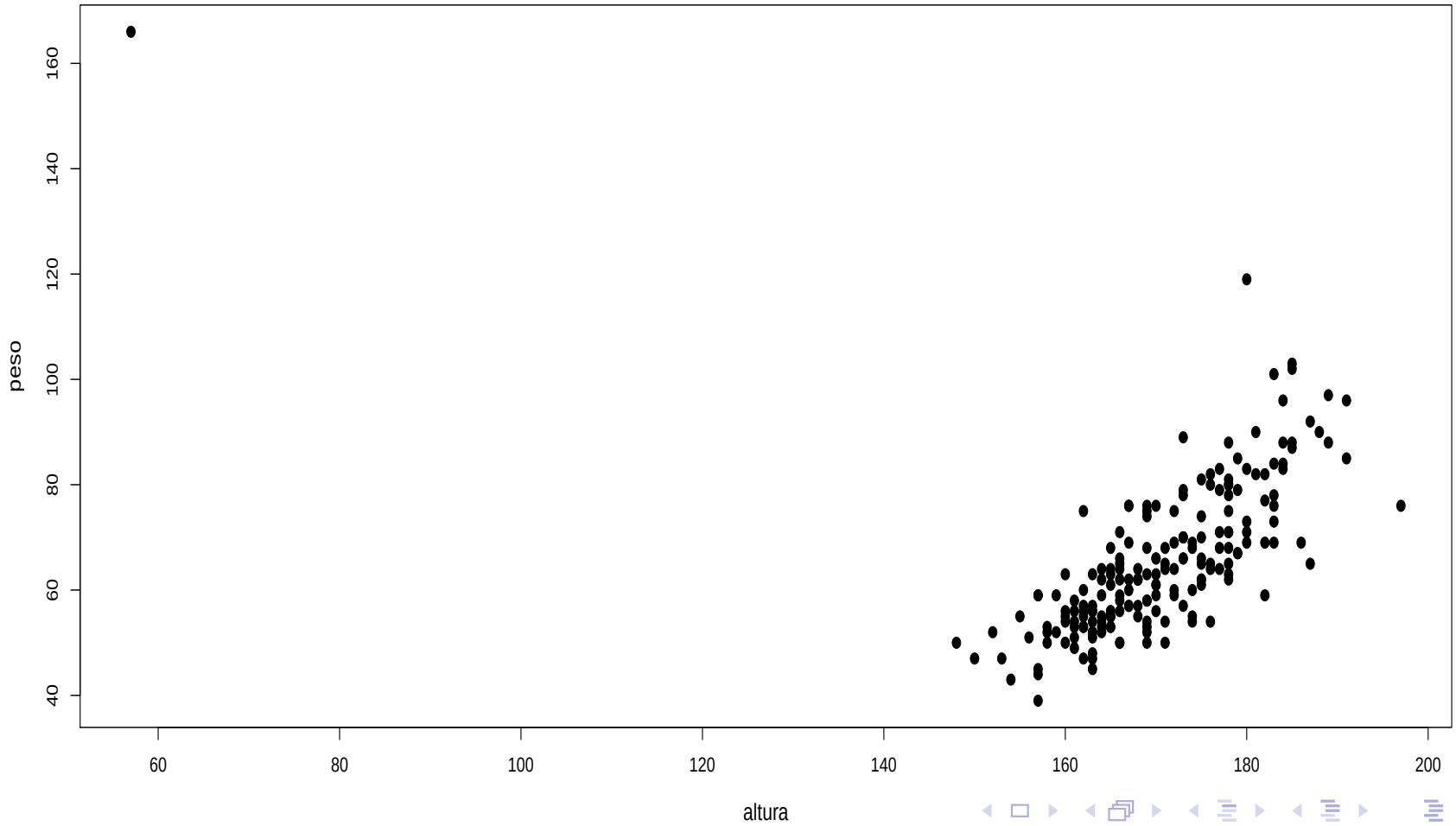
Análise de dados e métodos de diagnóstico em modelos de regressão normais lineares (parte 3)

Prof. Caio Azevedo

Exemplo 0: altura e peso de homens e mulheres

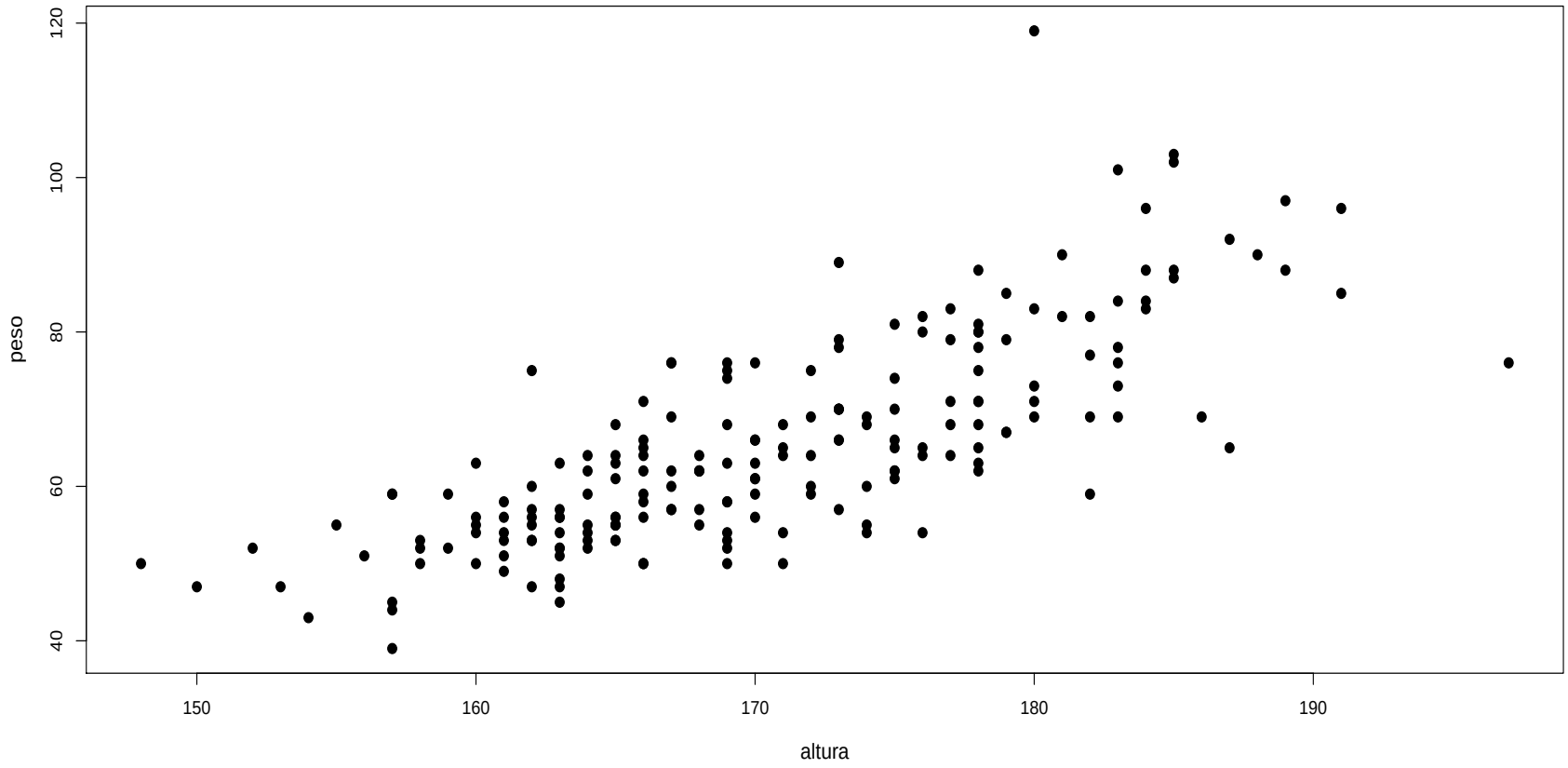
- Os dados correspondem aos pesos (em kg) e alturas (em cm) medidos e informados de 200 indivíduos.
- O sexo de cada indivíduo também foi coletado, sendo 112 mulheres e 88 homens.
- Este conjunto de dados está disponível em no R no pacote "car" sob o nome "Davis".

Exemplo 0: altura e peso de homens e mulheres



Exemplo 0: altura e peso de homens e mulheres

($\widetilde{\rho} = 0,7707$) sem a observação discrepante

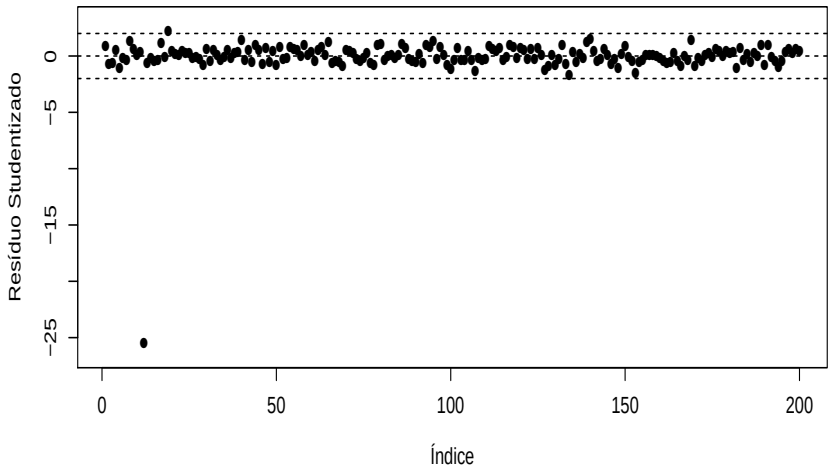


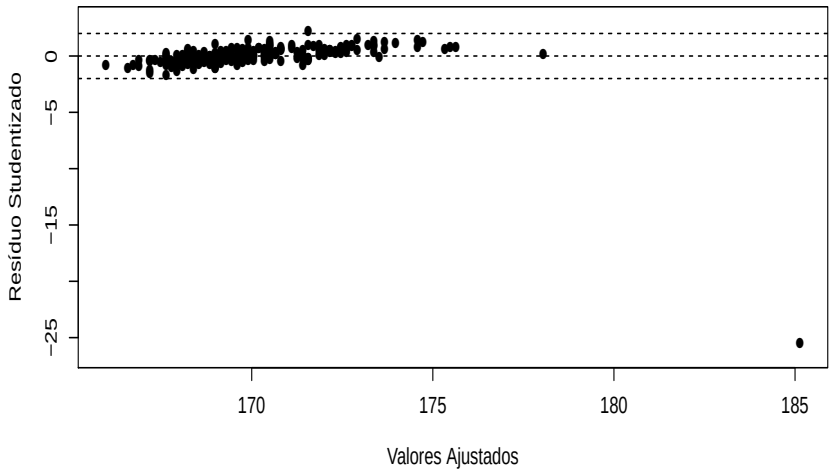
Exemplo 0: sem considerar o sexo

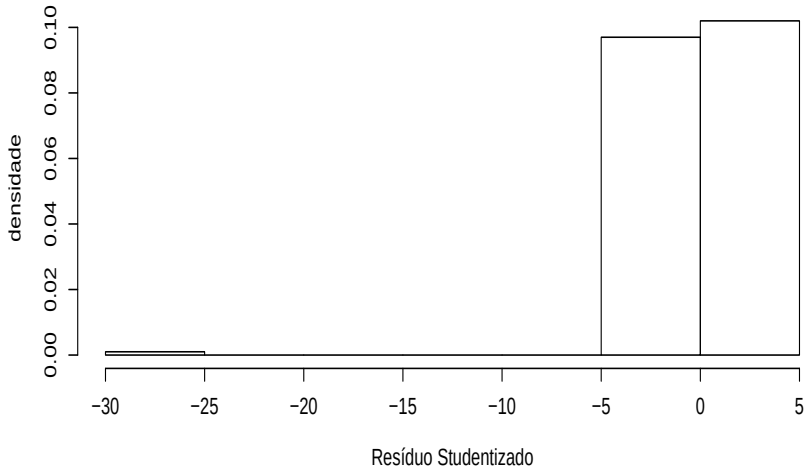
$$Y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \xi_i, i = 1, ..., 200$$

