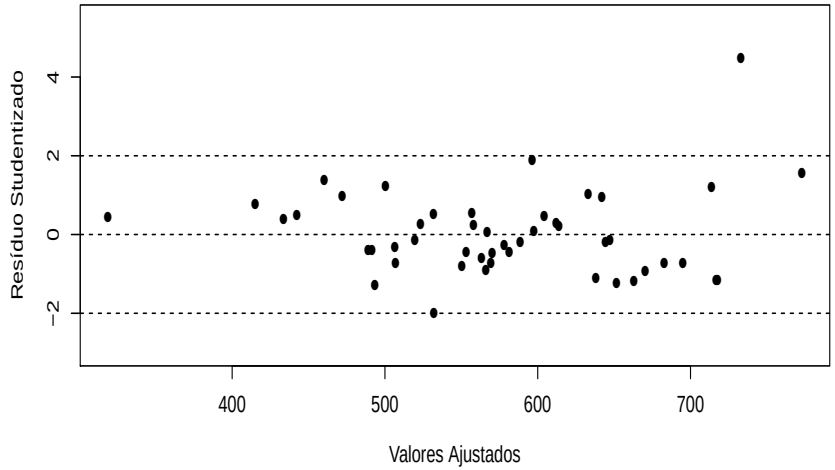
$$Y_i = \beta_0 + \beta_1 \frac{(x_{1i} - \overline{x}_1)}{s_1} + \beta_2 \frac{(x_{2i} - \overline{x}_2)}{s_2} + \beta_3 \frac{(x_{3i} - \overline{x}_3)}{s_3} + \beta_4 \frac{(x_{4i} - \overline{x}_4)}{s_4} + \xi_i,$$

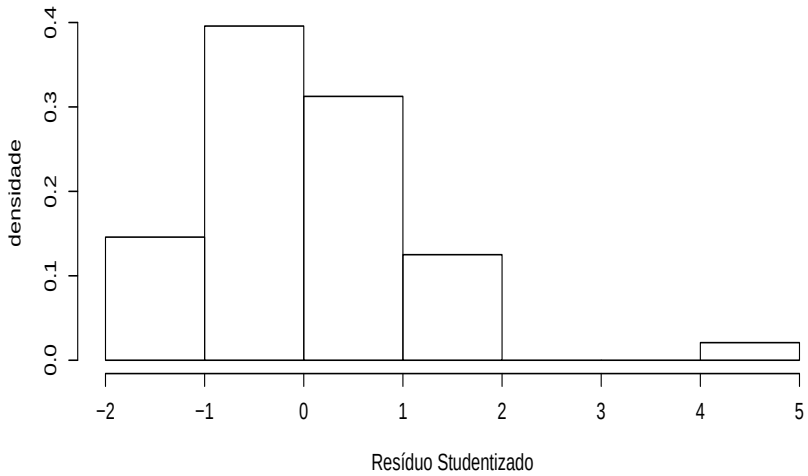
$i = 1, ..., 48, \overline{x}_j = \frac{1}{48} \sum_{i=1}^{48} x_{ij}, s_j = \sqrt{\frac{1}{47} \sum_{i=1}^{47} (x_{ij} - \overline{x}_j)^2}; j = 1, 2, 3, 4$

- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- β_0 : consumo esperado para estados com valor de cada covariável igual à sua respectiva média.
- β_j / s_j : incremento (positivo ou negativo) no consumo esperado para o aumento em uma unidade da variável $j, j = 1, 2, 3, 4$, mantendo-se as outras fixas.

Gráficos de resíduos







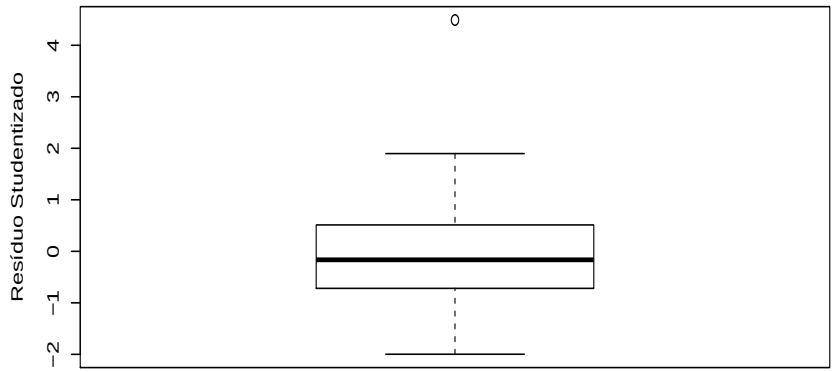
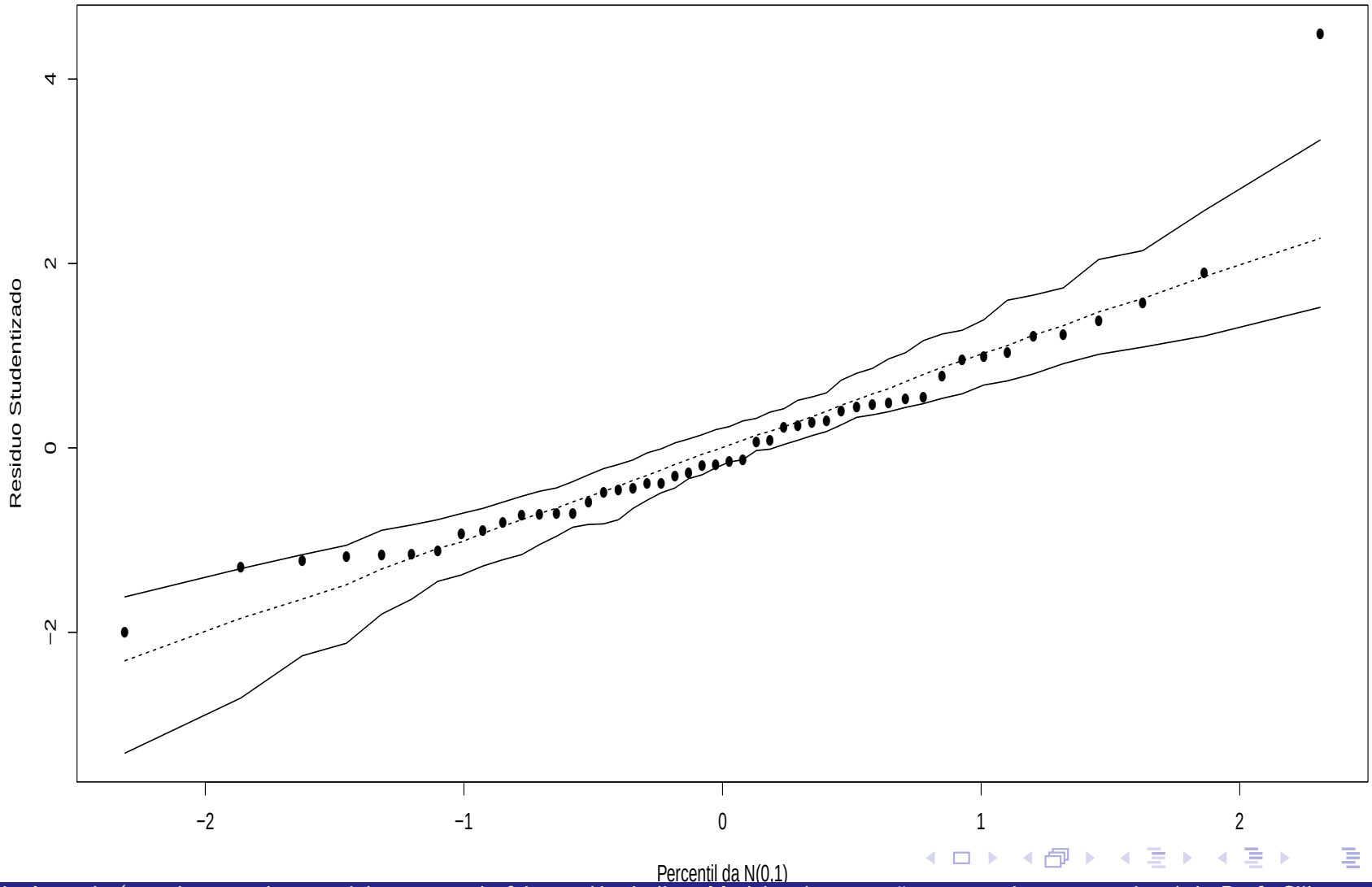


Gráfico de envelopes para os resíduos



Estimativas dos parâmetros

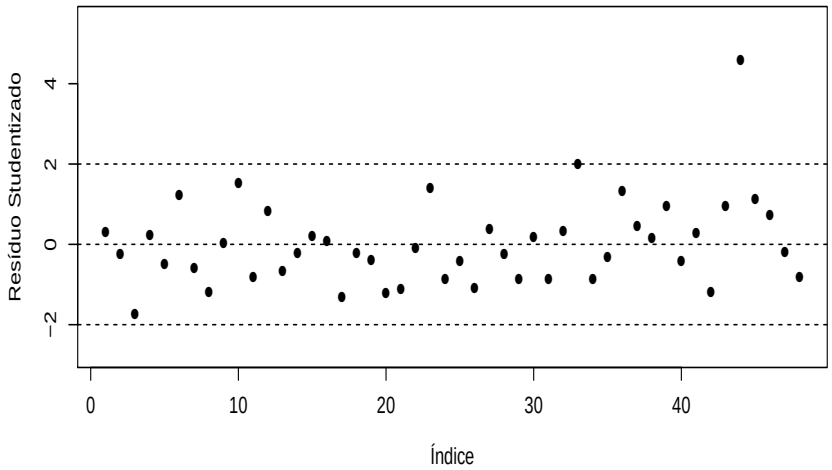
Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	576,771	9,570	[557,470 ; 596,072]	60,266	$<$0,0001
β_1	-33,077	12,332	[-57,947 ; -8,208]	-2,682	0,0103
β_2	74,133	10,667	[52,621 ; 95,645]	6,950	$<$0,0001
β_3	-38,197	9,879	[-58,119 ; -18,274]	-3,867	0,0004
β_4	-8,470	11,833	[-32,334 ; 15,394]	-0,716	0,4780

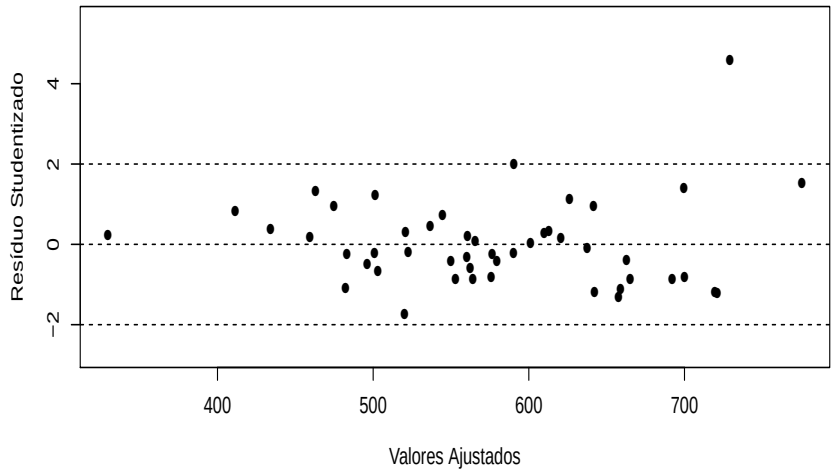
Todos os parâmetros são, aparentemente, significativos, à exceção, do parâmetro β_4 (estradas).

