Exemplo Seleção de Modelos

Gilberto A. Paula

Departamento de Estatística
IME-USP, Brasil
giapaula@ime.usp.br

2º Semestre 2022

Sumário

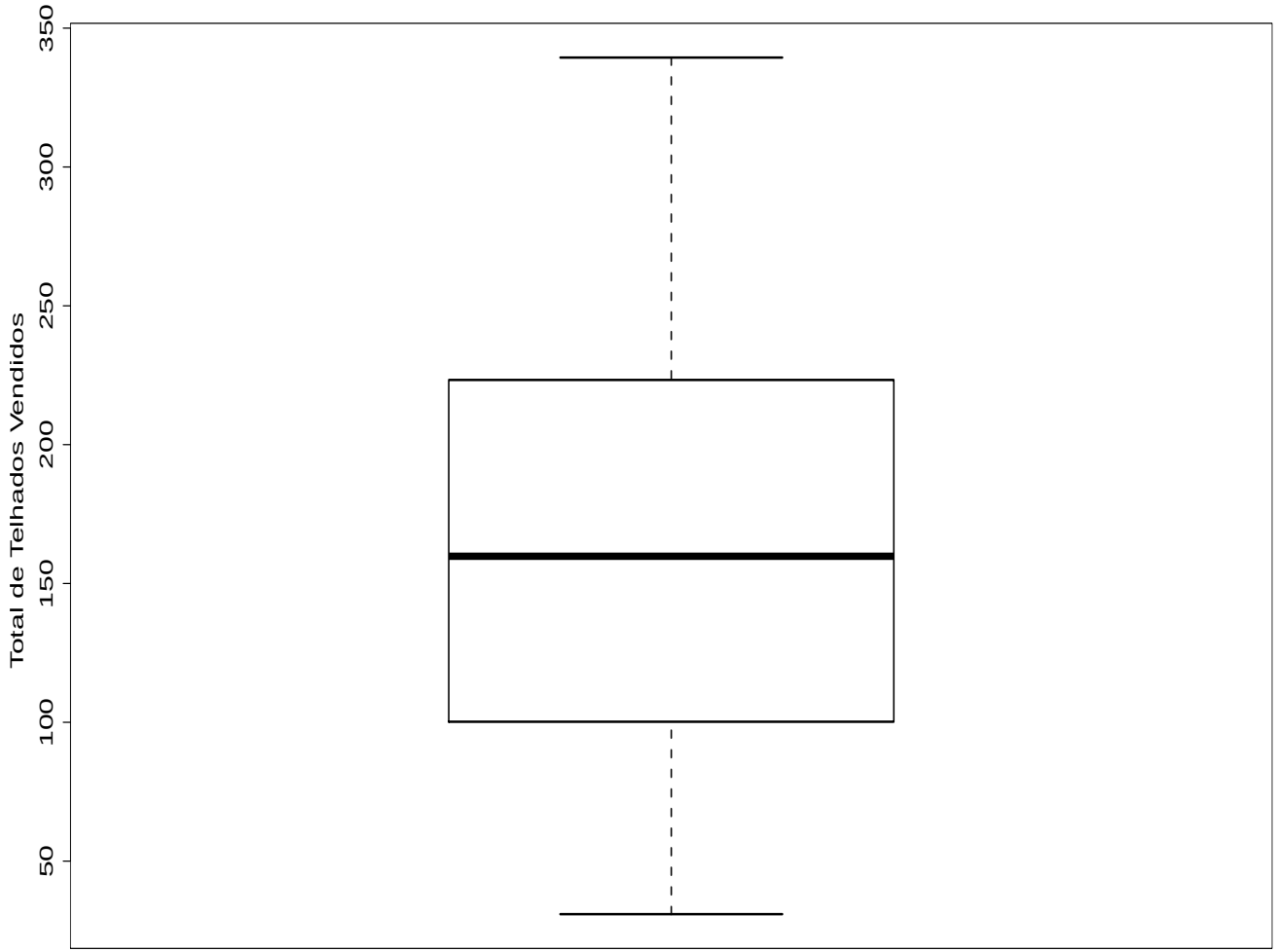
Descrição dos Dados

Para ilustrar um exemplo de seleção de modelos em regressão linear múltipla serão considerados os dados descritos em Neter et al. (1996, p.449) referentes à venda no ano anterior de um tipo de telhado de madeira em $n = 26$ filiais de uma rede de lojas de construção civil. As seguintes variáveis serão consideradas:

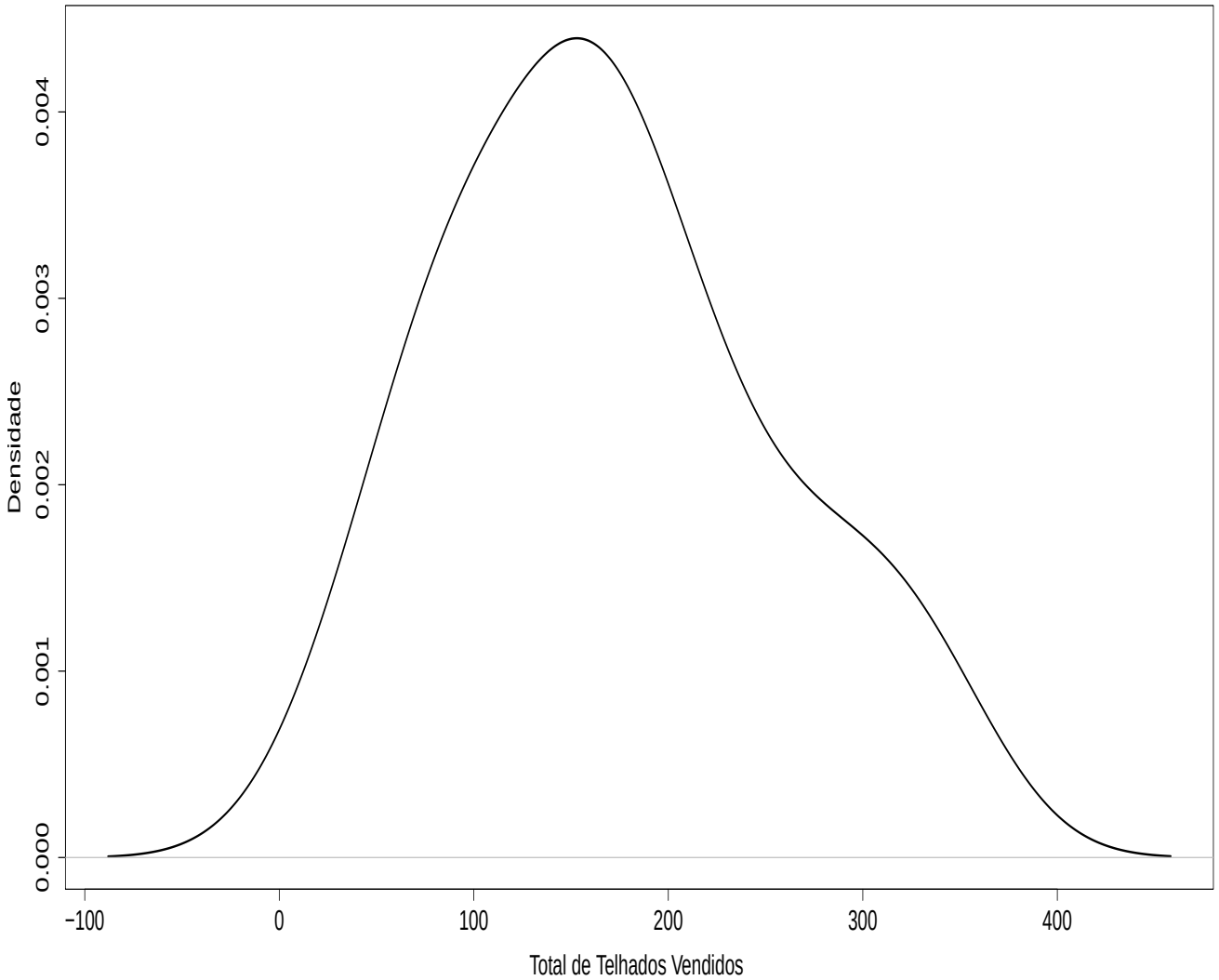
- Telhados, total de telhados vendidos (em mil metros quadrados)
- Gastos, gastos pela loja com promoções do produto (em mil USD)
- Clientes, número de clientes cadastrados na loja (em milhares)
- Marcas, número de marcas concorrentes do produto
- Potencial, potencial da loja (quanto maior o valor maior o potencial).

O interesse é explicar o número médio de telhados vendidos dadas as demais variáveis.