- $\overline{x} = \frac{1}{200}\sum_{i=1}^{200} x_i$.
- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- x_i: altura do i-ésimo indivíduo.
- β_{0i} : peso esperado para indivíduos com altura igual à $\overline{x} = 170,02$.
- β_{1i} : incremento (positivo ou negativo) no peso esperado para o aumento em uma unidade da altura (1 cm).

Gráficos de resíduos







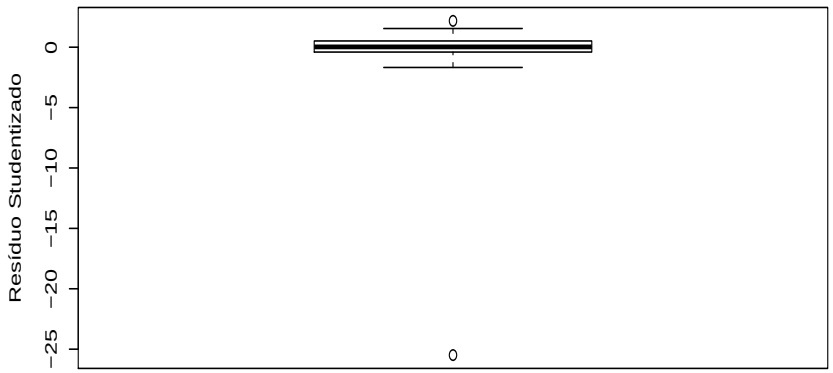
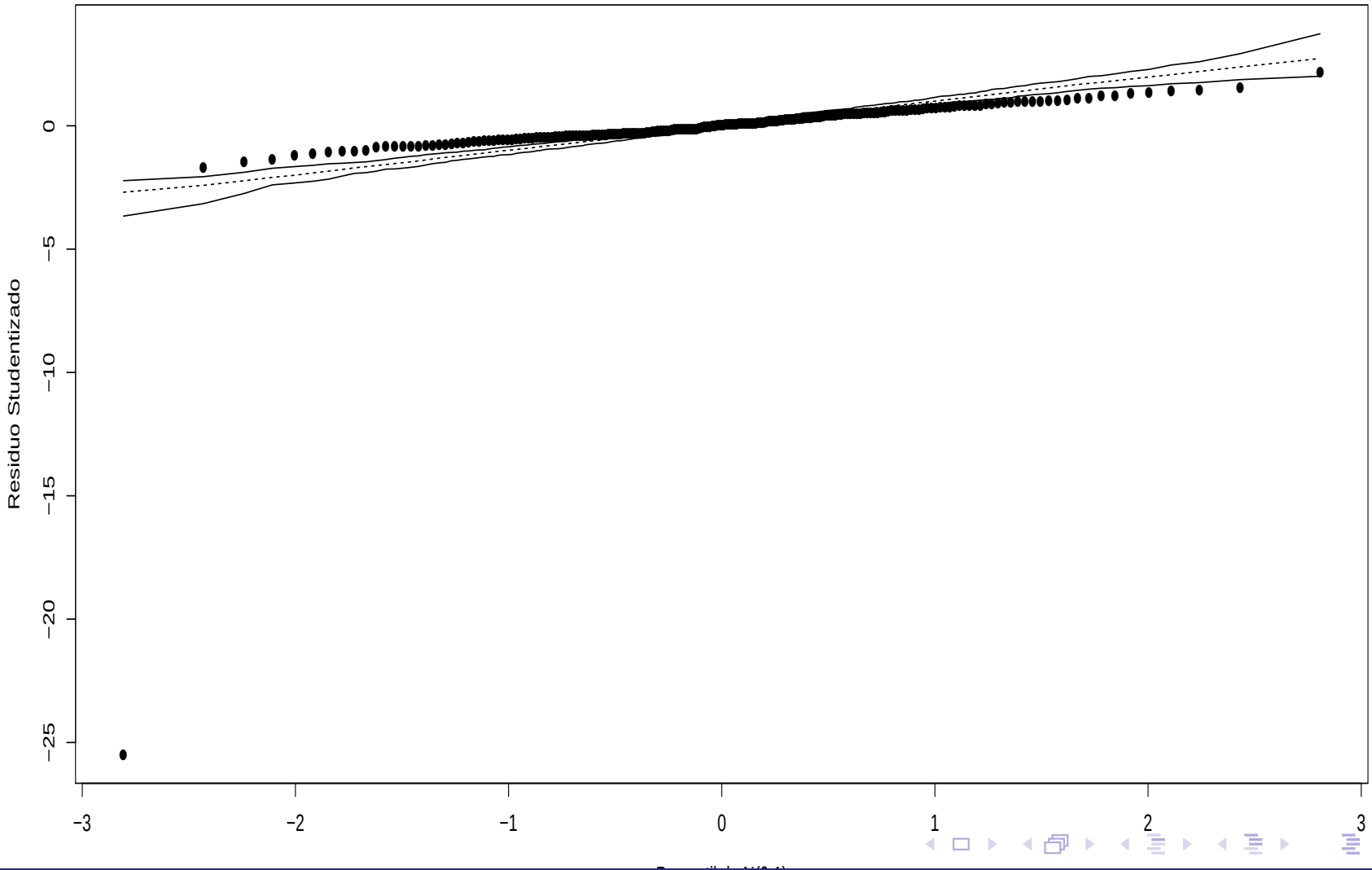


Gráfico de envelopes para os resíduos



Estimativas dos parâmetros

Parâmetro	Estimativa	EP	Estat. t	IC(95%)	p-valor
β_0	160,093	3,747	[152,704 ; 167,482]	42,728	$<$ 0,0001
β_1	0,151	0,056	[0,041 ; 0,260]	2,718	0,0072

$R^2 = 0,036$ e $\overline{R}^2 = 0,031$.

Comentários

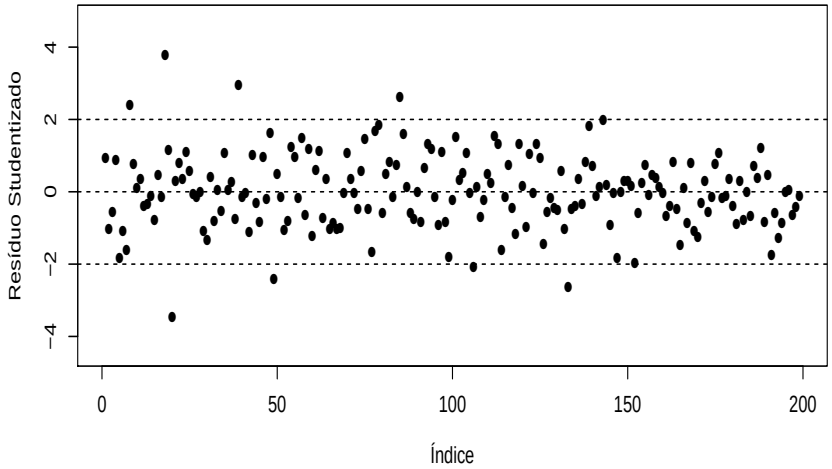
- Claramente, a observação destacada, influencia muito o ajuste do modelo.
- É necessário estudá-la em mais detalhes. Provavelmentem trata-se de um indivíduo com características muito peculirares (sub-população).
- Em geral, recomenda-se não retirar observações mas, neste caso, como não podemos contactar o(s) pesquisador(es) responsável(is) e devido aas restrições relativas à disciplina, vamos retirar tal observação.

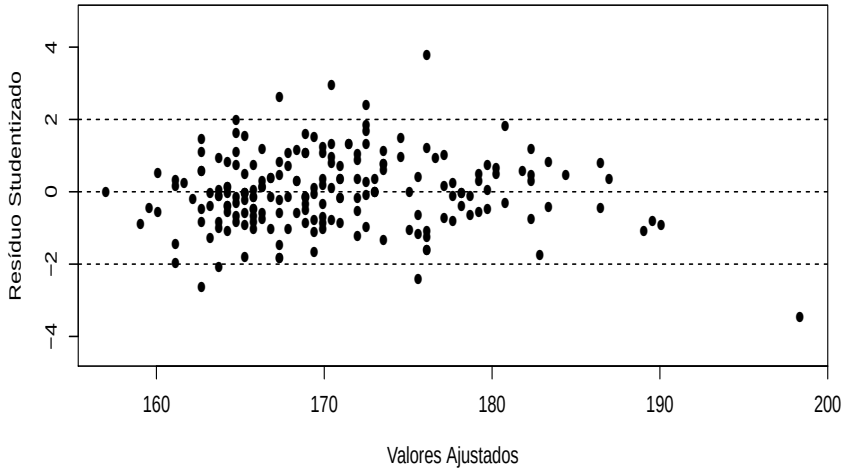
Sem a observação de # 12

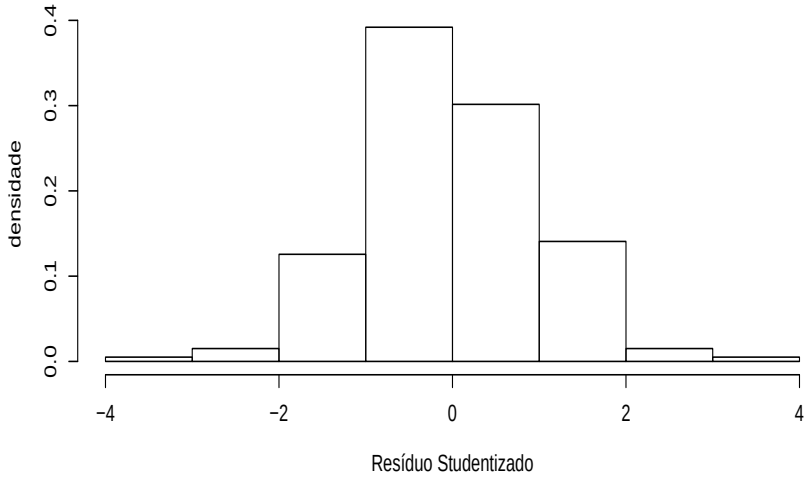
$$Y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \xi_i, i = 1, ..., 199$$

- $\overline{x} = \frac{1}{200}\sum_{i=1}^{200} x_i$.
- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- x_i: altura do i-ésimo indivíduo.
- β_{0i} : peso esperado para indivíduos com altura igual à $\overline{x} = 170,59$.
- β_{1i} : incremento (positivo ou negativo) no peso esperado para o aumento em uma unidade da altura (1 cm).