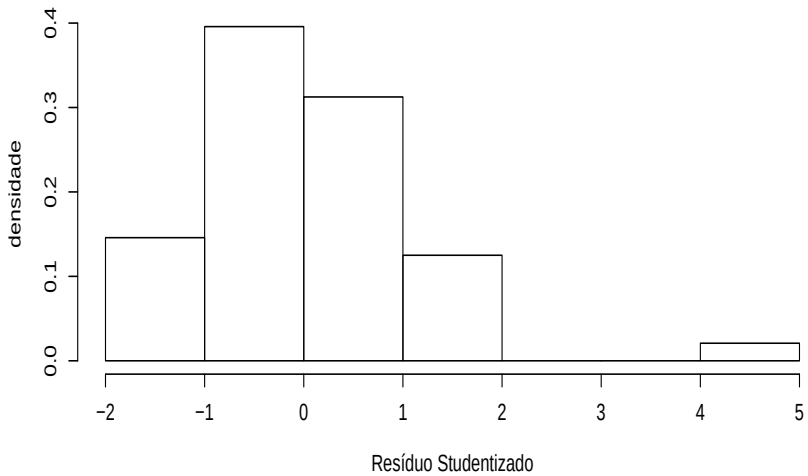
Seleção de modelos

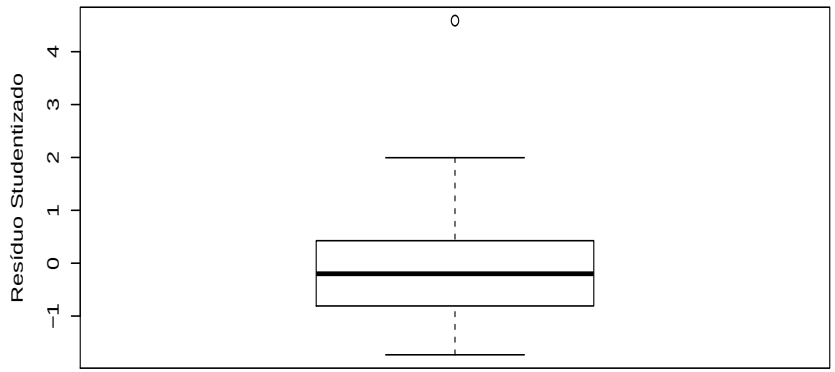
- A aplicação da metodologia stepwise, começando com o modelo só com o intercepto ou começando com o modelo completo, indicou, em ambos os casos, que a covariável "estradas" não é significativa.
- Ajustamos um modelo reduzido sem esta covariável.

Gráf. de resíduos para o modelo reduzido

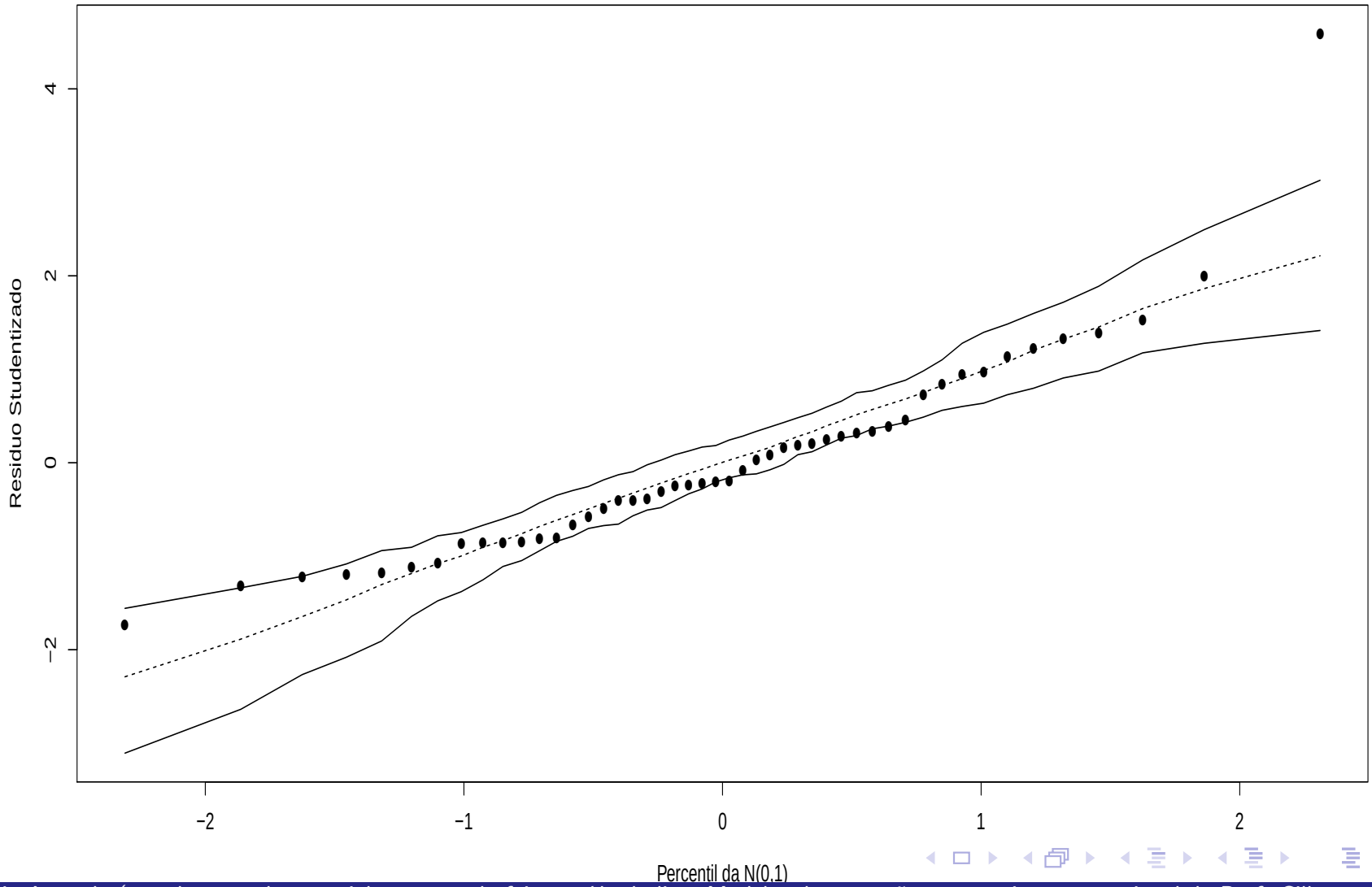








Gráf. de envelopes para os resíduos para o modelo reduzido



Estimativas dos parâmetros

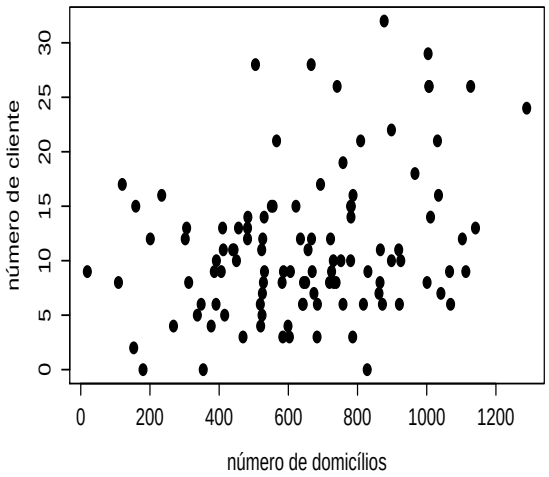
Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	576,771	9,517	[557,590 ; 595,952]	60,602	$<$ 0,0001
β_1	-28,032	10,063	[-48,312 ; -7,753]	-2,786	0,0078
β_2	76,259	10,188	[55,726 ; 96,792]	7,485	$<$ 0,0001
β_3	-39,020	9,757	[-58,684 ;-19,355]	-3,999	0,0002

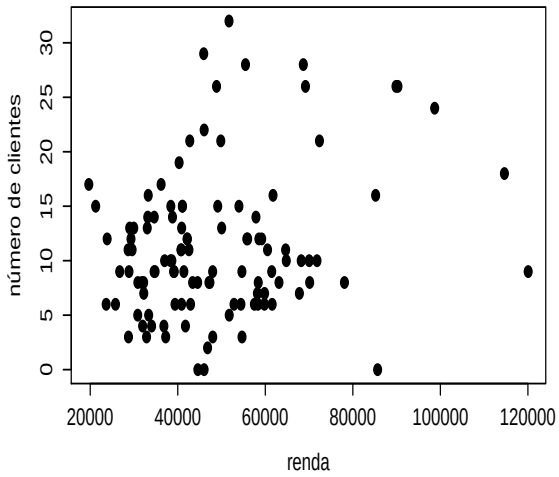
Todos os parâmetros são significativos.

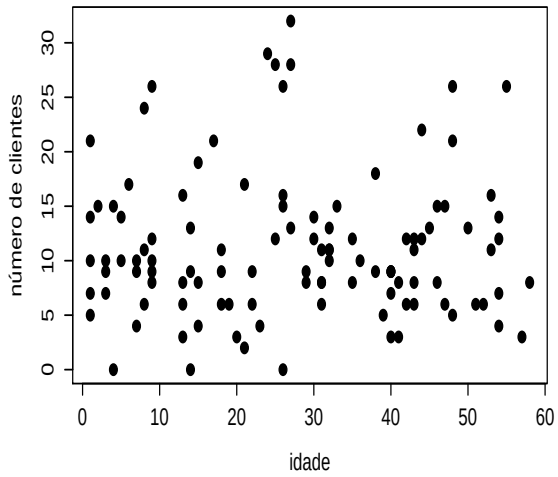
Exemplo 7: perfil dos clientes de uma loja

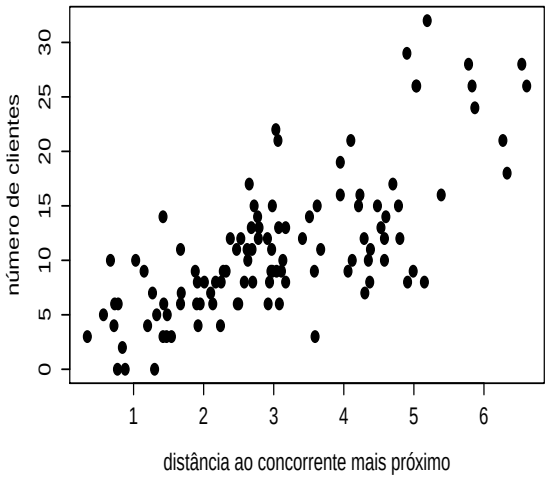
- Interesse: estudar o perfil dos clientes de uma determinada loja oriundos de 110 áreas de uma determinada cidade. Cada uma das 110 observações corresponde à uma área da cidade.
- Verificar como certas características (variáveis explicativas) afetam o número esperado de clientes em cada área (variável 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) (contagem).

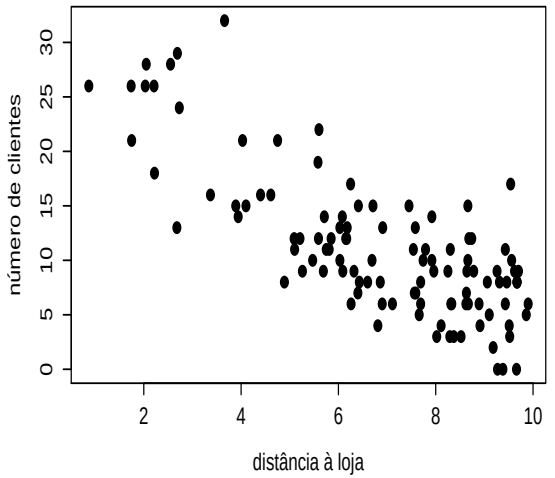
Gráficos de dispersão











Modelo (completo)

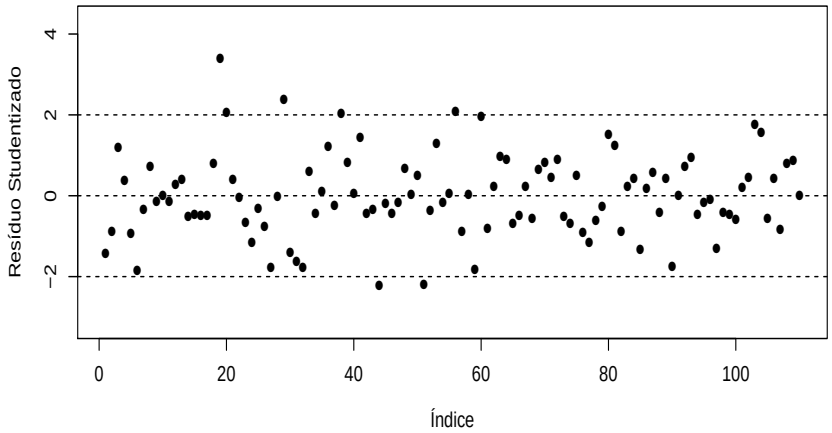
$$
\begin{aligned}
Y_i &= \mu_i + \xi_i, \xi_i \overset{i.i.d}{\sim} N(0, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \left(\frac{x_{1i} - \overline{x}_1}{s_1} \right) + \beta_2 \left(\frac{x_{2i} - \overline{x}_2}{s_2} \right) + \beta_3 \left(\frac{x_{3i} - \overline{x}_3}{s_3} \right) + \\
&+ \beta_4 \left(\frac{x_{4i} - \overline{x}_4}{s_4} \right) + \beta_5 \left(\frac{x_{5i} - \overline{x}_5}{s_5} \right), \\
\mu_i &= \exp\left\{ \beta_0 + \beta_1 \left(\frac{x_{1i} - \overline{x}_1}{s_1} \right) + \beta_2 \left(\frac{x_{2i} - \overline{x}_2}{s_2} \right) + \beta_3 \left(\frac{x_{3i} - \overline{x}_3}{s_3} \right) + \right. \\
&+ \left. \beta_4 \left(\frac{x_{4i} - \overline{x}_4}{s_4} \right) + \beta_5 \left(\frac{x_{5i} - \overline{x}_5}{s_5} \right) \right\}, i = 1, 2, ..., 110
\end{aligned}
$$