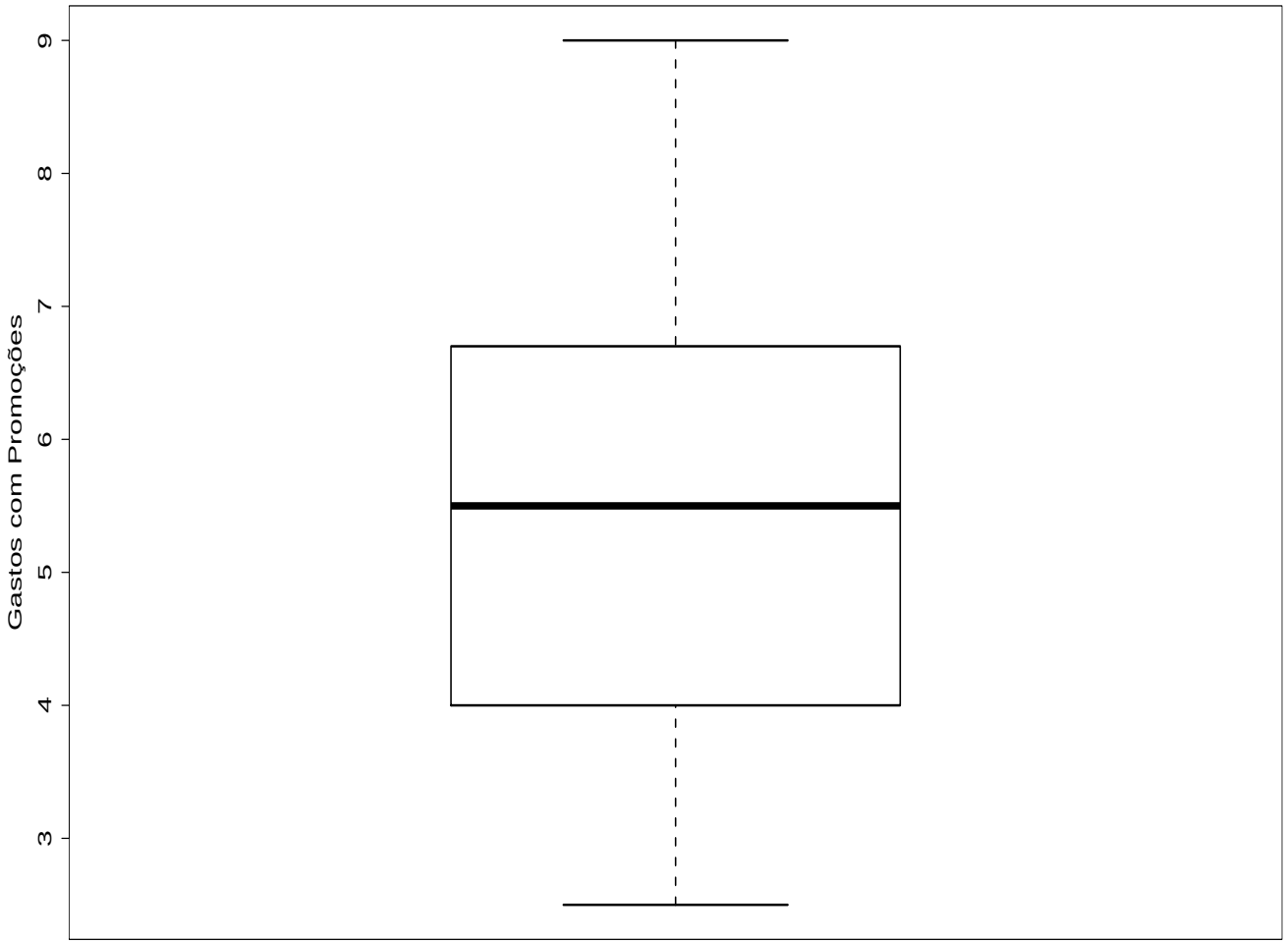
Boxplot Total de Telhados Vendidos



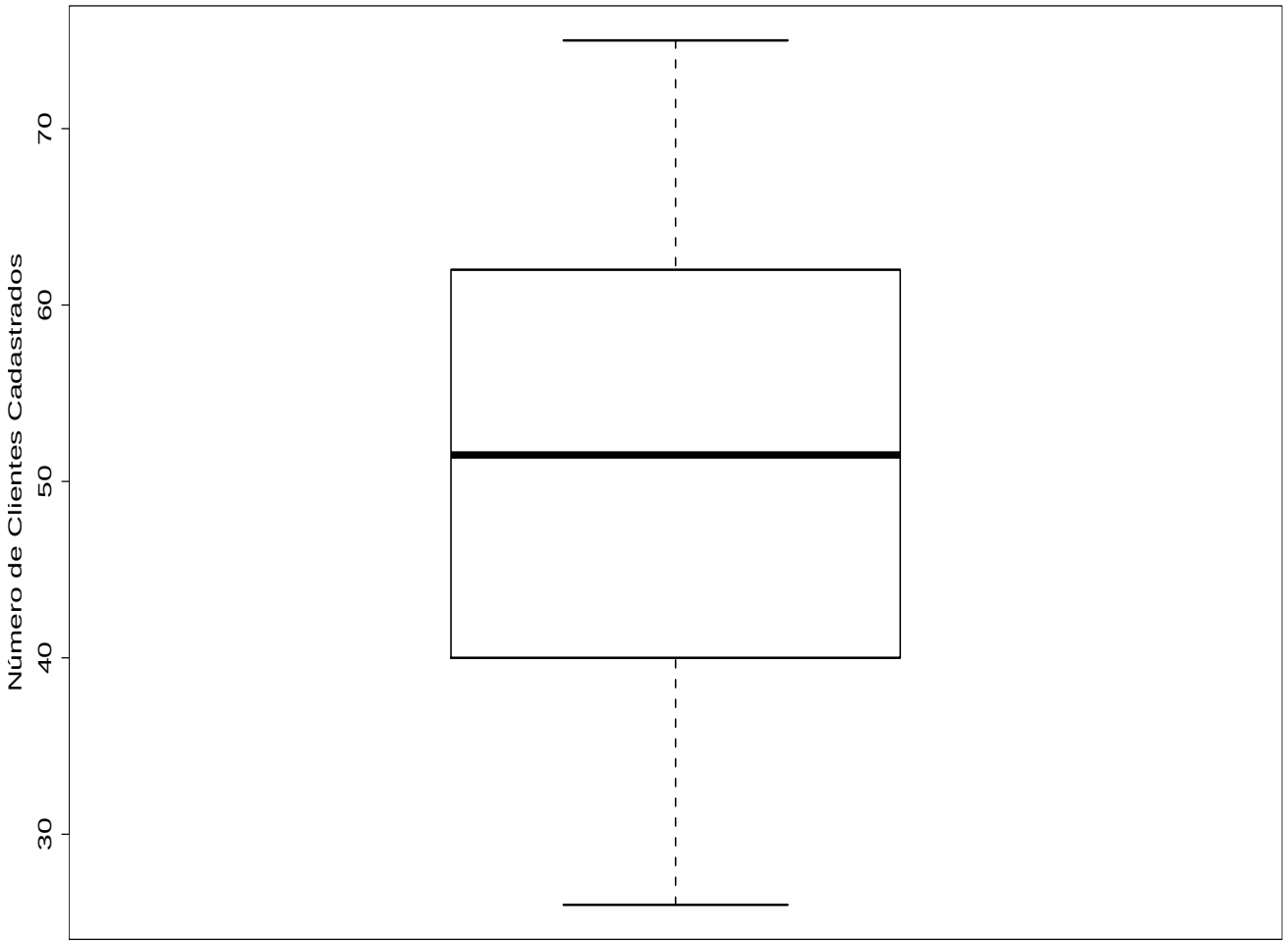