Gráficos de resíduos







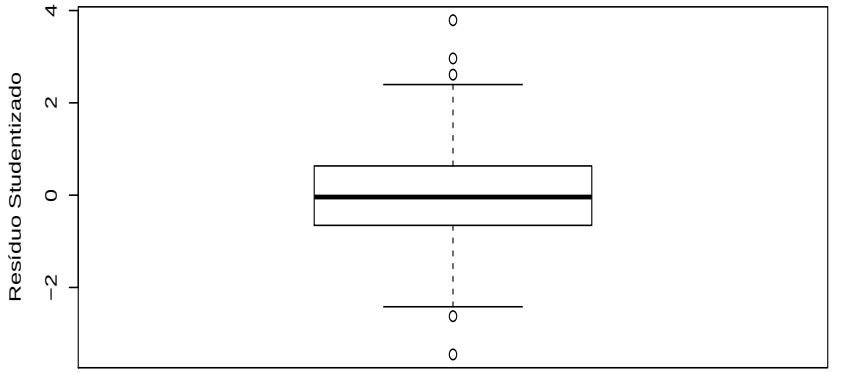
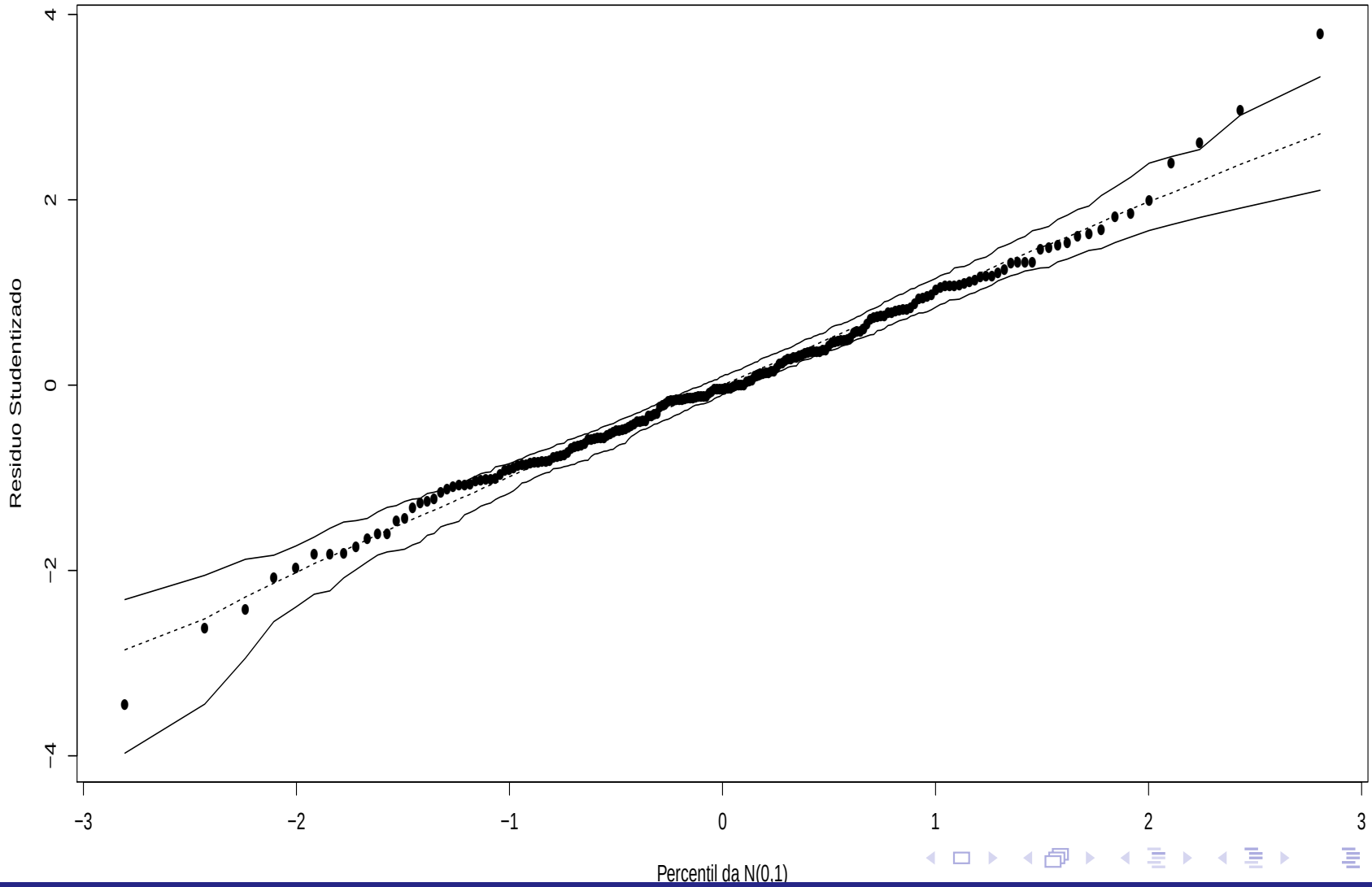


Gráfico de envelopes para os resíduos



Estimativas dos parâmetros

Parâmetro	Estimativa	EP	Estat. t	IC(95%)	p-valor
β_0	136,837	2,029	[132,836 ; 140,838]	67,446	$<$ 0,0001
β_1	0,517	0,030	[0,457 ; 0,577]	16,978	$<$ 0,0001

$R^2 = 0,594$ e $\overline{R}^2 = 0,592$.

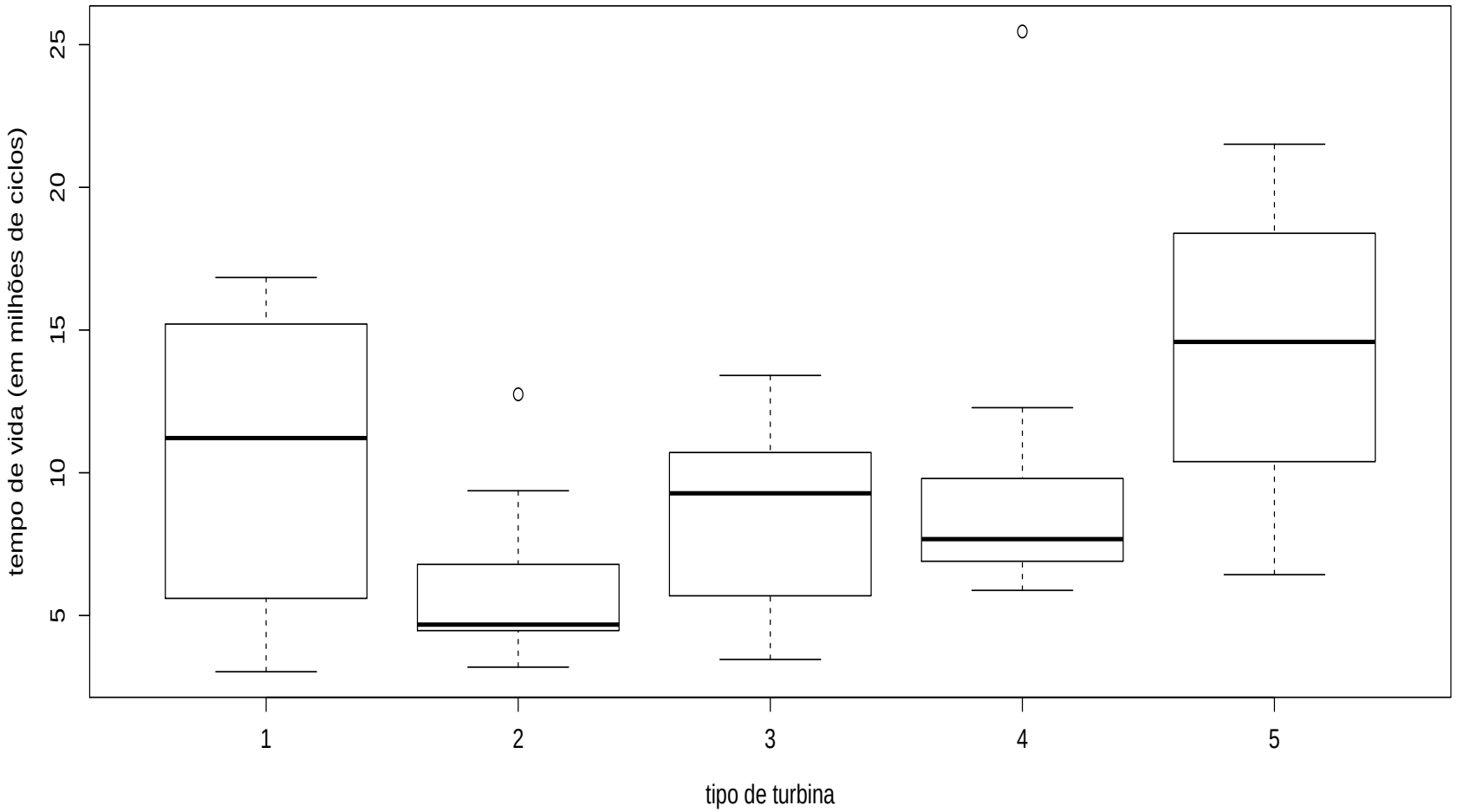
Comentários

- Presença de (leve) heterocedasticidade nos resíduos.
- Presencça de observações com valores elevados (em módulo) dos resíduos.
- Alternativa: modelo heterocedástico, para dados positivos (com caudas pesadas).

Exemplo 6: potência de turbinas de aviões

- Vamos considerar os 5 tipos de turbinas analisados no experimento, doravante tipos 1, 2, 3, 4 e 5.
- $n_i = 10, \forall i$ (tamanho amostral por grupo).
- Y_{ij} : tempo de vida (em milhões de ciclos) da j-ésima turbina do i-ésimo tipo.
- Quanto maior o número médio de ciclos, melhor o desempenho da turbina.