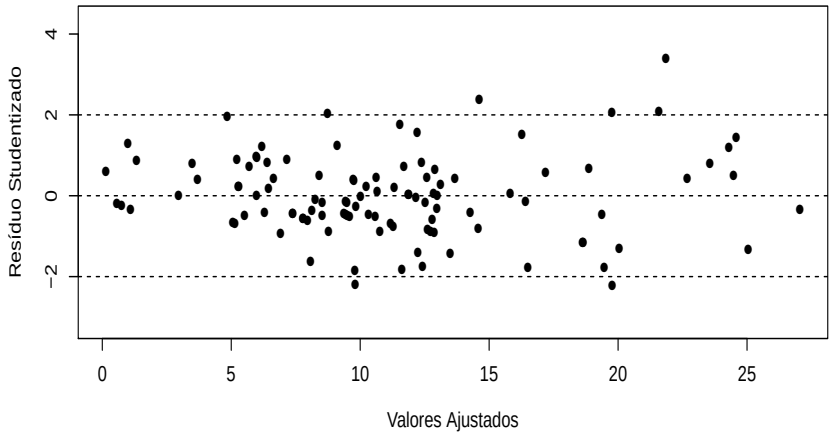
- x_{ji} : valor da variável explicativa j, $j = 1, 2, ..., 5$, associada à área i, $\overline{x}_j = \frac{1}{110} \sum_{i=1}^{110} x_{ji}$, e $s_j = \frac{\sum_{i=1}^{n} (x_{ji} - \overline{x}_j)^2}{109}$ $j = 1, 2, ..., 5$.

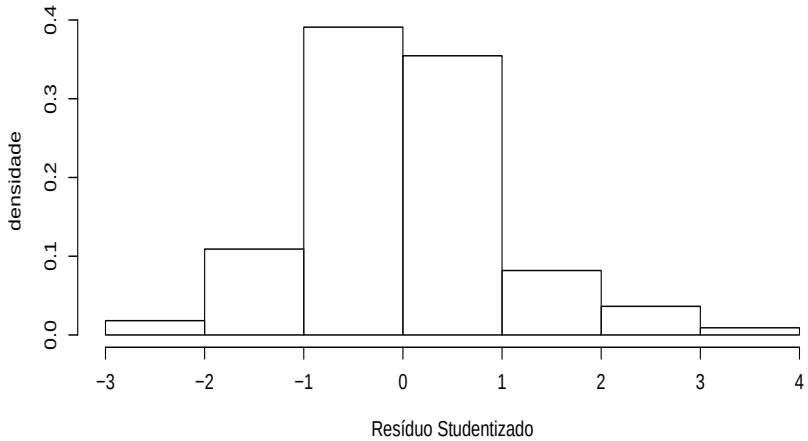
Modelo (completo)

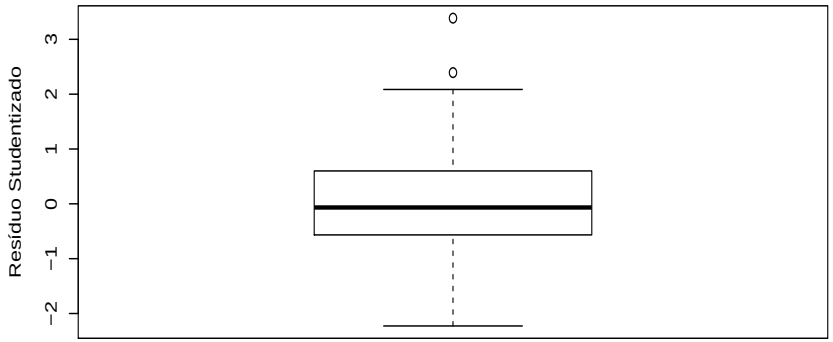
- β_0 número esperado de clientes para domicílios localizados em áreas com valor médio para cada uma das covariáveis.
- β_j / s_j : incremento (positivo ou negativo) no valor esperado do número de clientes, para o aumento em uma unidade no valor da covariável j, mantendo-se todas as outras fixas.
- Uma vez que cada uma das covariáveis está sendo introduzida no modelo com iguais média e variância (e de forma adimensional), as magnitudes dos respectivos coeficientes podem ser diretamente comparadas.

Modelo completo: gráficos de resíduos

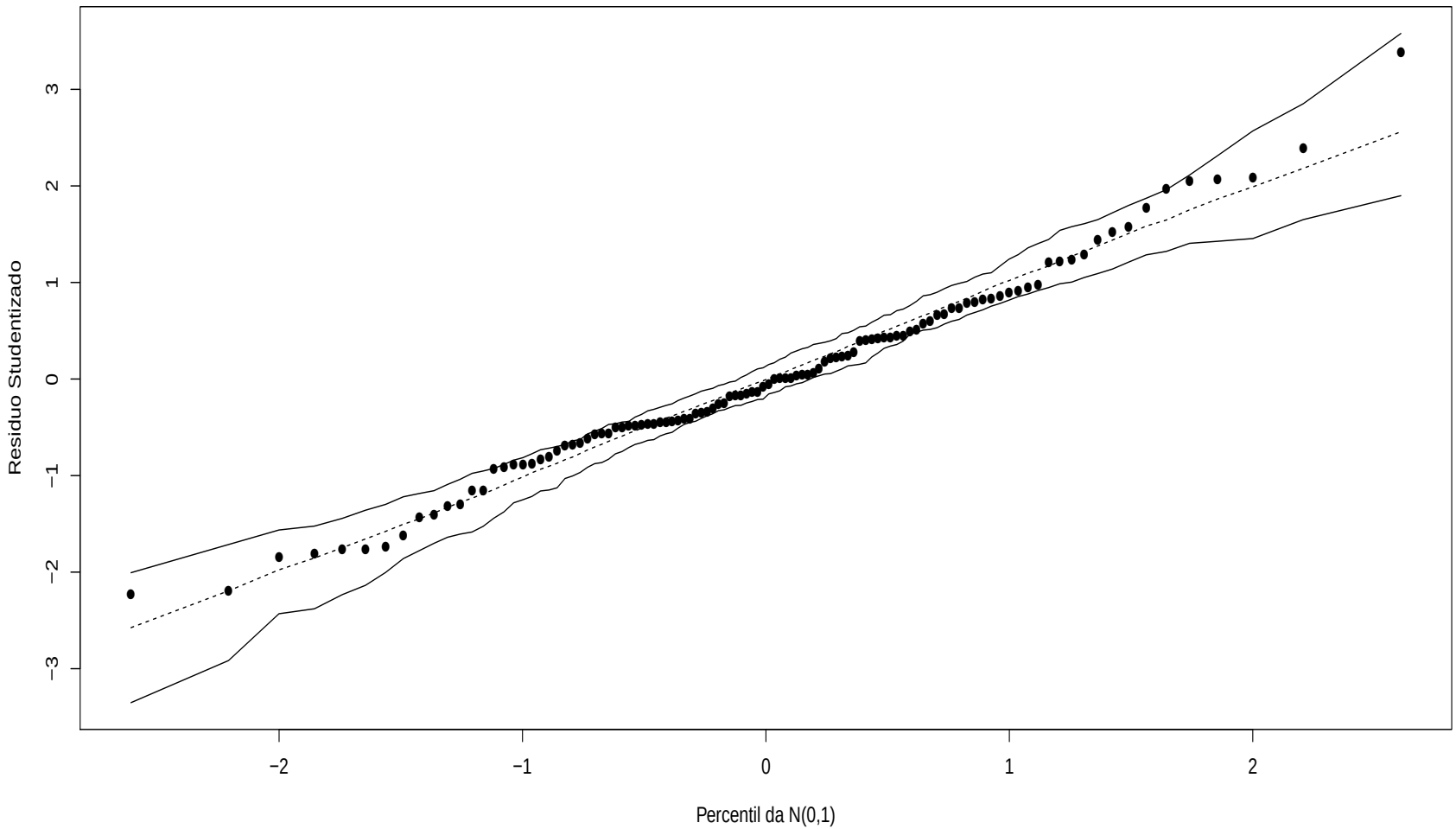








Modelo completo: gráficos de envelopes para os resíduos



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	1,736	0,400	[0,943 ; 2,530]	4,339	$<$ 0,0001
β_2	-2,153	0,423	[-2,992 ; -1,314]	-5,088	$<$ 0,0001
β_3	-0,599	0,314	[-1,221 ; 0,022]	-1,912	0,0587
β_4	2,863	0,388	[2,094 ; 3,632]	7,382	$<$ 0,0001
β_5	-3,918	0,398	[-4,708 ; -3,129]	-9,837	$<$ 0,0001

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

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.

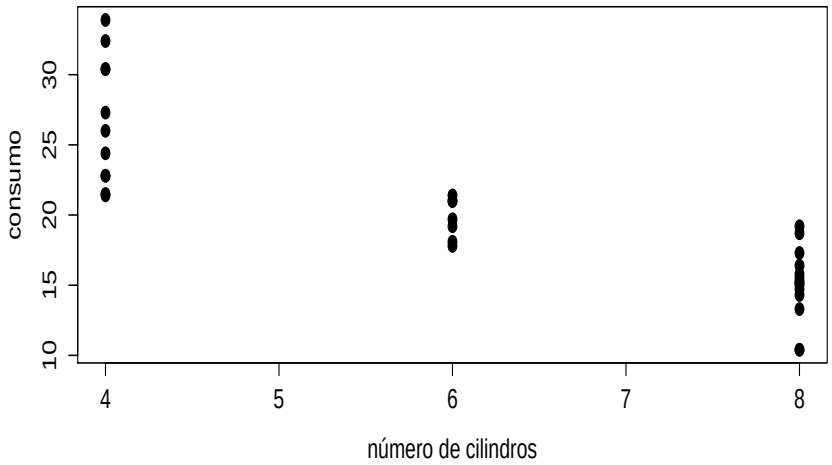
Exemplo 9: dados sobre automóveis

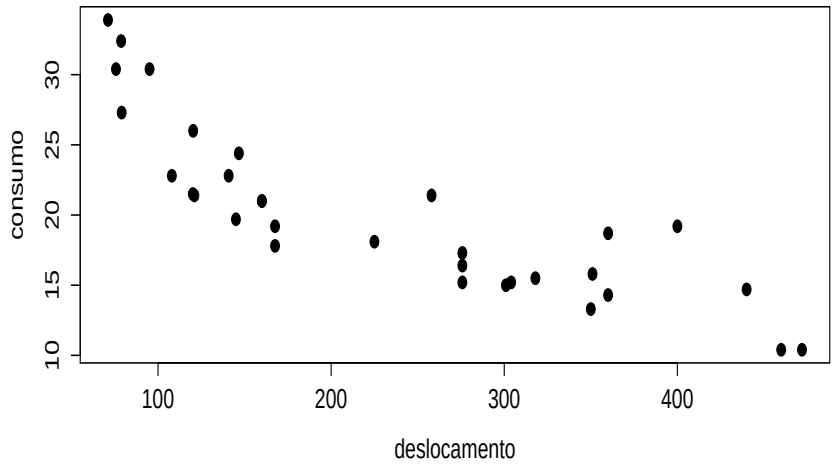
- Os dados foram extraídos da revista "1974 Motor Trend US", e compreende o consumo de combustível e 10 aspectos do design e desempenho automotivo para 32 automóveis (1973-74 modelos).
- Disponível no pacoge R sob o nome "mtcars".