Densidade Total de Telhados Vendidos



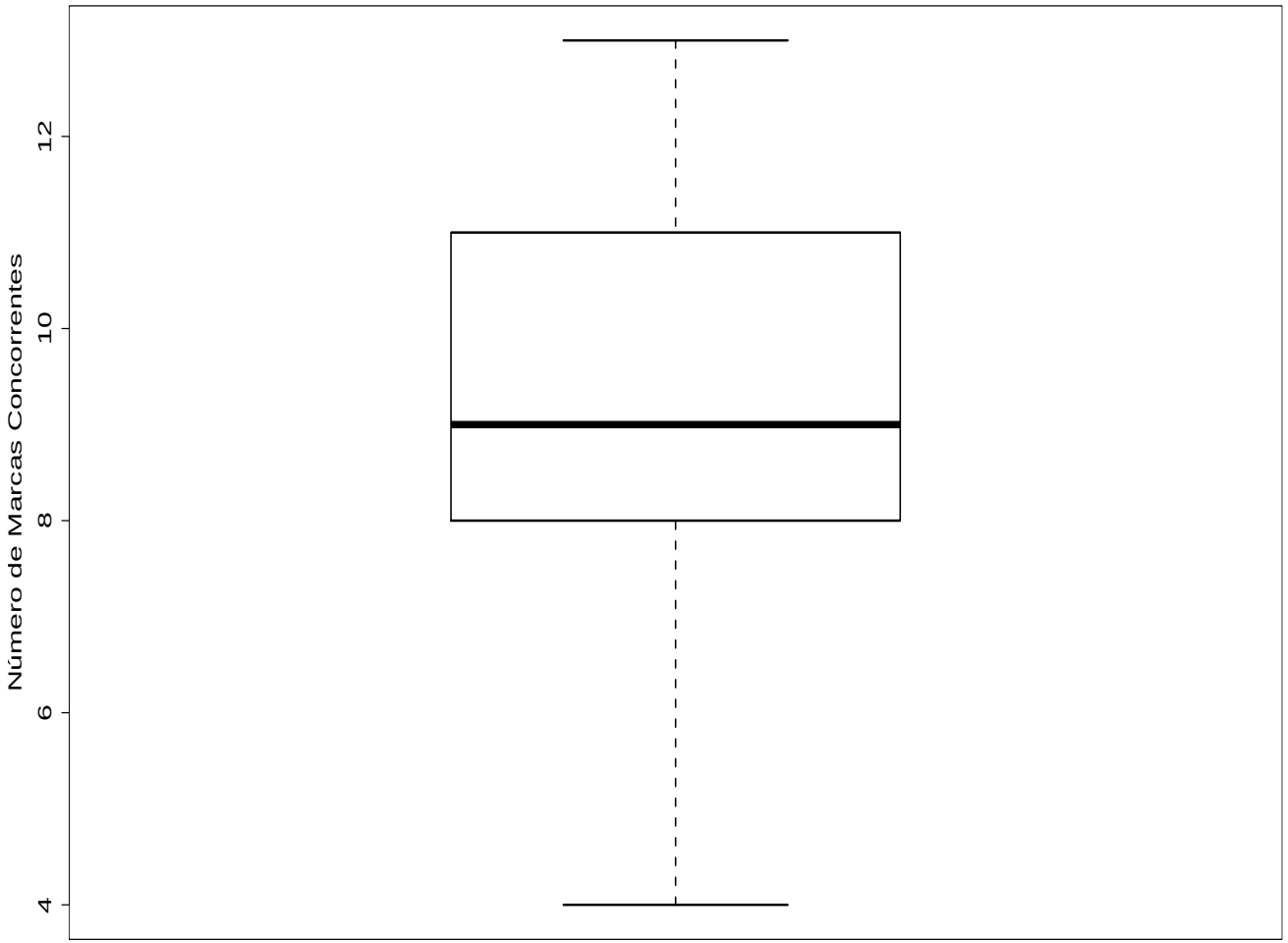
Boxplot Gastos com Promoções