Análise descritiva



Análise descritiva

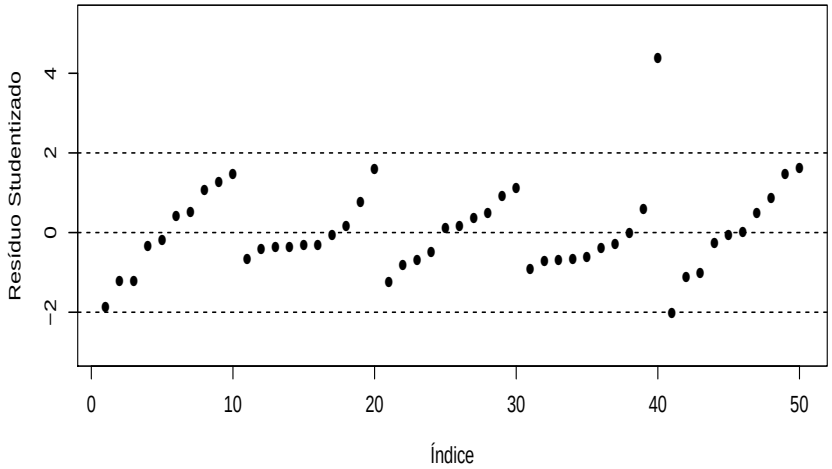
Tipo de turbina	Média	DP	Var.	CV(%)	CA	Mín.	Máx.
1	10,69	4,82	23,23	45,07	-0,20	3,03	16,84
2	6,05	2,92	8,50	48,18	1,20	3,19	12,75
3	8,64	3,29	10,83	38,10	-0,08	3,46	13,41
4	9,80	5,81	33,71	59,26	1,89	5,88	25,46
5	14,71	4,86	23,65	33,07	-0,13	6,43	21,51

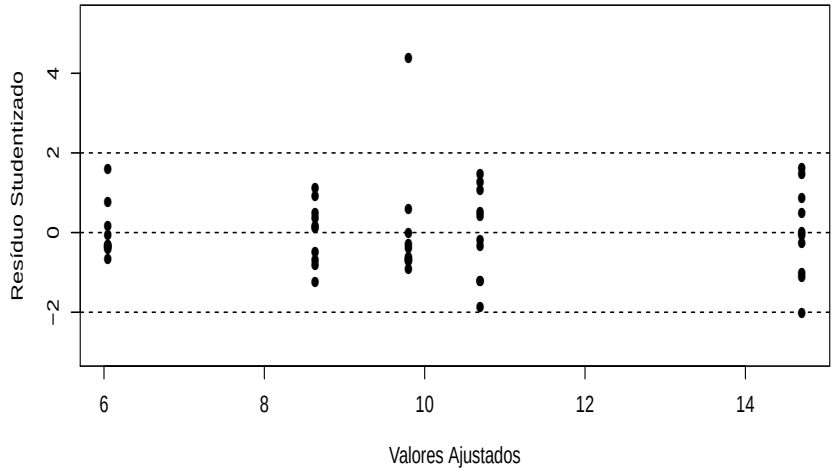
Modelo

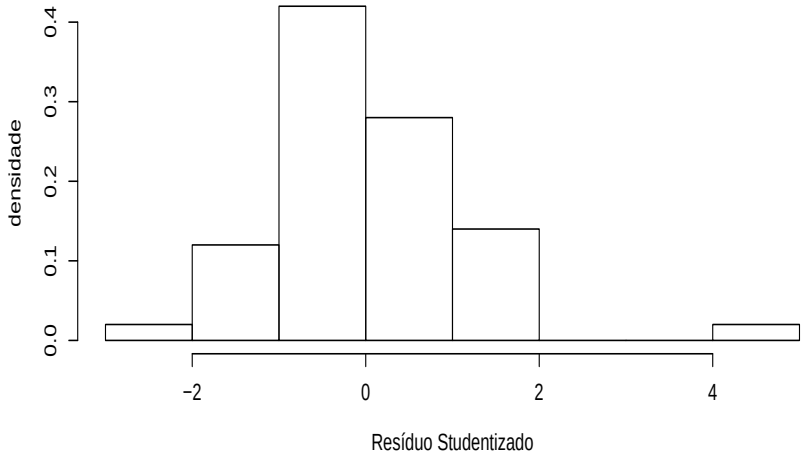
$$
\begin{aligned}
Y_{ij} &= \mu_i + \xi_i, \xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \\
i &= 1, ..., 5 \text{ (tipo de turbina)} \\
j &= 1, 2, ..., 10 \text{ (turbina - unidade experimental)} \\
\mu_i &= \alpha + \beta_i, \beta_1 = 0
\end{aligned}
$$

- $\beta_i = \mu_i - \mu_1, i = 2, .., 5$: incremento (aditivo) da média do tipo de turbina i com relação ao tipo de turbina 1 (referência).

Gráficos de resíduos







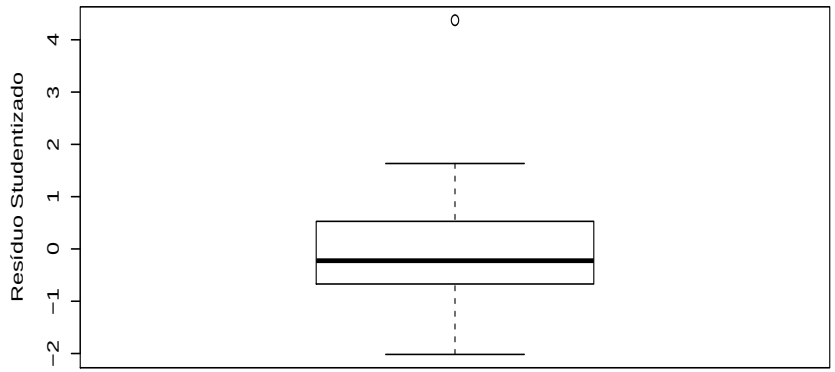
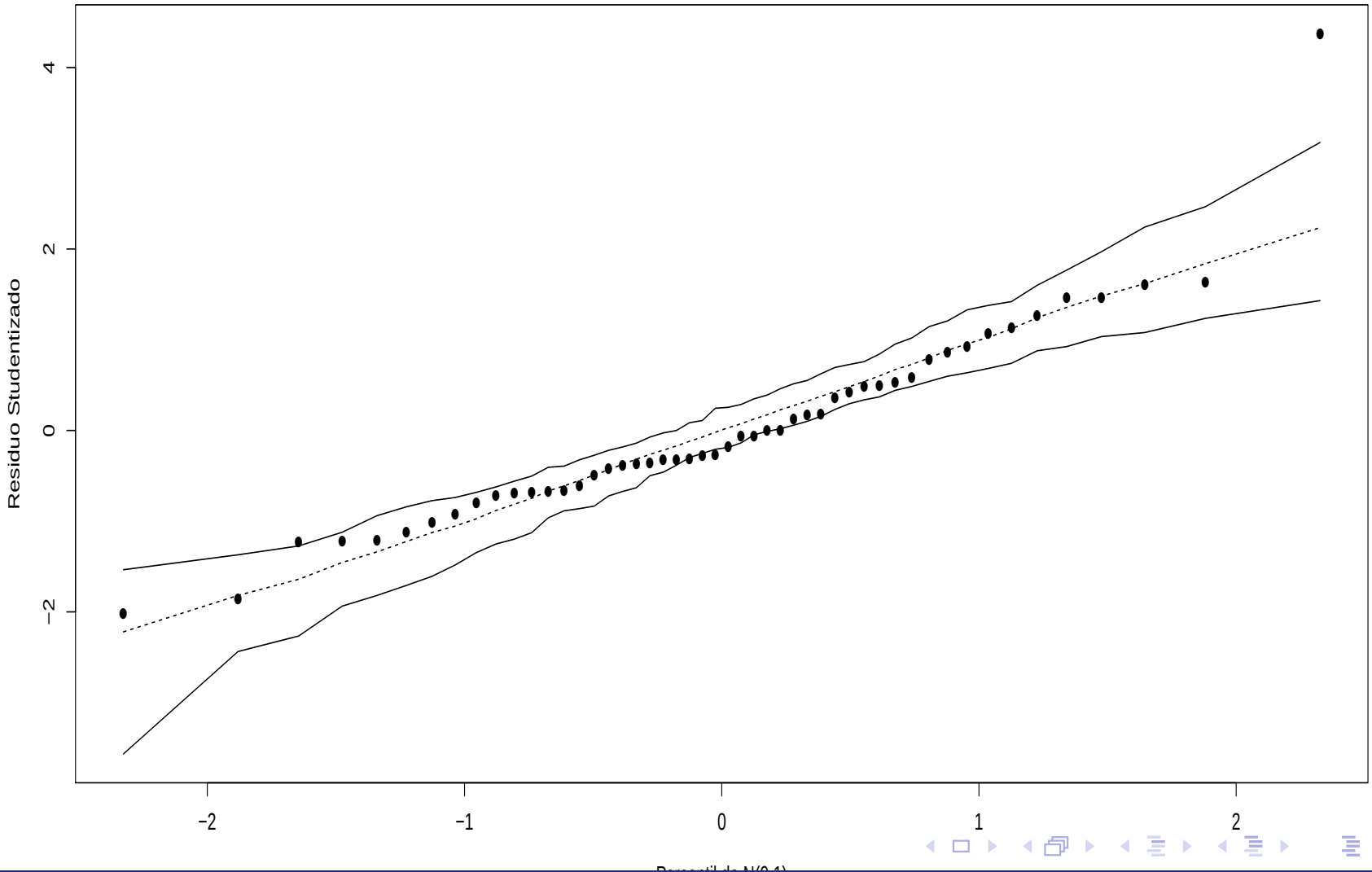


Gráfico de envelopes para os resíduos



Estimativas dos parâmetros

Parâmetro	Estimativa	EP	Estat. t	IC(95%)	p-valor
α	10,693	1,414	[7,846 ; 13,540]	7,564	<0,0001
β_2	-4,643	1,999	[-8,669 ; -0,617]	-2,322	0,0248
β_3	-2,057	1,999	[-6,083 ; 1,969]	-1,029	0,309
β_4	-0,895	1,999	[-4,921 ; 3,131]	-0,448	0,657
β_5	4,013	1,999	[-0,013 ; 8,039]	2,007	0,0507

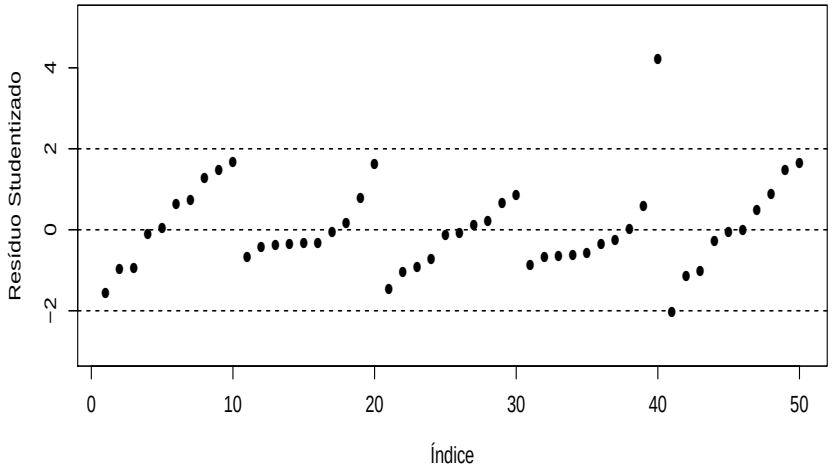
$R^2 = 0,309$ e $\overline{R}^2 = 0,247$.