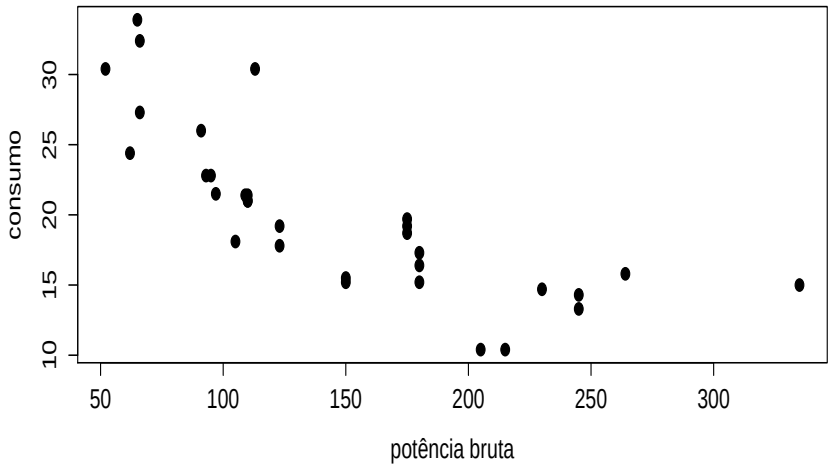
Variáveis

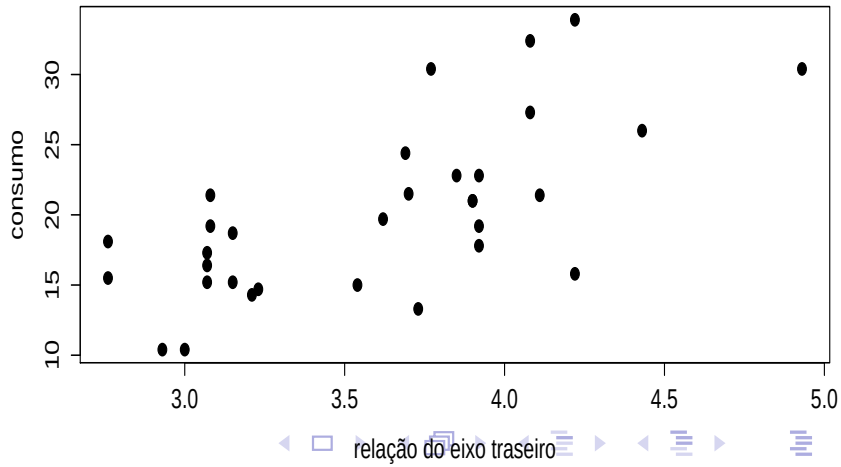
- mpg : consumo (milhas por galão) (Y).
- cyl : número de cilindros (x_1).
- disp : deslocamento (motor-volume relativos aos pistões) (x_2).
- hp : potência bruta (x_3).
- drat : relação do eixo traseiro (x_4).
- wt: peso (1000 lbs) (x_5).
- qsec : 1/4 de milha de tempo (aceleração) (x_6).
- vs: configuração do motor (0 - motor em V ; 1 - motor reto) (x_7).
- am: transmissão (0 = automático, 1 = manual) (x_8).
- gear: número de marchas dianteiras (x_9).
- carb: número de carburadores (x_{10}).