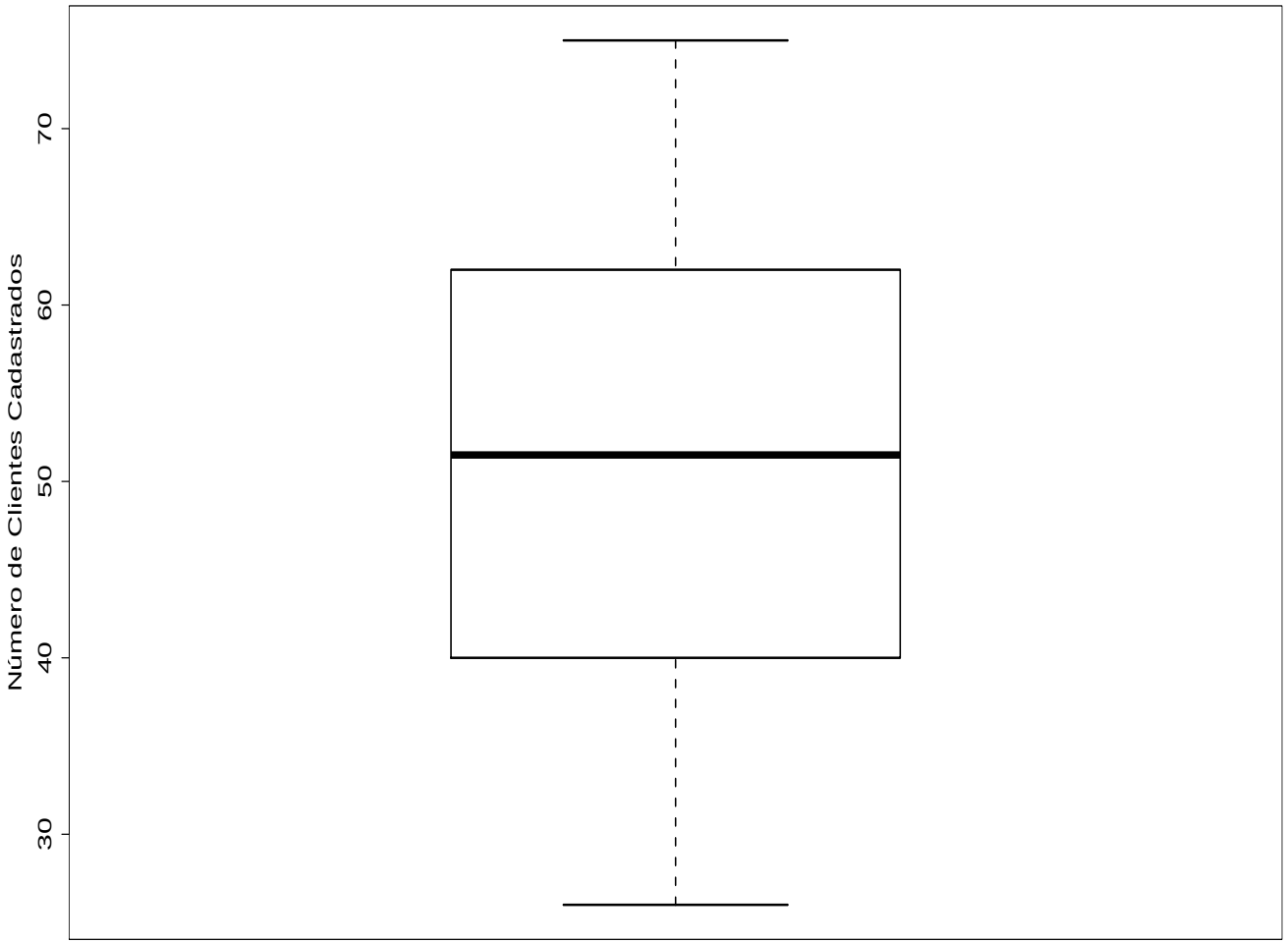
Boxplot Número de Clientes Cadastrados



Boxplot Número de Marcas Concorrentes



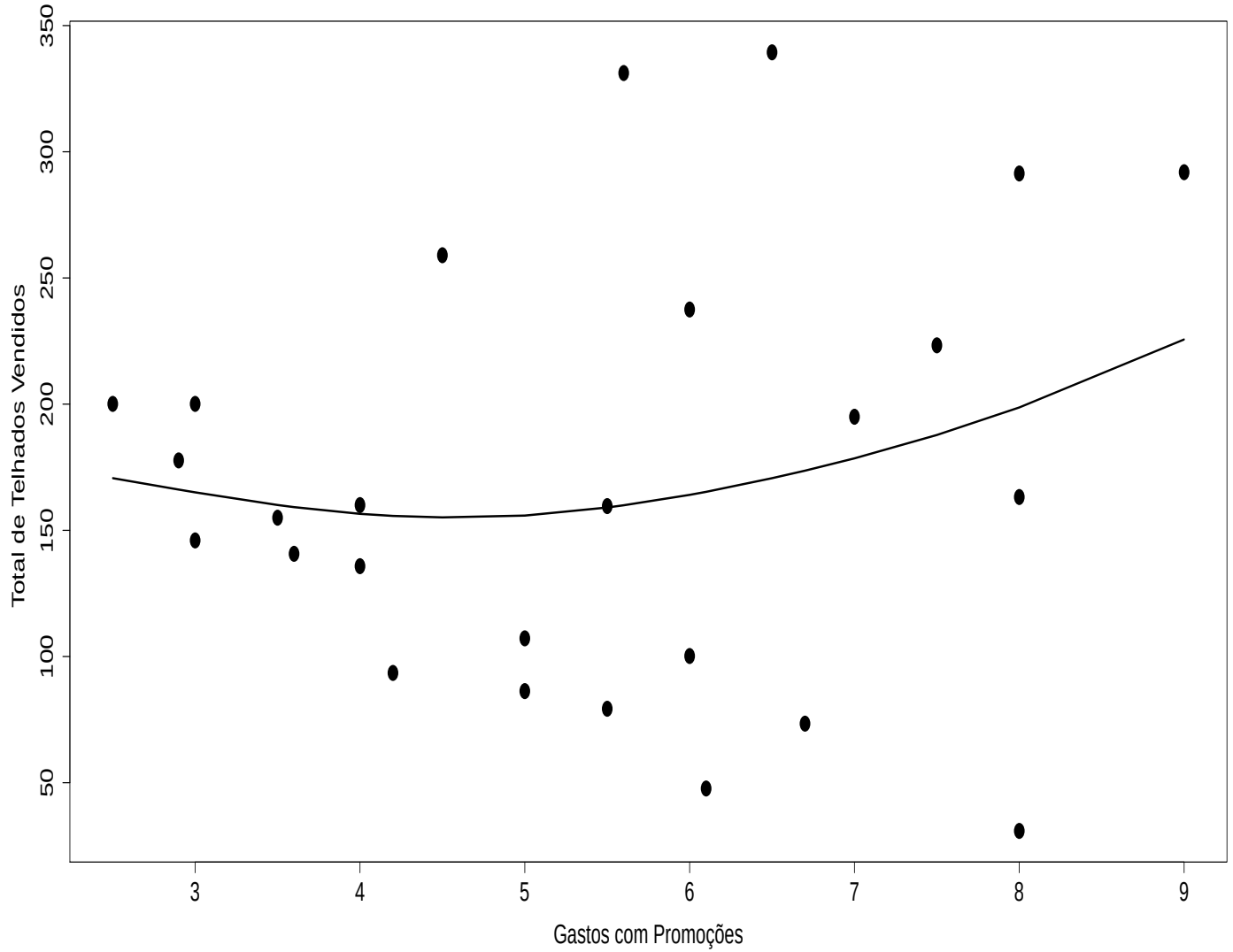