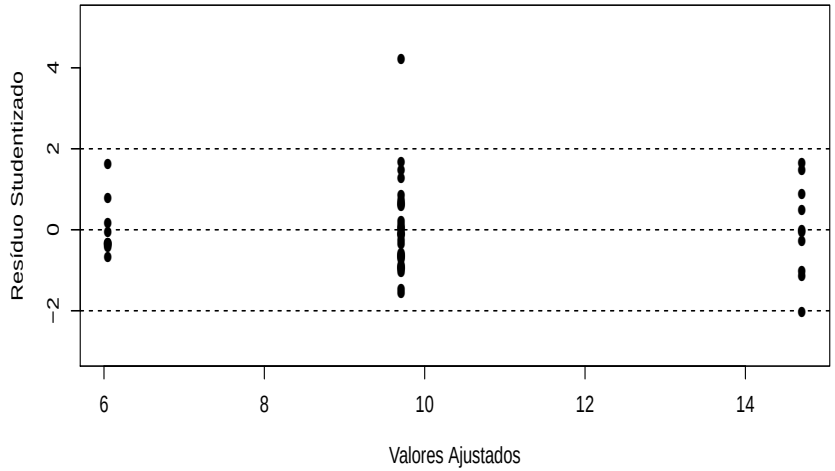
Modelo reduzido

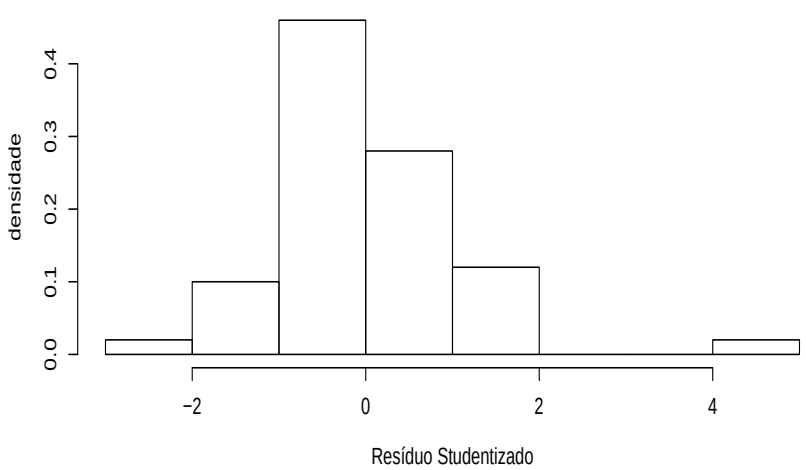
$$
\begin{aligned}
Y_{ij} &= \mu_i + \xi_i, \xi_i \overset{i.i.d.}{\sim} N(, \sigma^2), \\
i &= 1, ..., 5 \text{ (tipo de turbina)} \\
j &= 1, 2, ..., 10 \text{ (turbina - unidade experimental)} \\
\mu_i &= \alpha + \beta_i, \beta_1 = \beta_3 = \beta_4 = 0
\end{aligned}
$$

- $\beta_i = \mu_i - \mu_1, i = 2, .., 5$: incremento (aditivo) da média do tipo de turbina i com relação ao tipo de turbina 1 (referência).

Gráficos de resíduos







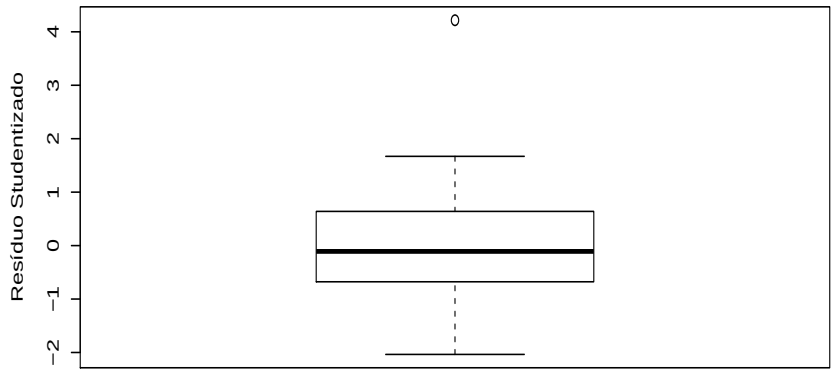
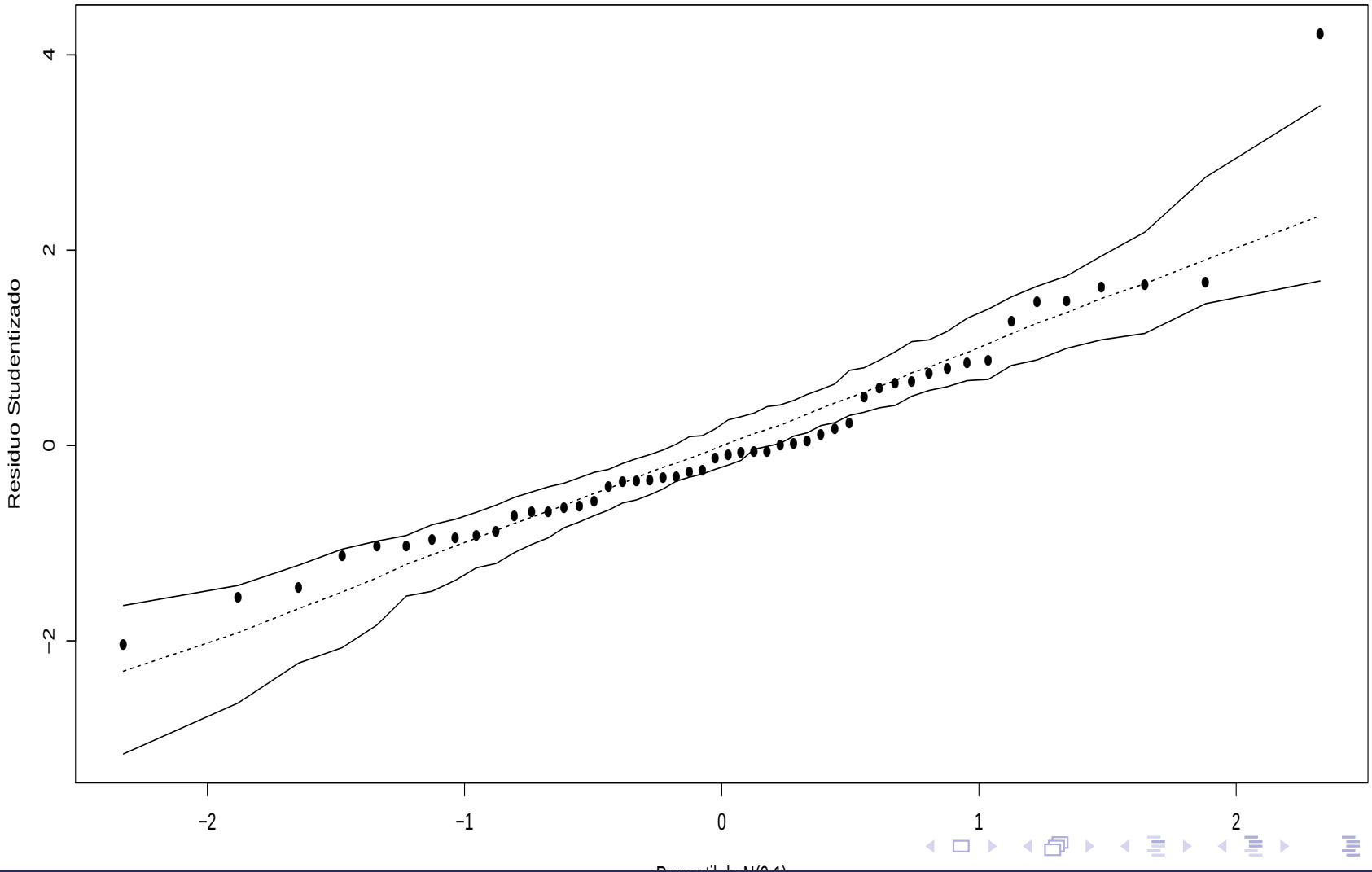