Gráficos de dispersão

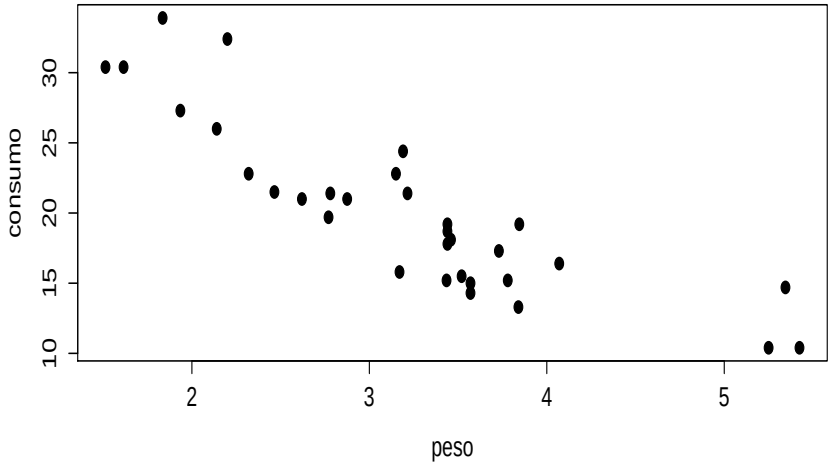


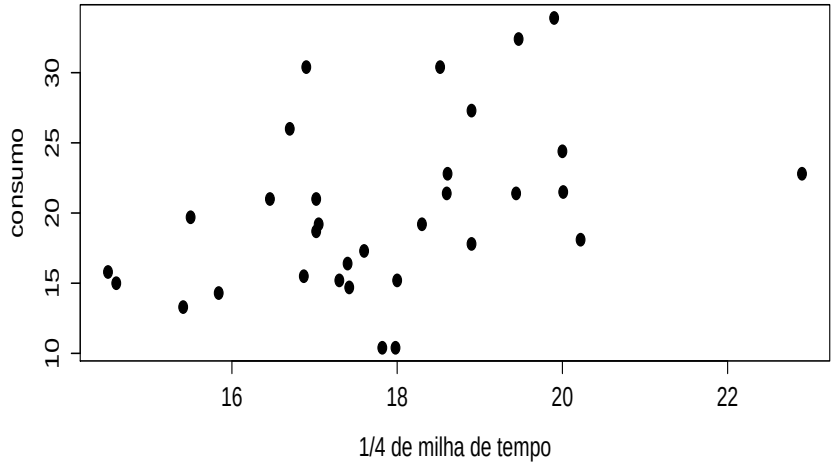


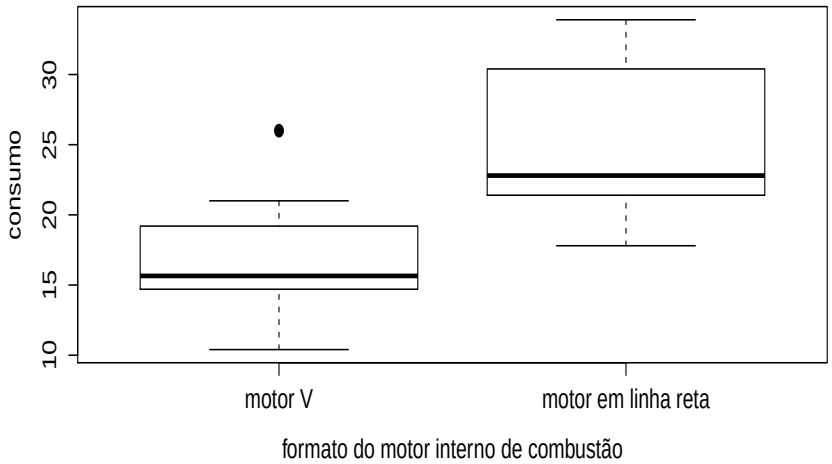


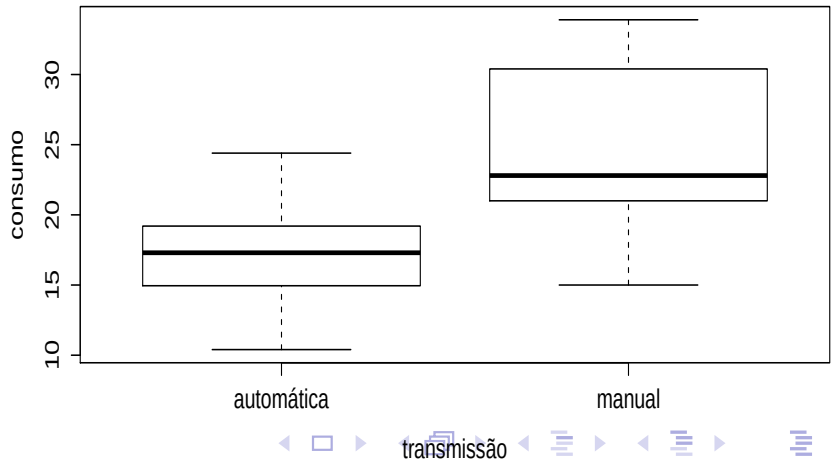


Gráficos de dispersão

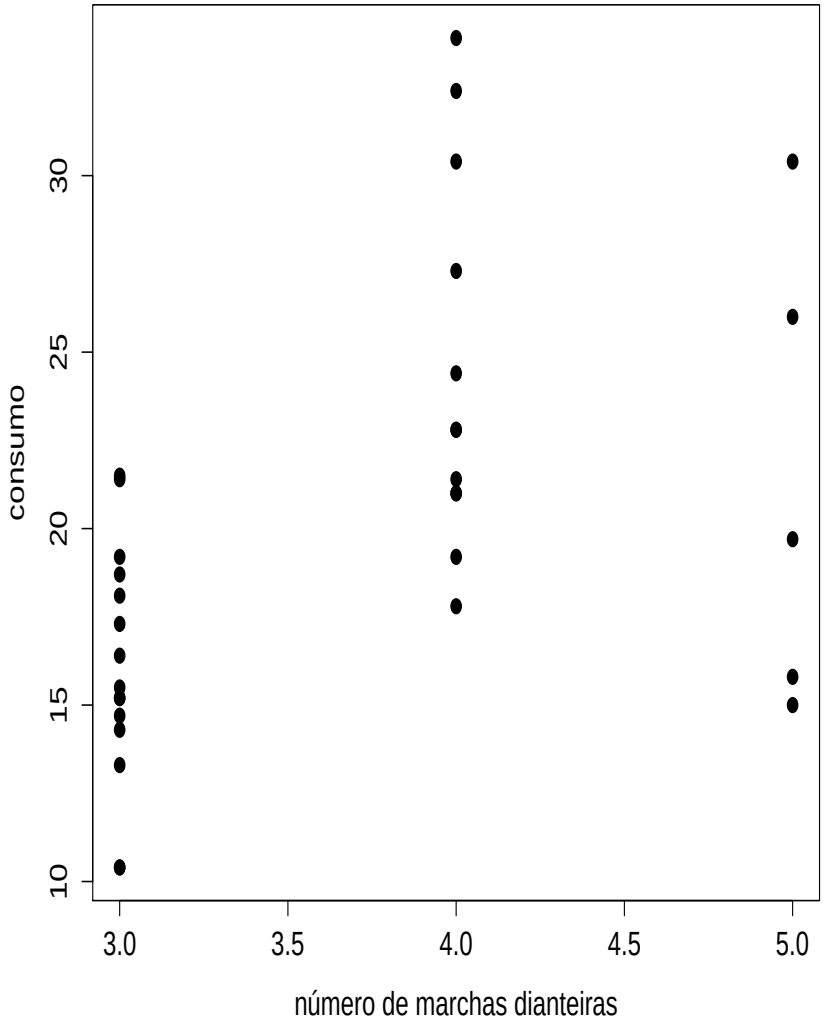


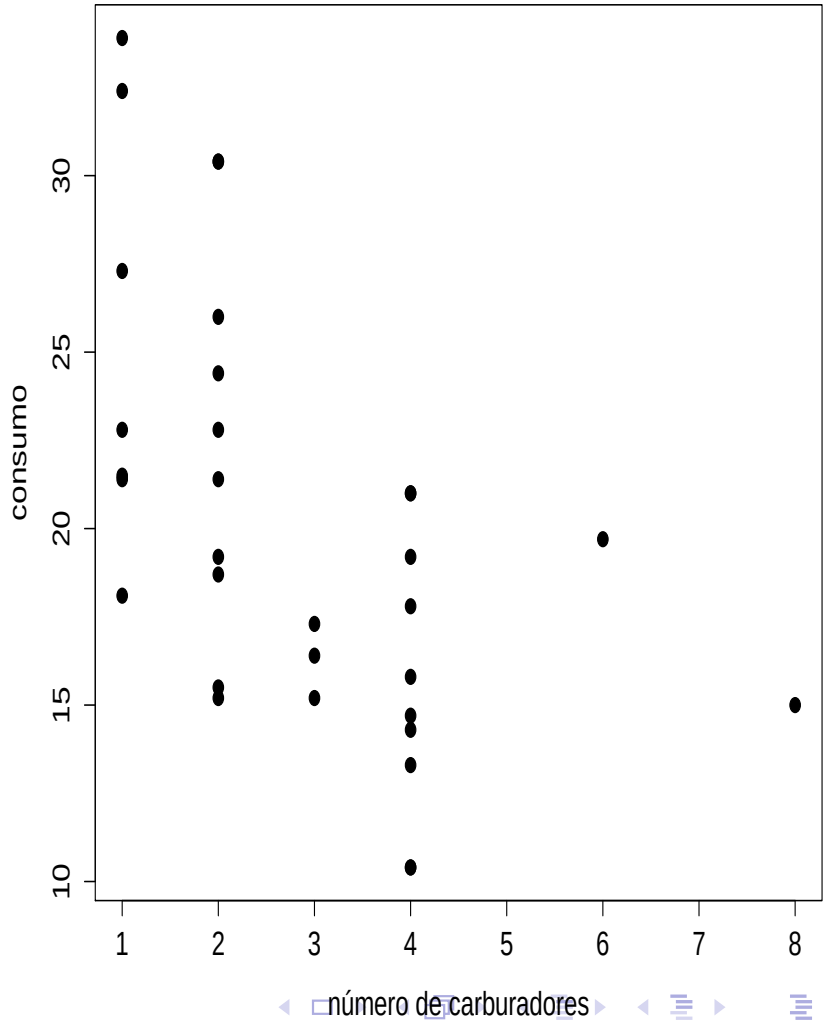






Gráficos de dispersão





Matriz de correlações

	mpg	cyl	disp	hp	drat	wt	qsec	gear	carb
mpg	1,00	-0,85	-0,85	-0,78	0,68	-0,87	0,42	0,48	-0,55
cyl	.	1,00	0,90	0,83	-0,70	0,78	-0,59	-0,49	0,53
disp	.	.	1,00	0,79	-0,71	0,89	-0,43	-0,56	0,39
hp	.	.	.	1,00	-0,45	0,66	-0,71	-0,13	0,75
drat	.	.	.	.	1,00	-0,71	0,09	0,70	-0,09
wt	.	.	.	.	.	1,00	-0,17	-0,58	0,43
qsec	.	.	.	.	.	.	1,00	-0,21	-0,66
gear	.	.	.	.	.	.	.	1,00	0,27
carb	.	.	.	.	.	.	.	.	1,00

Modelo

$$Y_i = \beta_0 + \sum_{j=1}^{10} \beta_j \frac{(x_{ji} - \overline{x}_j)}{s_j} \mathbb{1}_{\{1,2,\ldots,6,9,10\}}(j) + \sum_{j=1}^{10} \beta_j x_{ji} \mathbb{1}_{\{7,8\}}(j) + \xi_i$$

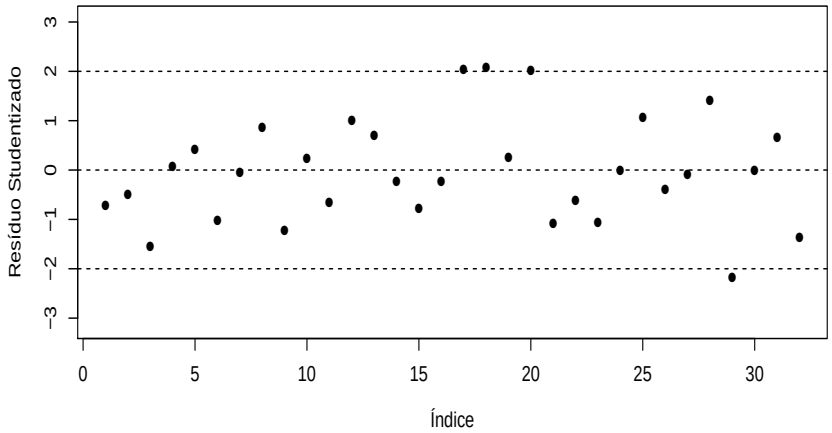
$i = 1, ..., 32, \overline{x}_j = \frac{1}{32} \sum_{i=1}^{32} x_{ij}, s_j = \sqrt{\frac{1}{31} \sum_{i=1}^{31} (x_{ij} - \overline{x}_j)^2}; j = 1, 2, ..., 6, 9, 10$, $x_{i7} = 1$, se o carro tiver motor reto e 0 se tiver motor V, $x_{i8} = 1$ se o carro tiver transmissão manual e 0 se automática.

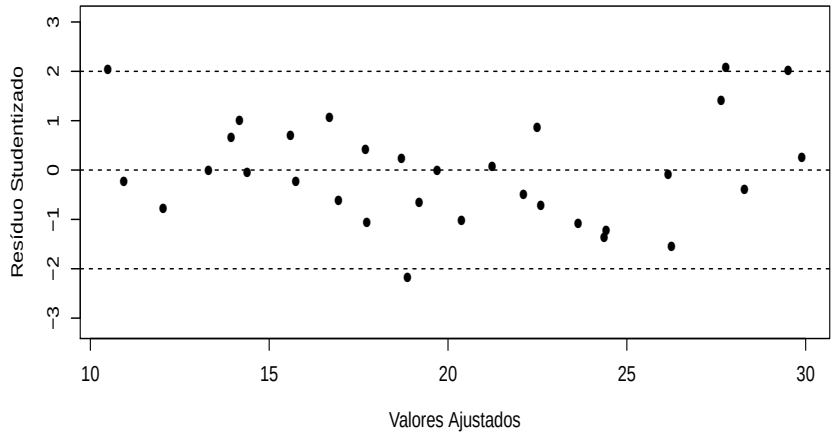
- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- β_0 : consumo esperado para carros com valor de cada covariável quantitativa igual à sua respectiva média com motor reto e transmissão automática.

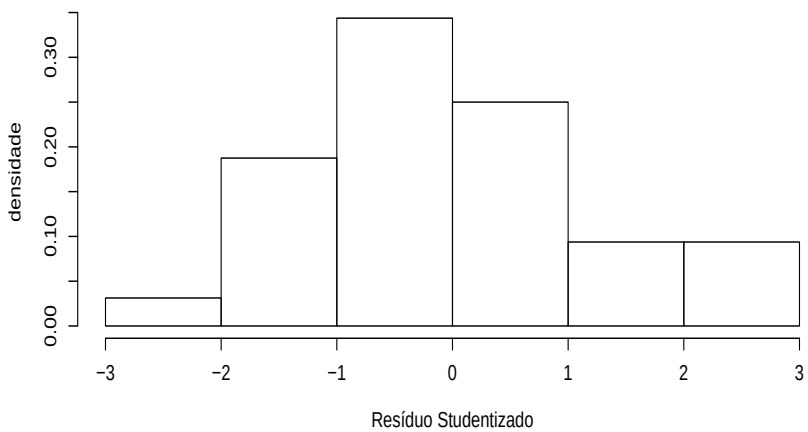
Modelo

- β_j/s_j : incremento no consumo esperado para o aumento em uma unidade da variável j, mantendo-se as outras fixas entre as quantitativas.
- β_7/s_7 : incremento no consumo esperado para carros com motor reto em relação àqueles com motor V, mantendo-se as outras covariáveis fixas.
- β_8/s_8 :incremento no consumo esperado para carros transmissão manual em relação àqueles com transmissão automática, mantendo-se as outras covariáveis fixas.
- Ajuste do modelo completo: $R^2 = 0,869$ e $\overline{R}^2 = 0,807$.

Gráficos de resíduos







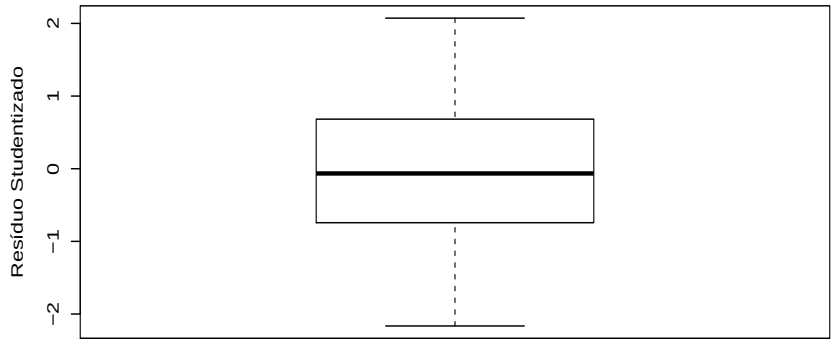
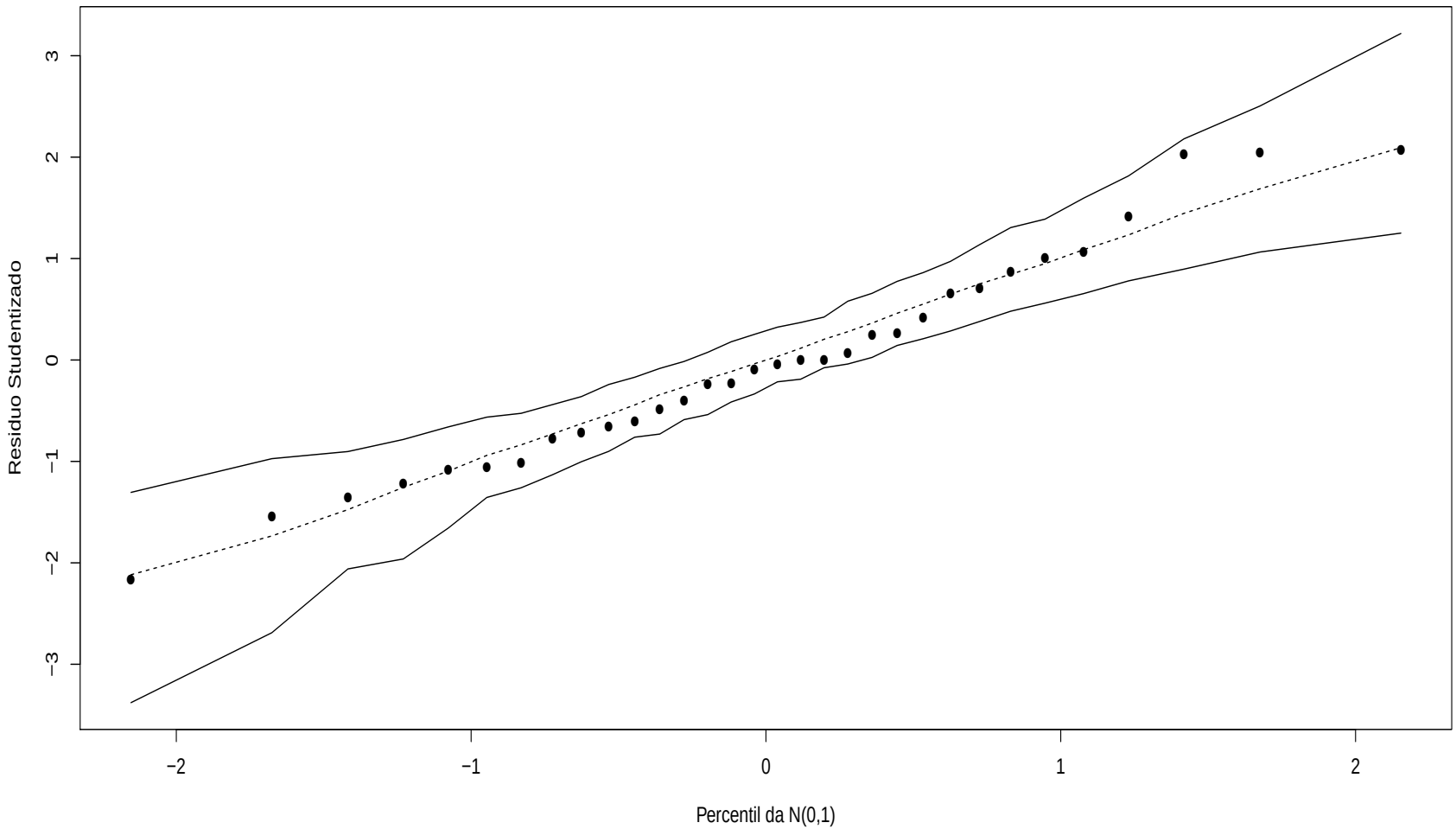
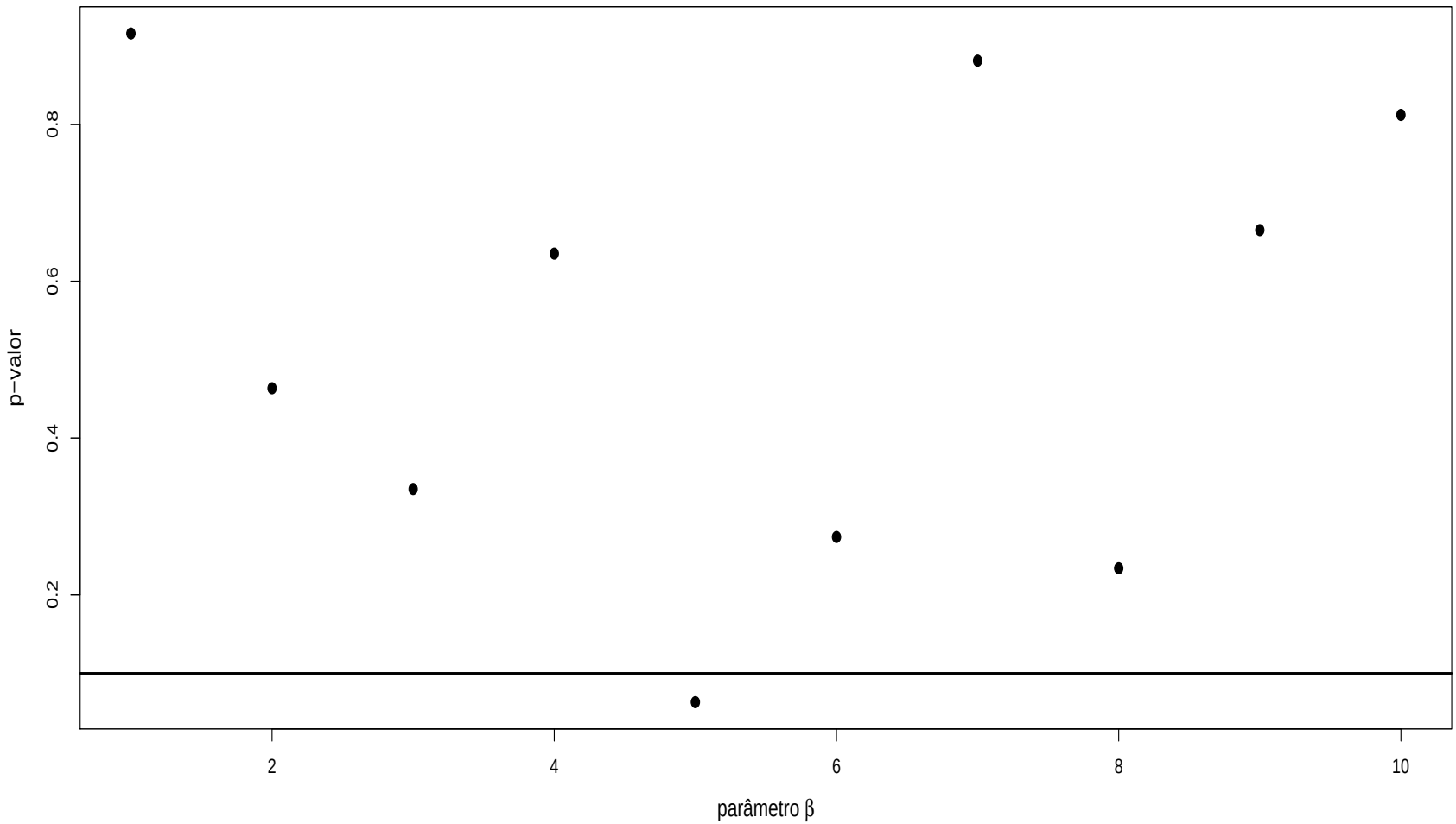