Boxplot Potencial da Loja



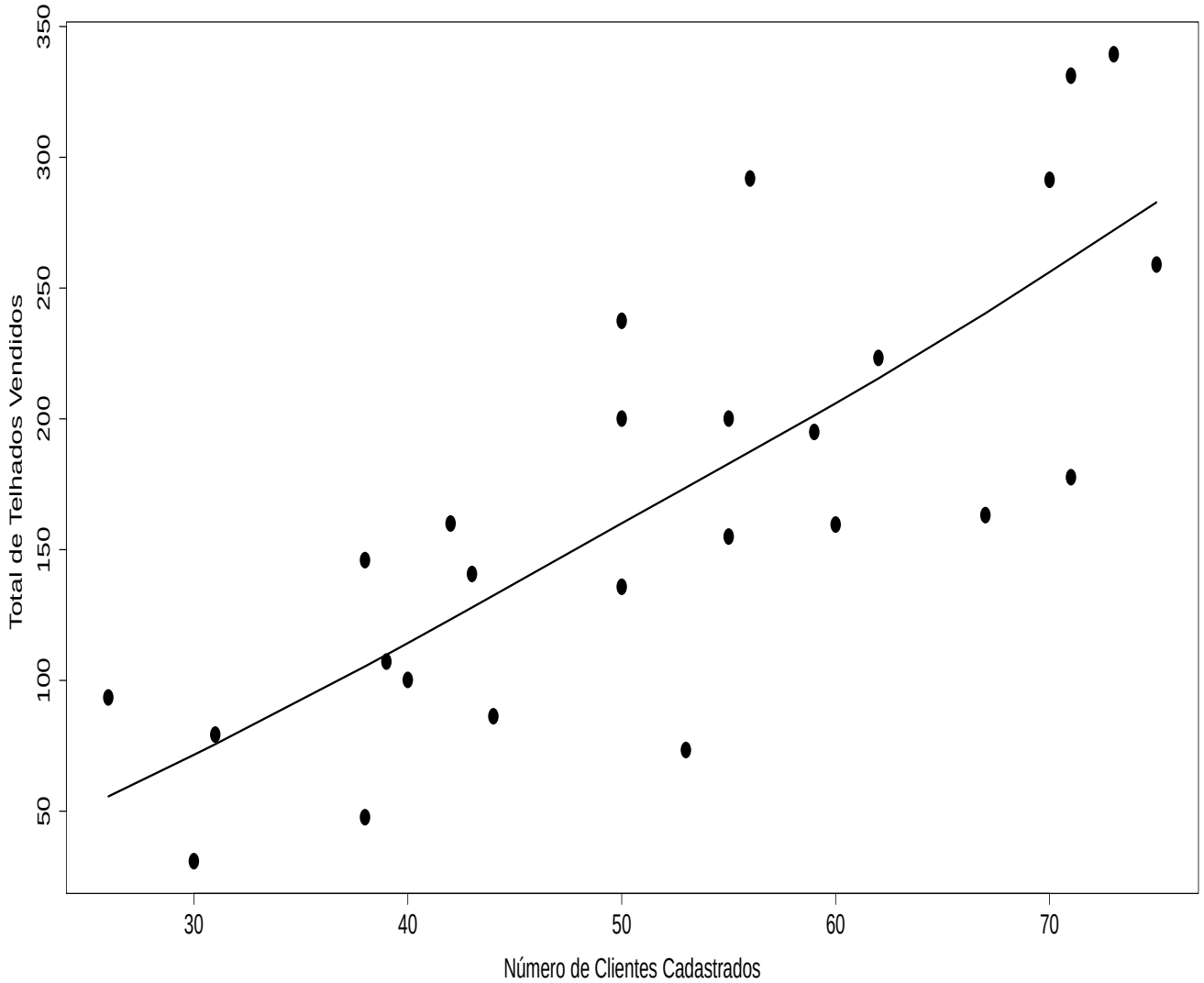
Matriz de Correlações Lineares