Gráfico de envelopes para os resíduos

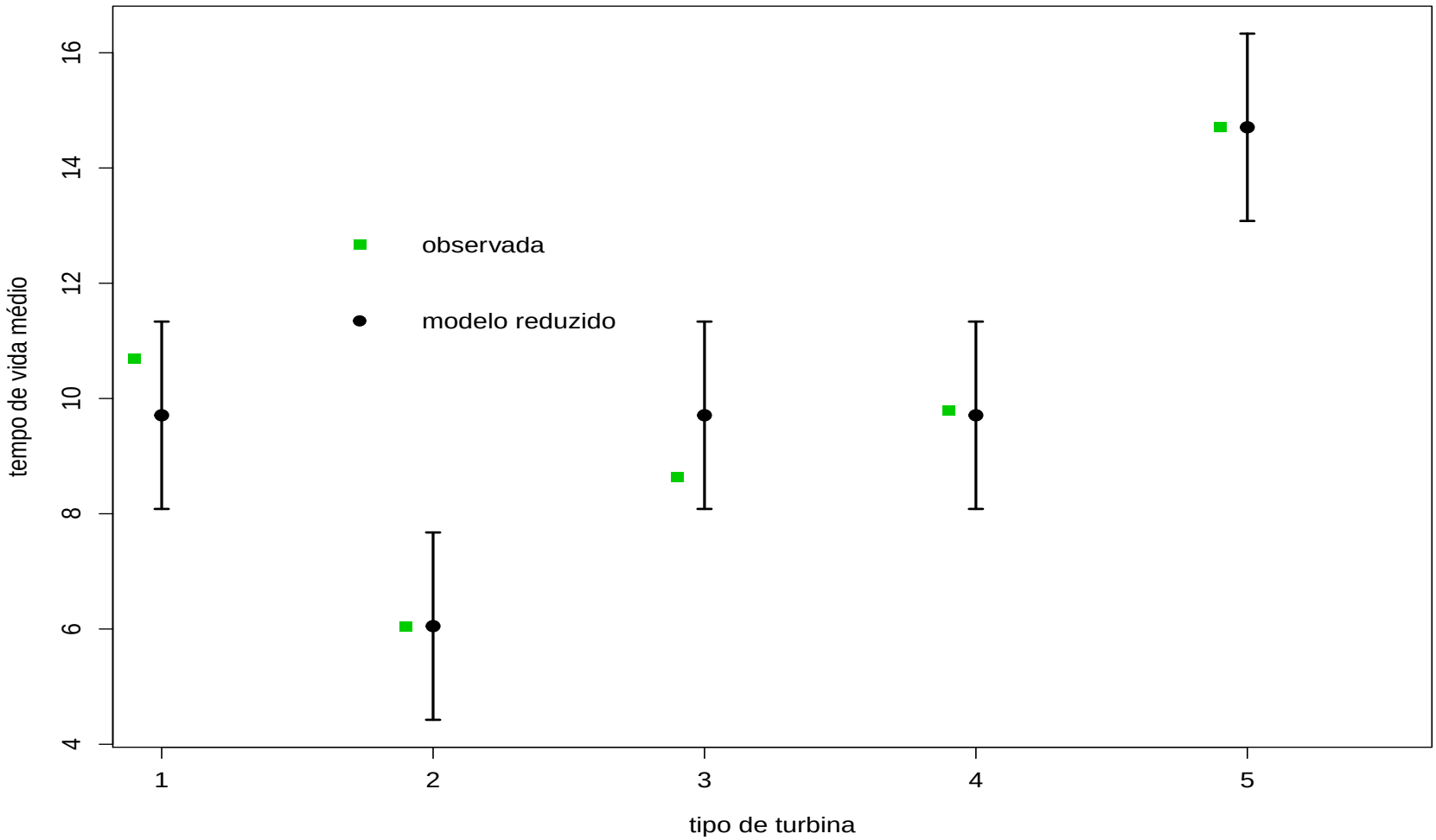


Estimativas dos parâmetros

Parâmetro	Estimativa	EP	Estat. t	IC(95%)	p-valor
β_0	9,709	0,808	[8,084 ; 11,334]	12,016	, 0,0001
β_2	-3,659	1,616	[-6,910 ; -0,408]	-2,264	0,0282
β_5	4,997	1,616	[1,746 ; 8,248]	3,092	$<$ 0,0033

$R^2 = 0,292$ e $\overline{R}^2 = 0,262$.

Médias preditas pelo modelo reduzido

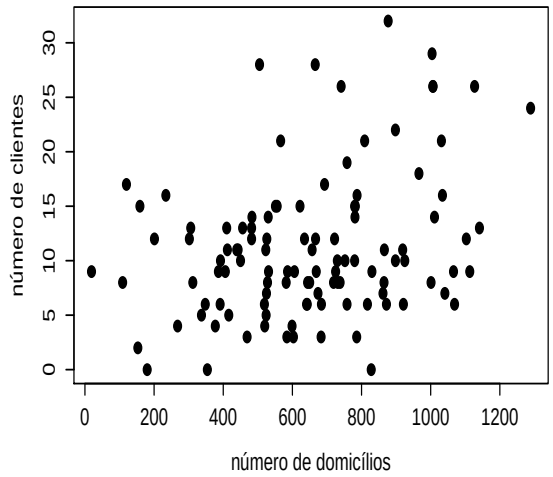


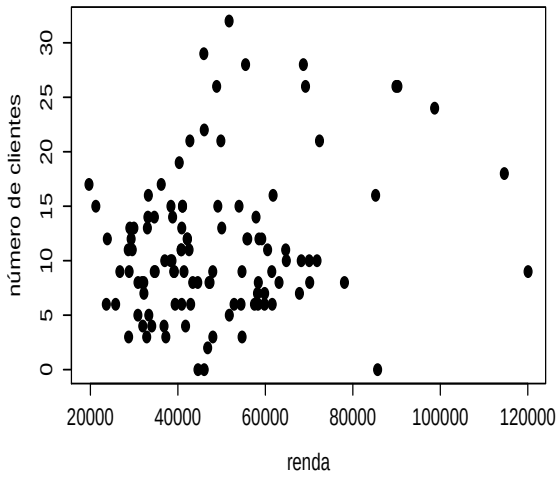
Exemplo 7: perfil dos clientes de uma loja

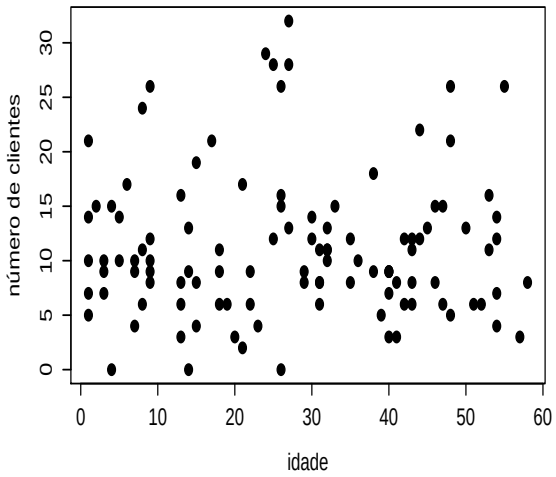
- Interesse: estudar o perfil dos clientes de uma determinada loja oriundos de 110 áreas de uma determinada cidade. Cada uma das 110 observações corresponde à uma área da cidade.
- Verificar como certas características (variáveis explicativas) afetam o número esperado de clientes em cada área (variável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

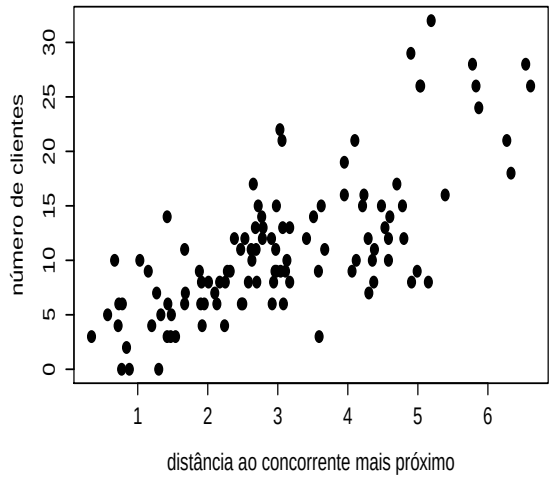
Mapa de Fortaleza (ilustração)

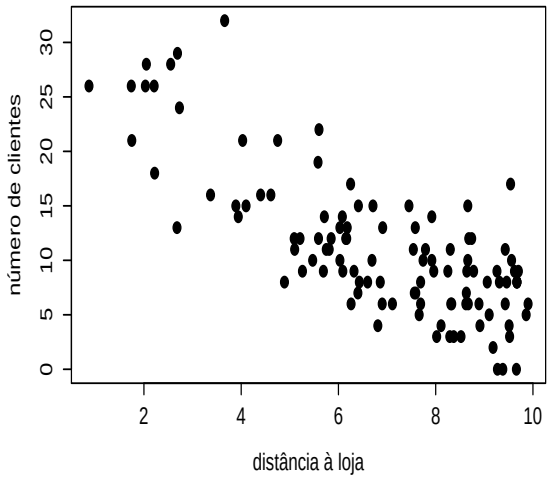
Gráficos de dispersão











Legenda

- ndom - número de domicílios.
- renda - renda média anual.
- idade - idade média dos domicílios.
- disc - distância ao concorrente mais próximo.
- disl - distância à loja

Medidas resumo

	Variável				
Medida-resumo	ndom	renda	idade	dist	disl
Média	647,76	48836,78	27,43	3,07	6,83
DP	263,03	18531,06	16,68	1,50	2,29
CV(%)	40,61	37,94	60,83	49,02	33,54
Mediana	647,00	44564,50	27,00	2,93	7,28
Mínimo	19,00	19673,00	1,00	0,34	0,87
Máximo	1289,00	120065,00	58,00	6,61	9,90

Modelo (completo)

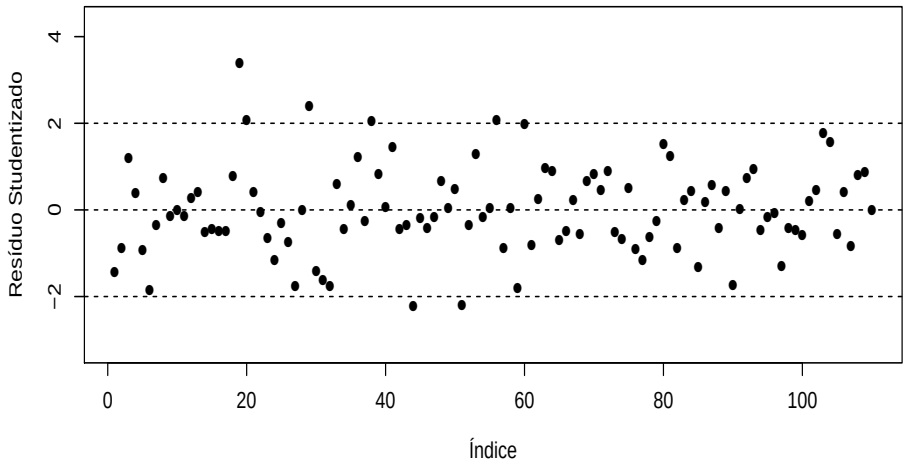
$$
\begin{aligned}
Y_i &= \mu_i + \xi_i, \xi_i \overset{i.i.d}{\sim} N(0, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \left(\frac{x_{1i} - \overline{x}_1}{s_1} \right) + \beta_2 \left(\frac{x_{2i} - \overline{x}_2}{s_2} \right) + \beta_3 \left(\frac{x_{3i} - \overline{x}_3}{s_3} \right) + \\
&+ \beta_4 \left(\frac{x_{4i}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

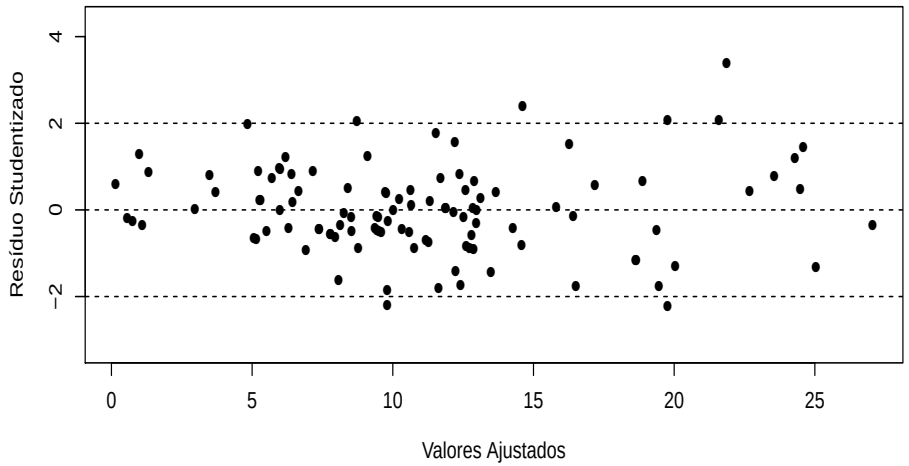
- x_{ji} : valor da variável explicativa j, $j = 1, 2, ..., 5$, associada à área i, $\overline{x}_j = \frac{1}{110} \sum_{i=1}^{110} x_{ji}$, e $s_j = \frac{\sum_{i=1}^{n} (x_{ji} - \overline{x}_j)^2}{109}$ $j = 1, 2, ..., 5$.