Gráfico de envelopes para os resíduos



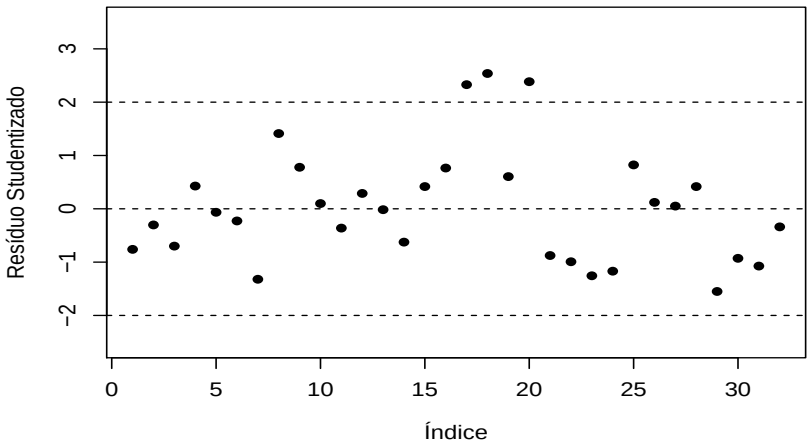
p-valores associados às estimativas dos par. de regressão

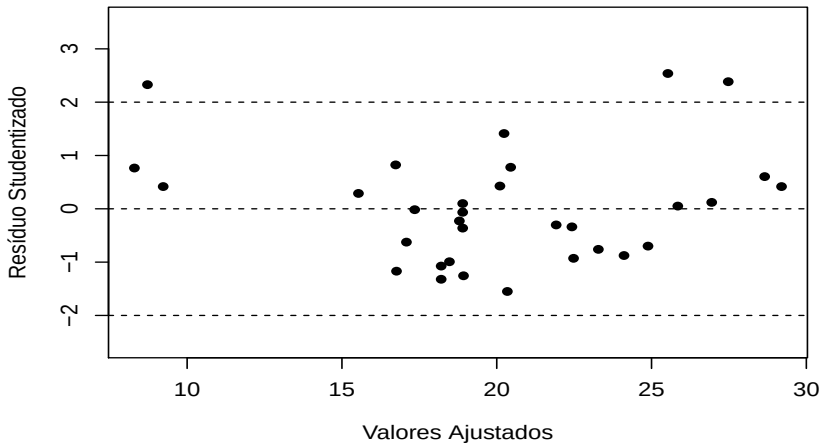


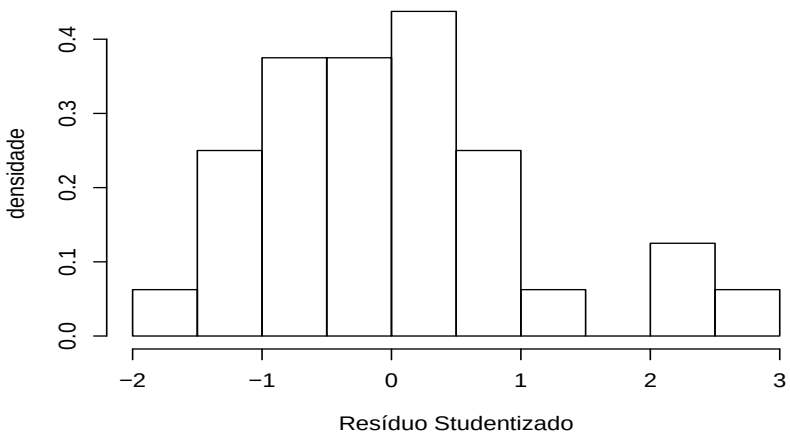
Modelo completo e reduzido

- Modelo completo: $R^2 = 0,87$ e $\overline{R}^2 = 0,81$
- Modelo reduzido (MR): Ajustou-se um modelo com somente a covariável wt (peso), a qual mostrou-se significativa -5,23 (0,55) (p-valor $<$ 0,0001).
- Além disso: $R^2 = 0,75$ e $\overline{R}^2 = 0,74$.

Gráficos de resíduos (MR)







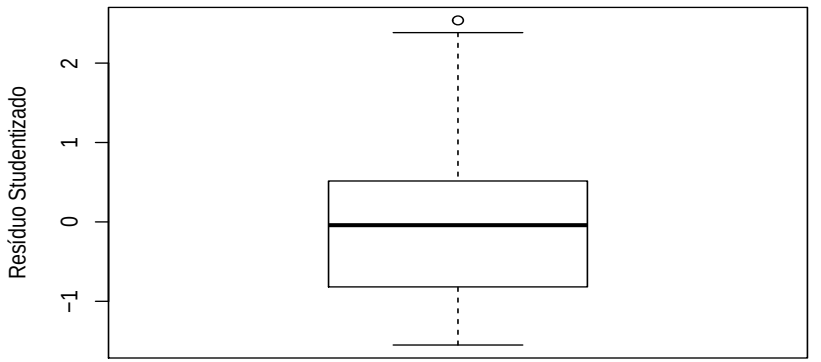
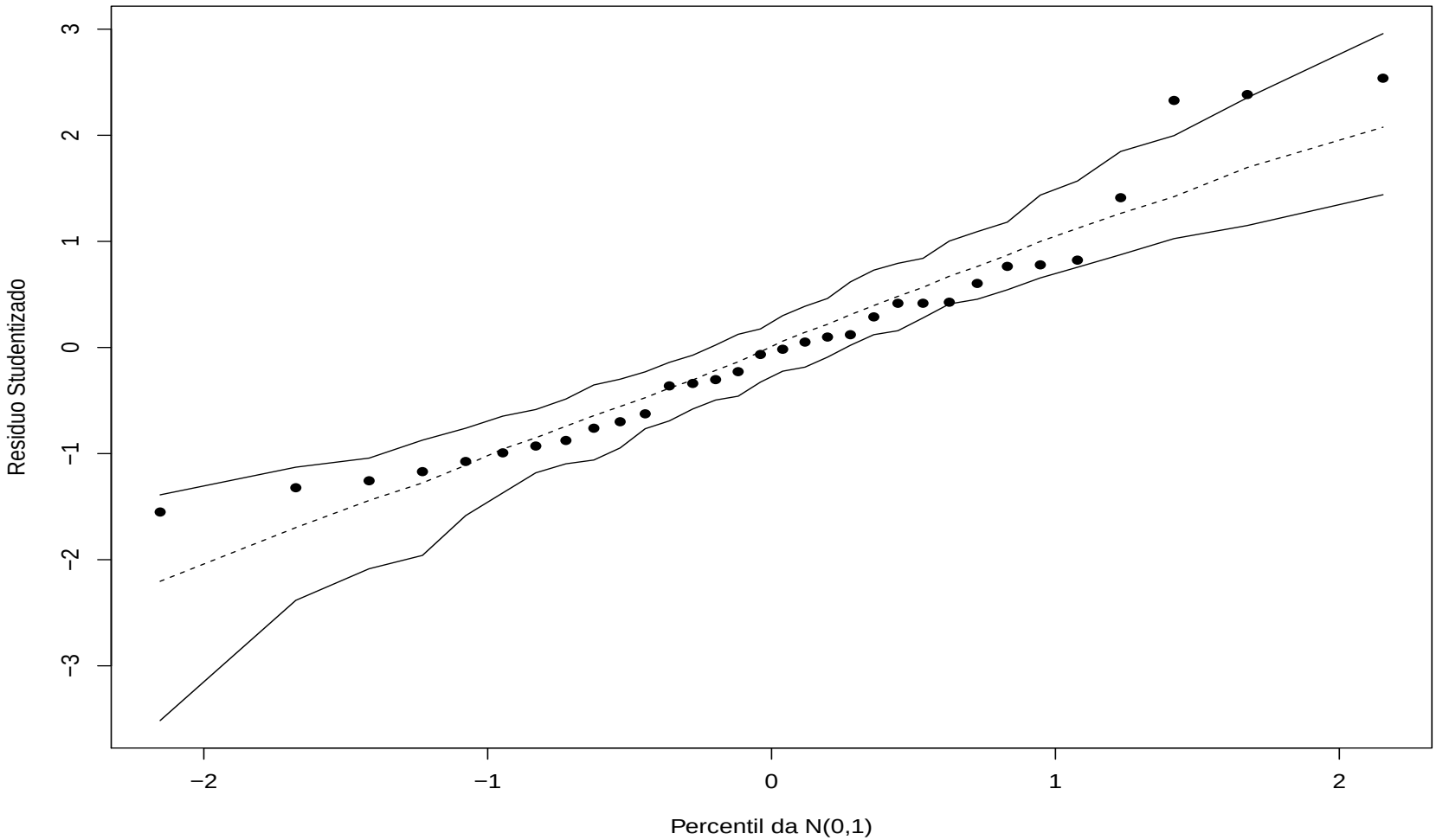


Gráfico de envelopes para os resíduos (MR)



Seleção de modelos

- A aplicação da metodologia stepwise, começando com o modelo só com o intercepto indicou que as covariáveis relevantes são: número de cilindros, potência bruta e peso (modelo 1).
- A aplicação da metodologia stepwise começando com o modelo completo, indicou que as covariáveis são: peso, 1/4 de milha de tempo e transmissão (modelo 2).

Seleção de modelos

- Estatísticas de comparação de modelos:

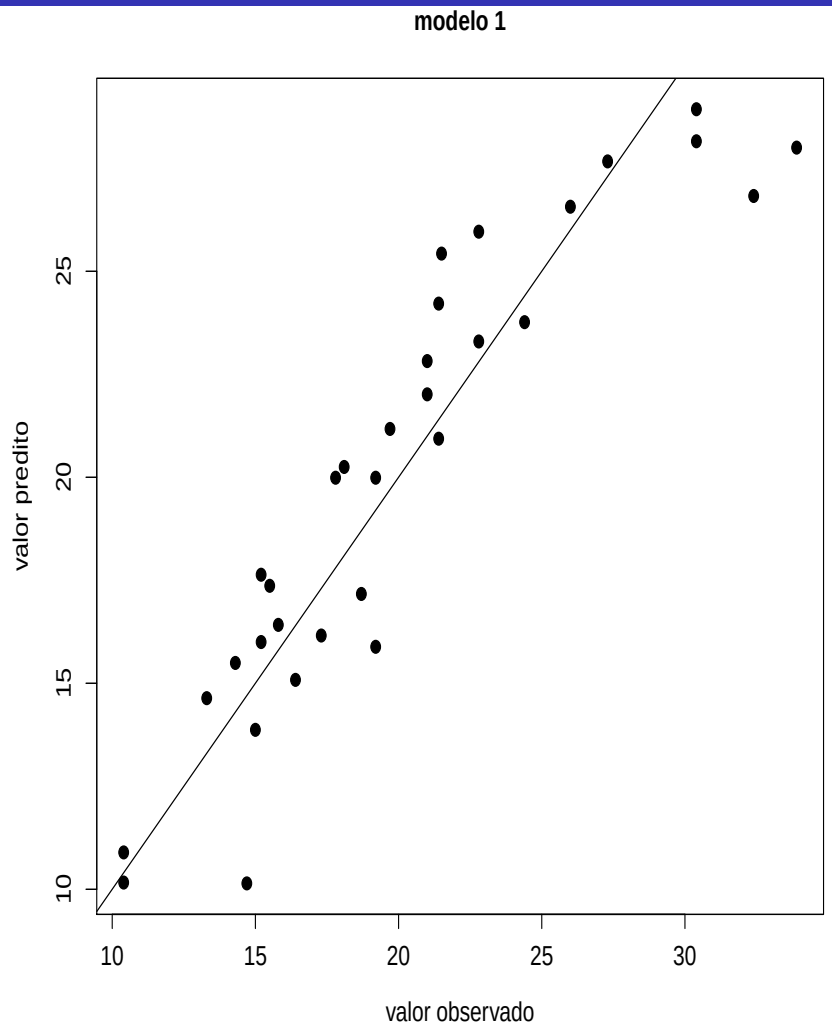
modelo	AIC	BIC	R2	R2 ajustado
1	155,48	155,48	0,84	0,83
2	154,12	154,12	0,85	0,83

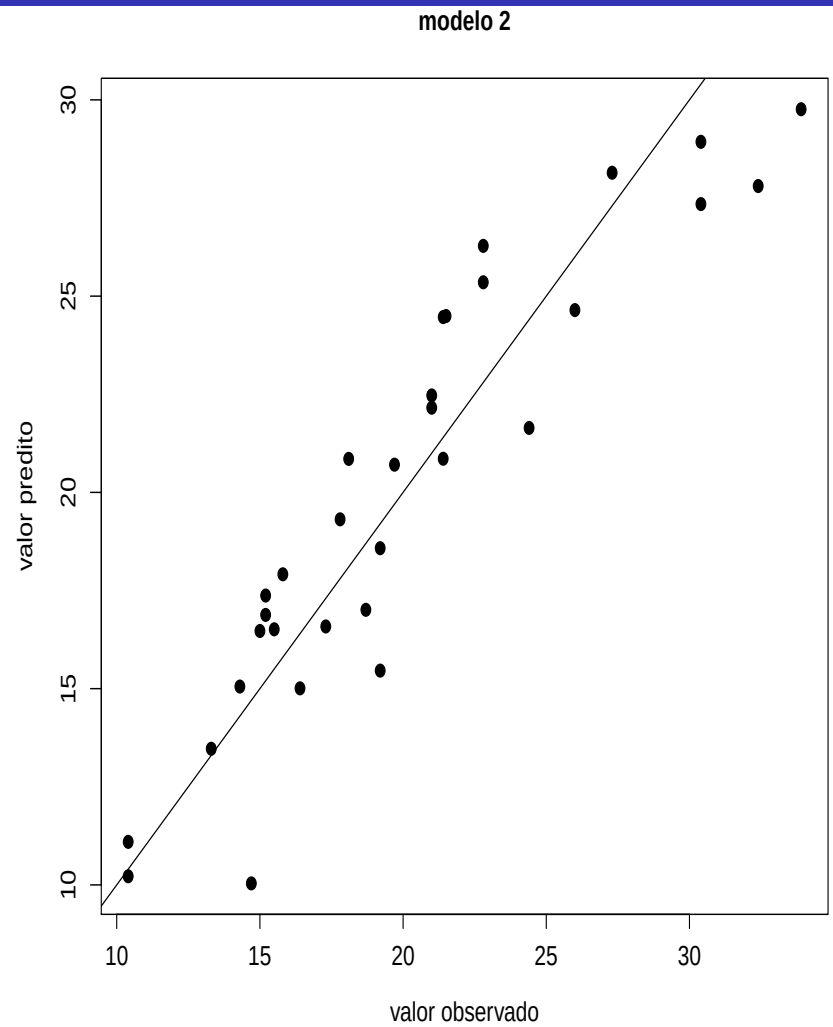
- Média do vício absoluto ponderado. $MVP = \frac{1}{32}\sum_{i=1}^{32}\frac{|y_i - \widetilde{y}_i|}{y_i}$.
- $MVP_1/MVP_2 = 95,91\%$.

Seleção de modelos

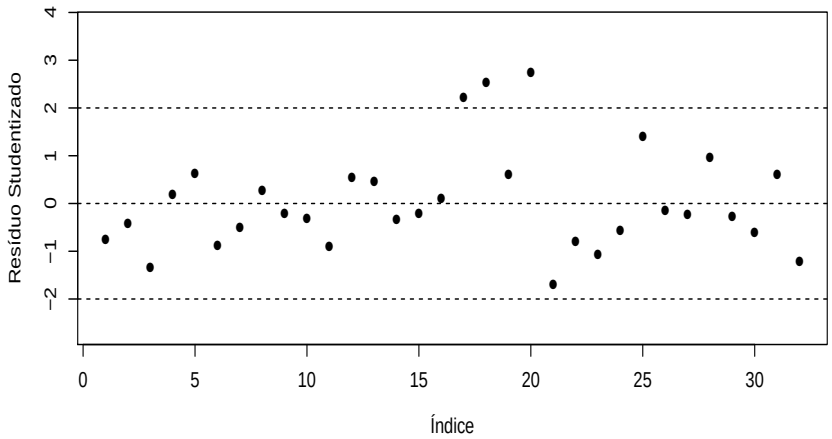
- $MVP_1/MVP_2 = 95,91\%$. Modelo 1: $R^2 = 0,84$ e $\overline{R}^2 = 0,83$ e Modelo 2: $R^2 = 0,85$ e $\overline{R}^2 = 0,83$.
- Pelos resultados da análise residual bem como das estatísticas de comparação de modelos, optou-se pelo modelo e (note que, apesar do modelo original ter apresentado uma adequabilidade melhor, através da análise residual, apenas uma única covariável fora significativa).

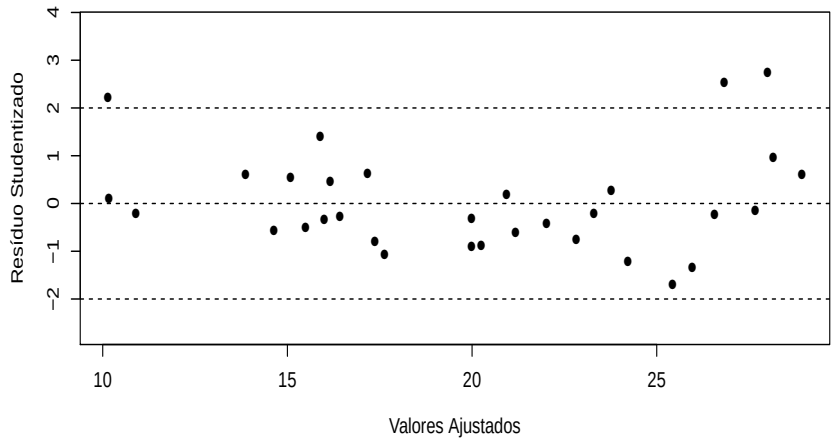
Valores preditos e observados

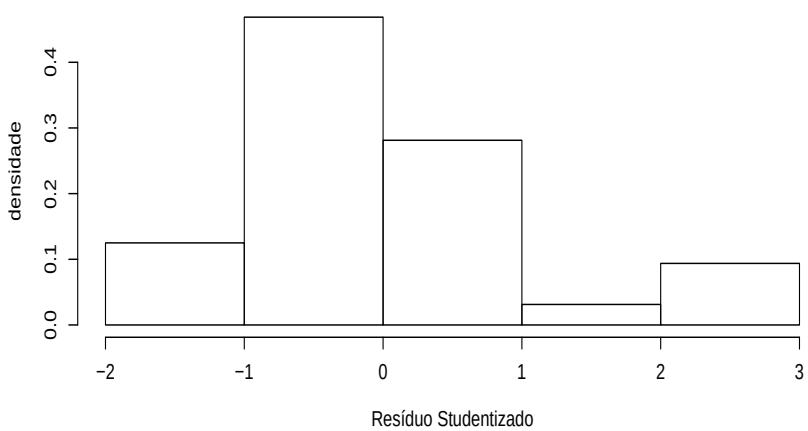




Gráficos de resíduos (modelo 1)