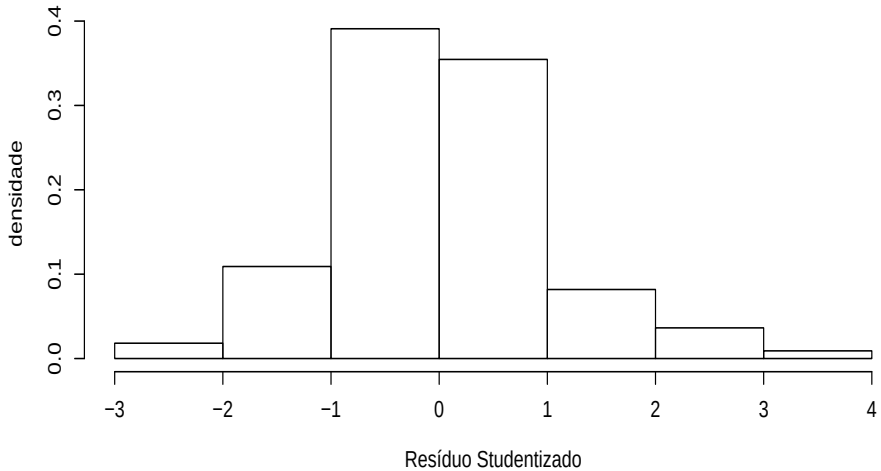
Modelo (completo)

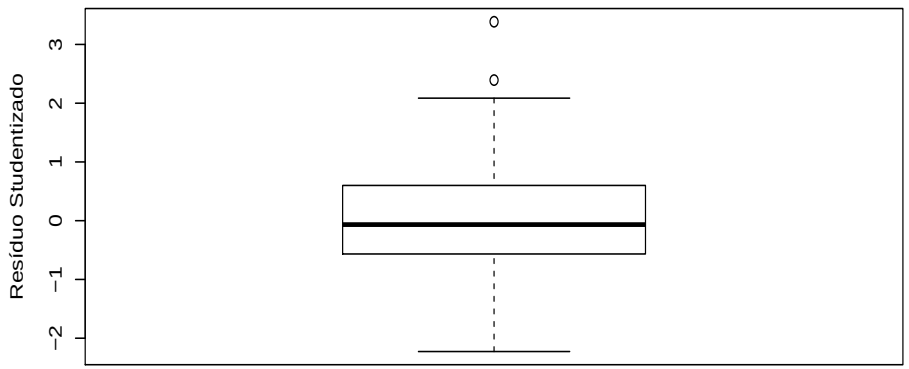
- β_0 número esperado de clientes para domicílios localizados em áreas com valor médio para cada uma das covariáveis.
- β_j/s_j : incremento (positivo ou negativo) no valor esperado do número de clientes, para o aumento em uma unidade no valor da covariável j, mantendo-se todas as outras fixas.
- Uma 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.

Gráficos de diagnóstico

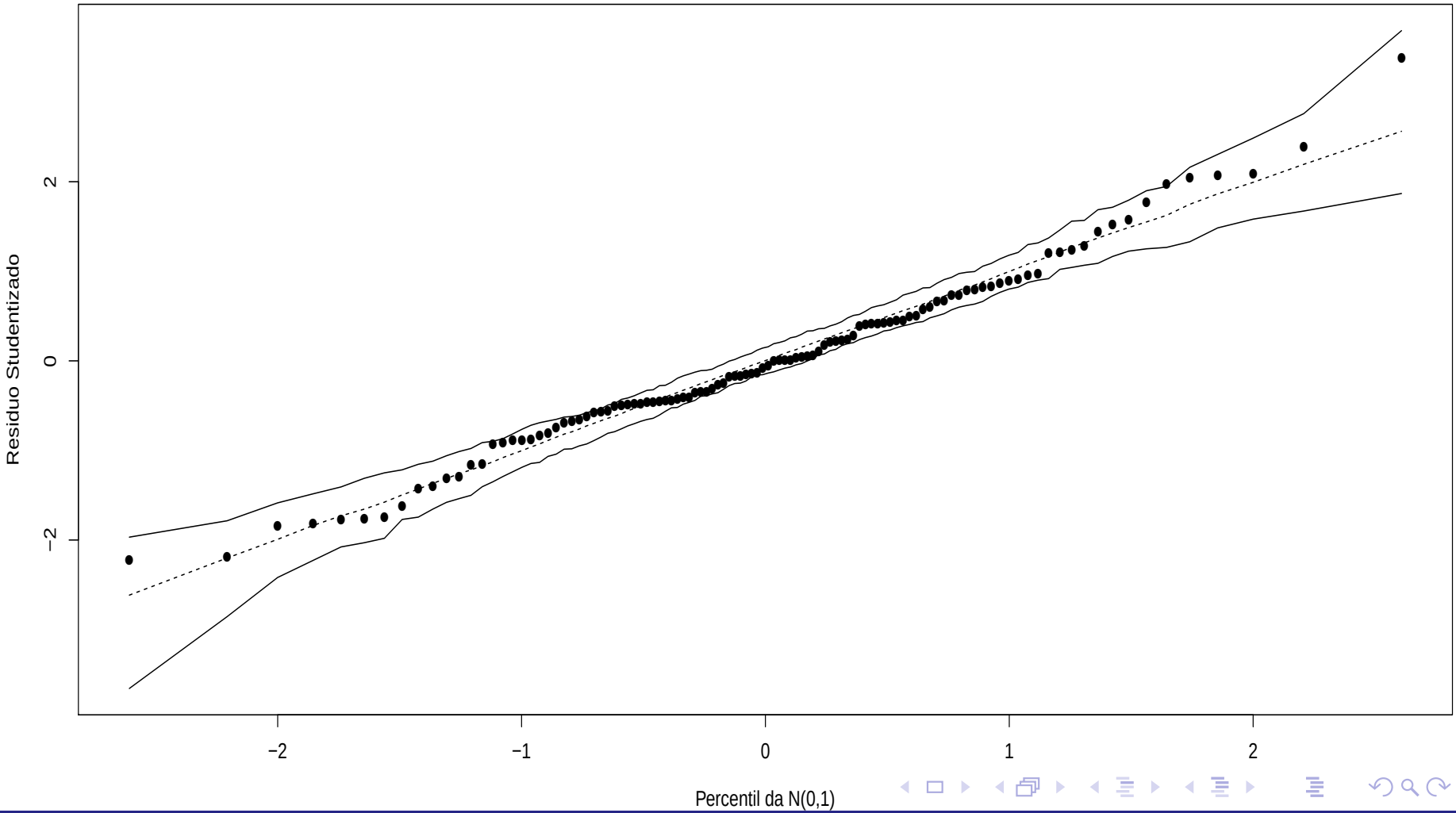








Gráficos de envelope



Comentários

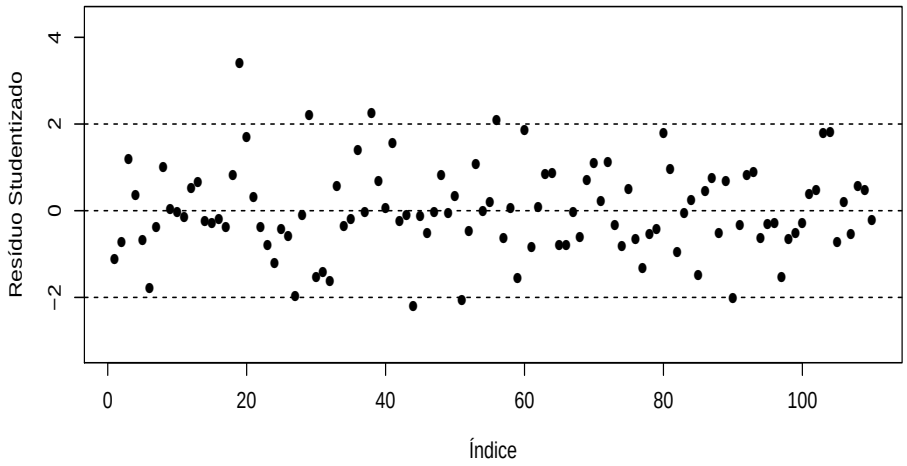
- Presença de leve assimetria positiva e heterocedasticidade.
- Alternativa: modelos para dados de contagem.

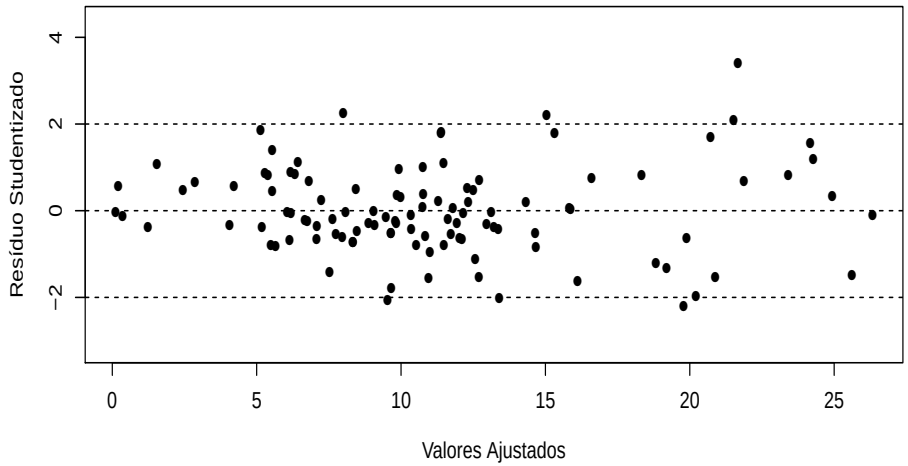
Estimativas dos parâmetros

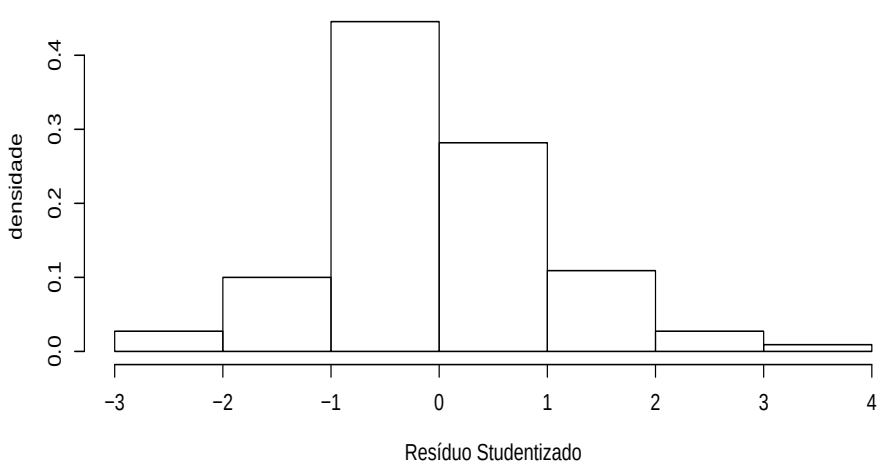
Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	11,200	0,306	[10,593 ; 11,807]	36,564	<0,0001
β_1	1,736	0,400	[0,943 ; 2,530]	4,339	<0,0001
β_2	-2,153	0,423	[-2,992 ; -1,314]	-5,088	<0,0001
β_3	-0,599	0,314	[-1,221 ; 0,022]	-1,912	0,0587
β_4	2,863	0,388	[2,094 ; 3,632]	7,382	<0,0001
β_5	-3,918	0,398	[-4,708 ; -3,129]	-9,837	<0,0001

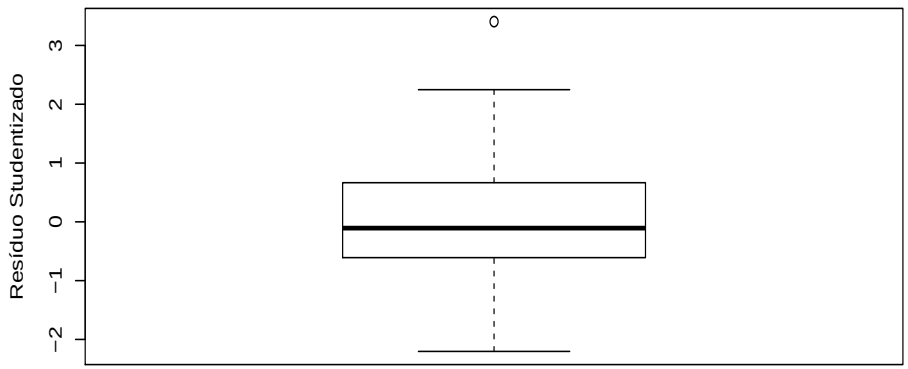
$R^2 = 0,777$ e $\overline{R}^2 = 0,766$. Ajustar um modelo reduzido, sem a variável idade.

Gráficos de diagnóstico

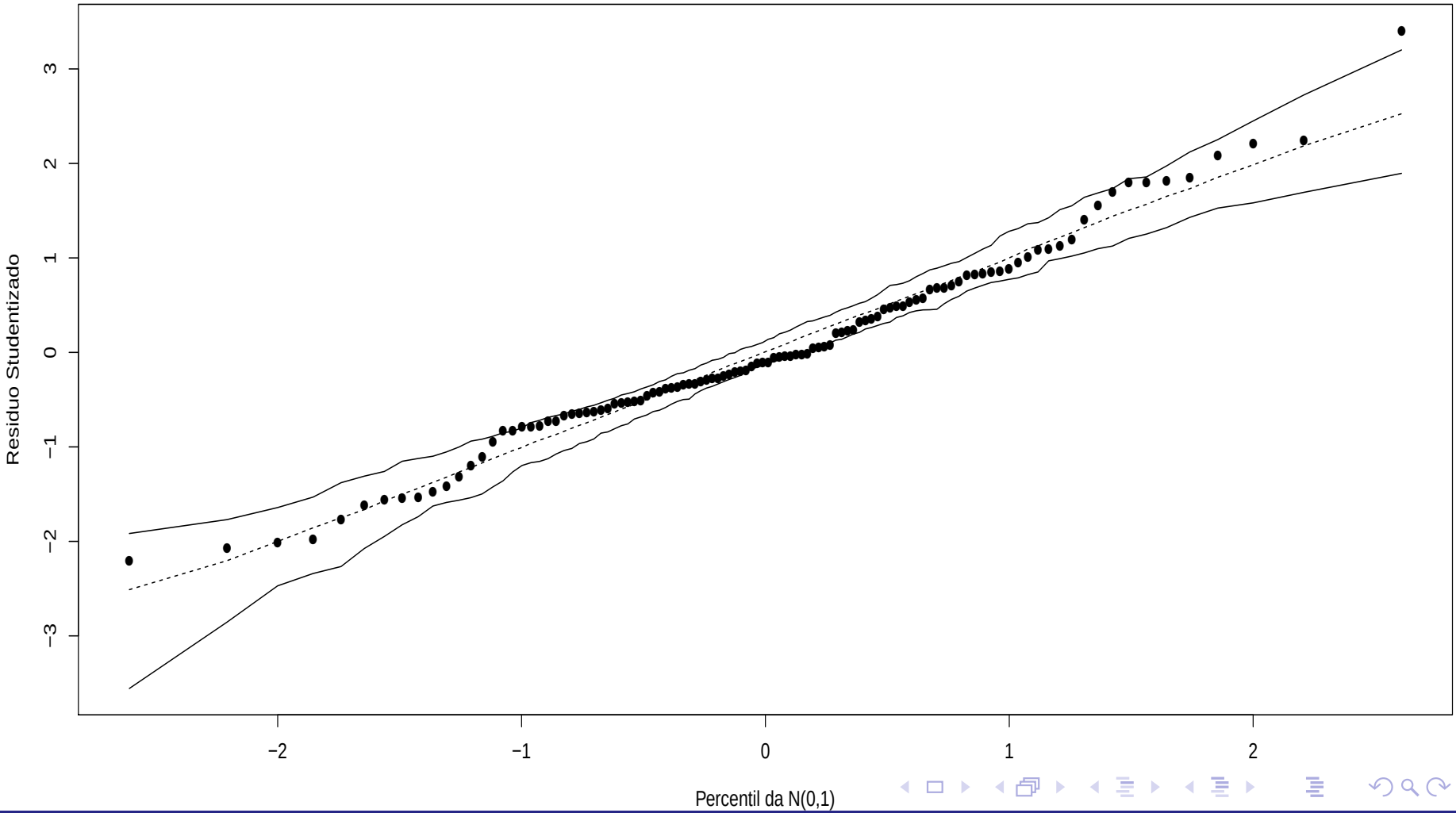








Gráficos de envelope



Comentários

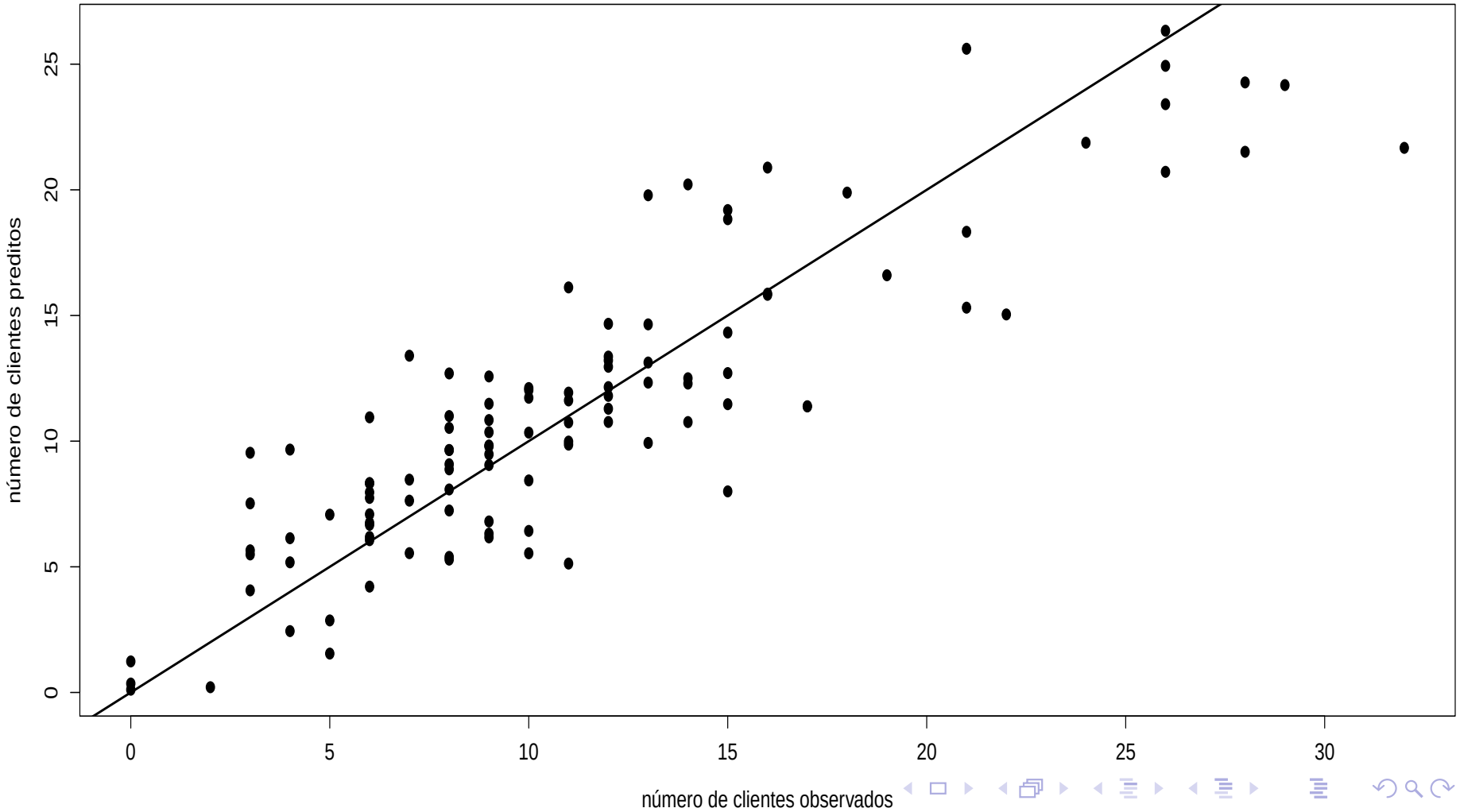
- Presença de leve assimetria positiva e heterocedasticidade.
- Alternativa: modelos para dados de contagem.

Estimativas dos parâmetros

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	11,200	0,310	[10,585 ; 11,815]	36,110	<0,0001
β_1	1,602	0,399	[0,811 ; 2,393]	4,017	<0,0001
β_2	-2,038	0,424	[-2,879 ; -1,197]	-4,806	<0,0001
β_4	2,857	0,393	[2,078 ; 3,635]	7,275	<0,0001
β_5	-3,871	0,403	[-4,670 ; -3,073]	-9,616	<0,0001

$R^2 = 0,769$ e $\overline{R}^2 = 0,760$.

Valores preditos e observados da resposta



Valores preditos e observados da resposta