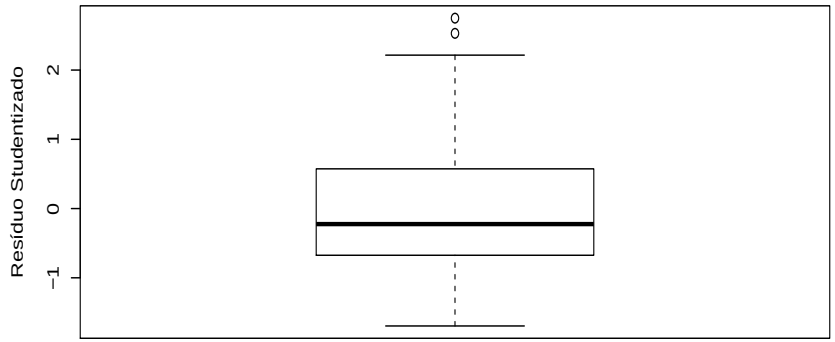
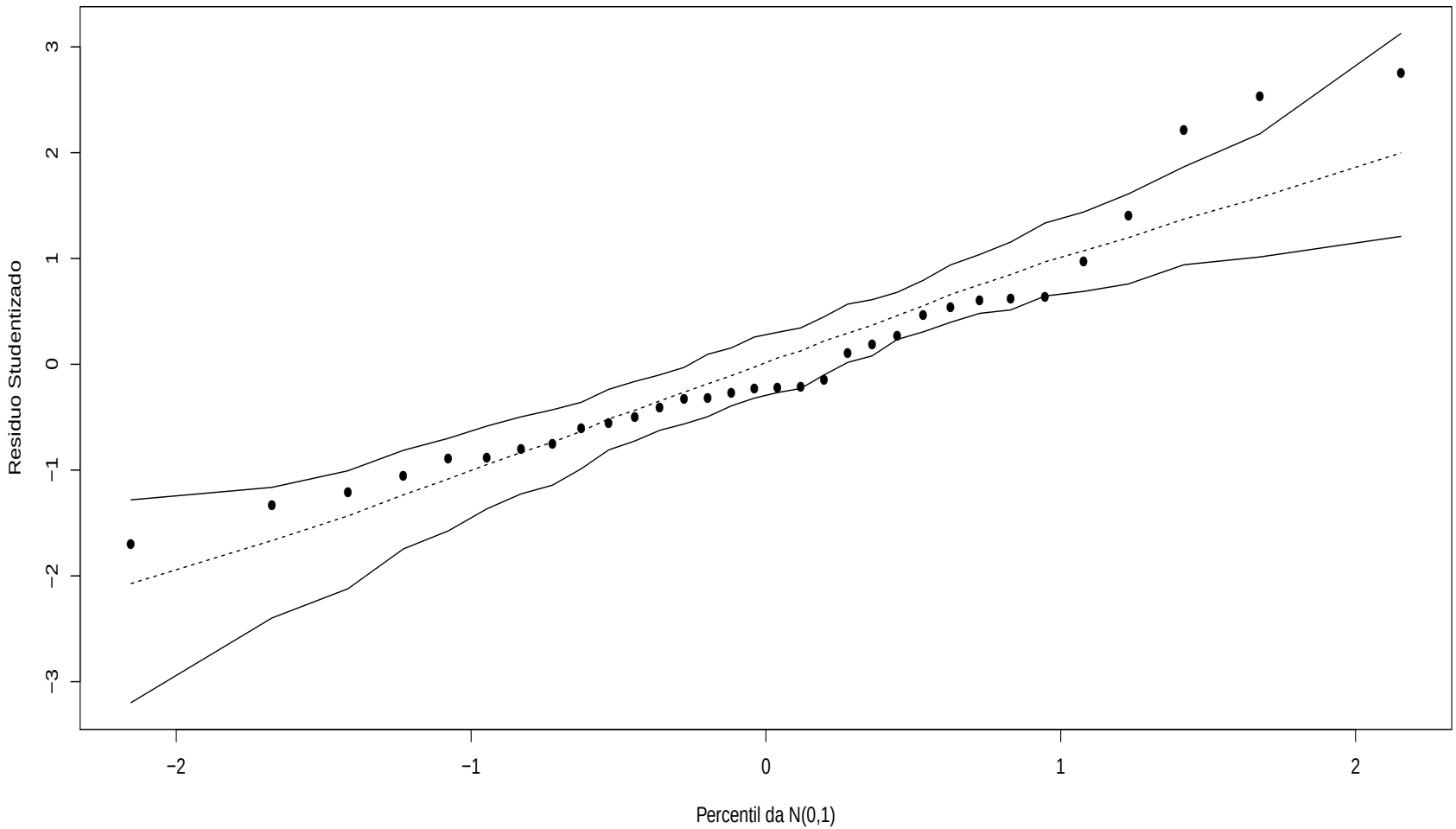
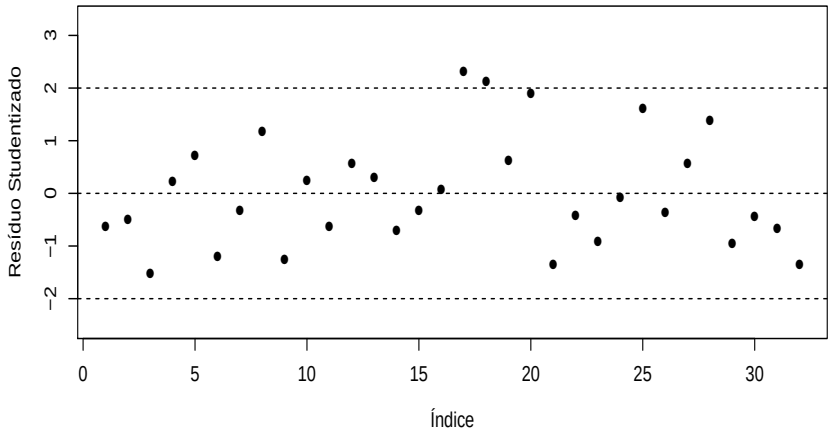
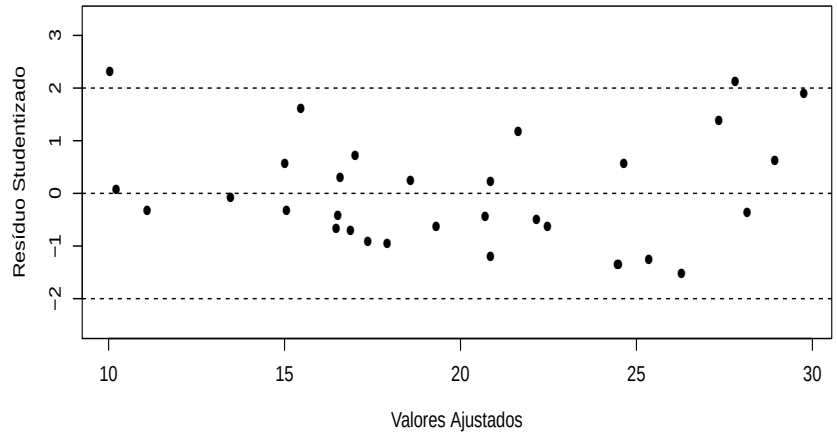


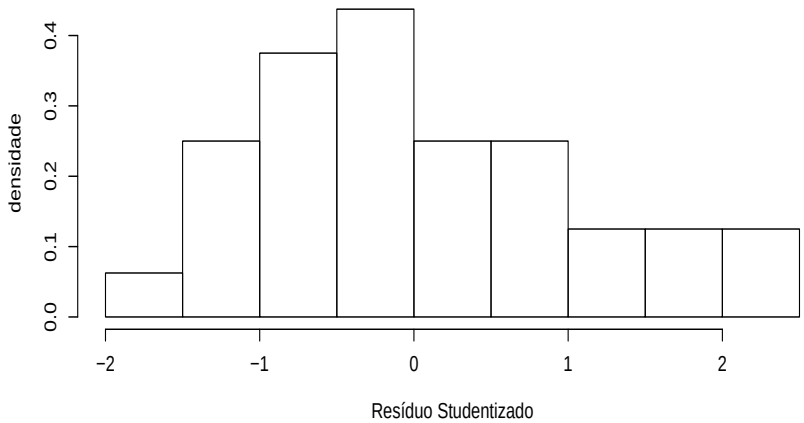
Gráfico de envelopes para os resíduos (modelo 1)



Gráficos de resíduos (modelo 2)







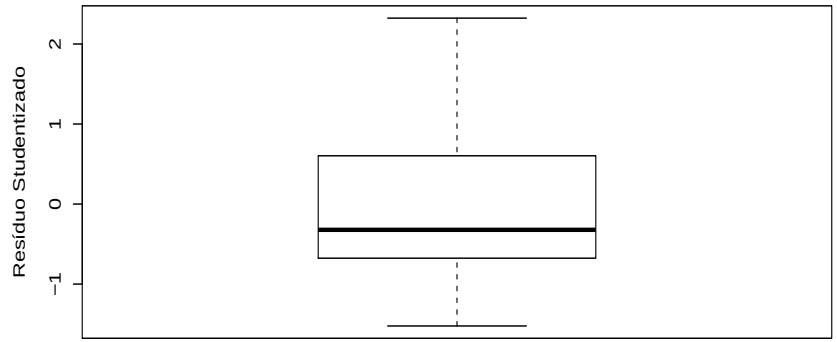
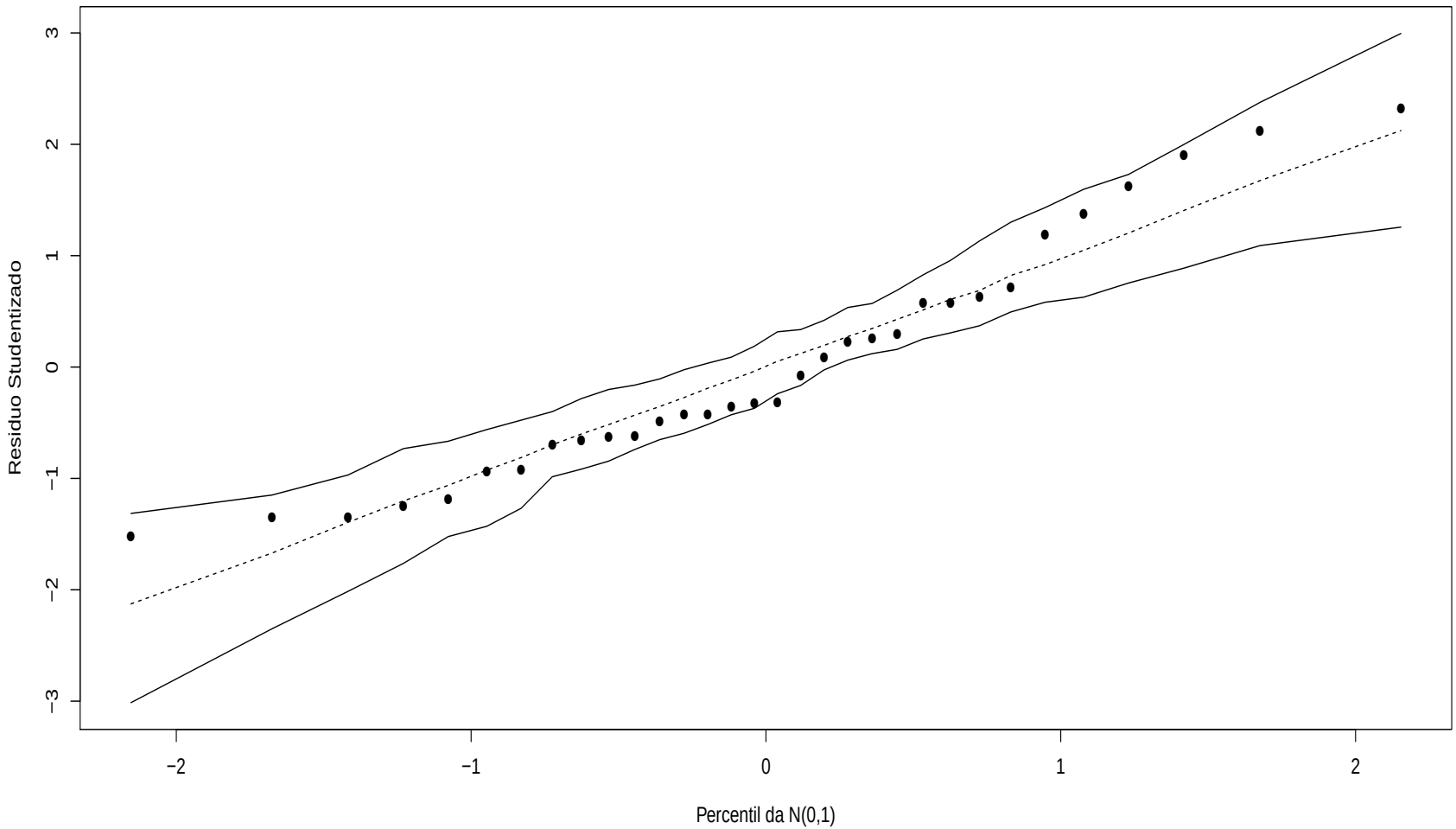


Gráfico de envelopes para os resíduos (modelo 2)



Resultados relativos ao modelo 2

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	18,898	0,719	[17,424 ; 20,371]	26,271	<0,0001
β_5	-3,832	0,696	[-5,258 ; -2,407]	-5,507	<0,0001
β_6	2,191	0,516	[1,134 ; 3,247]	4,247	0,0002
β_8	2,936	1,411	[0,046 ; 5,826]	2,081	0,0467

Quanto maior o peso e menor a aceleração, maior o consumo. Por outro lado, carros com câmbio manual consomem, em média, menos combustível.