	Telhados	Gastos	Clientes	Marcas	Potencial
Telhados	1,0	0,159	0,783	-0,833	0.407
Gastos		1,0	0,173	-0,038	-0,070
Clientes			1,0	-0,324	0,468
Marcas				1,0	-0,202
Potencial					1,0

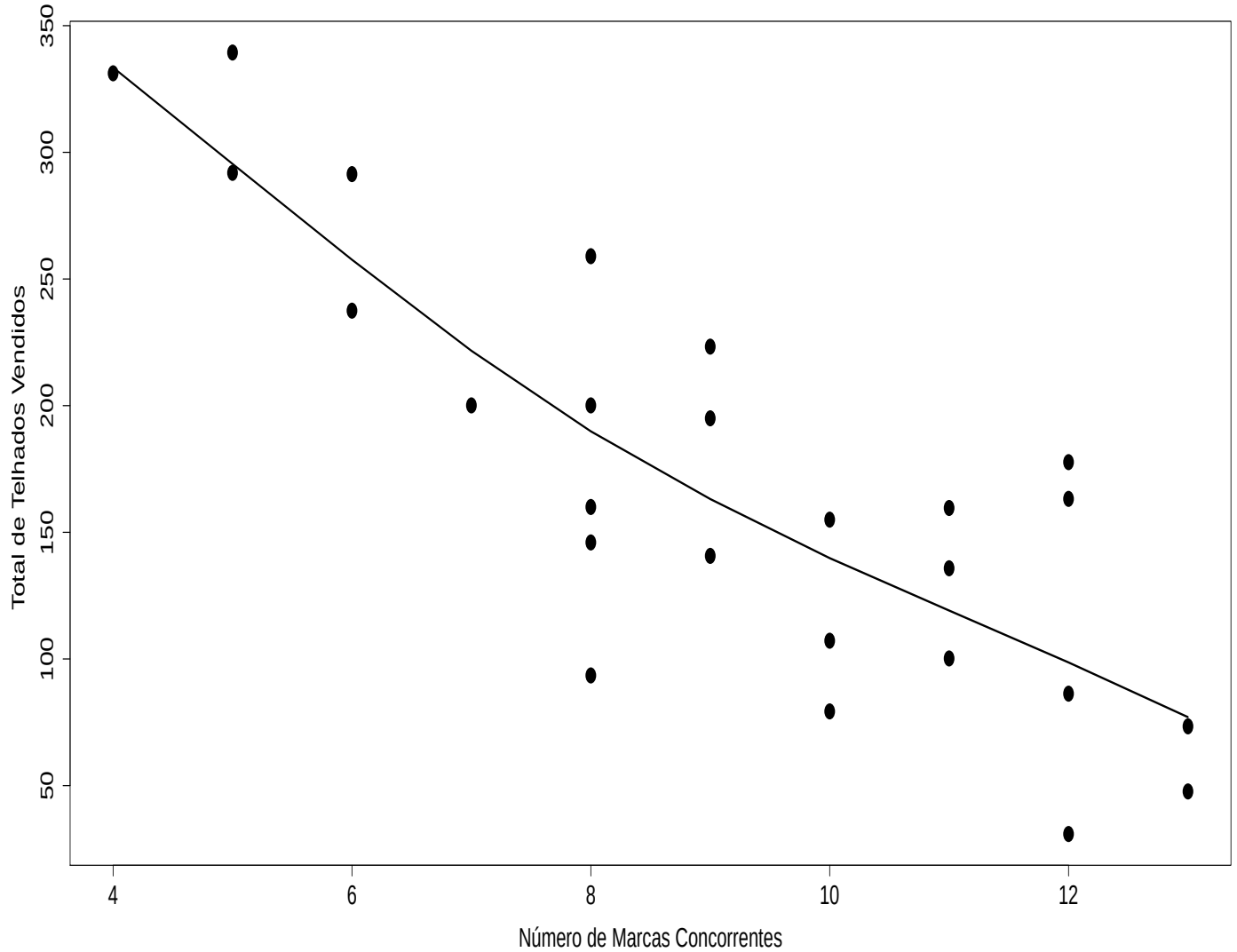
Dispersão Telhados versus Gastos



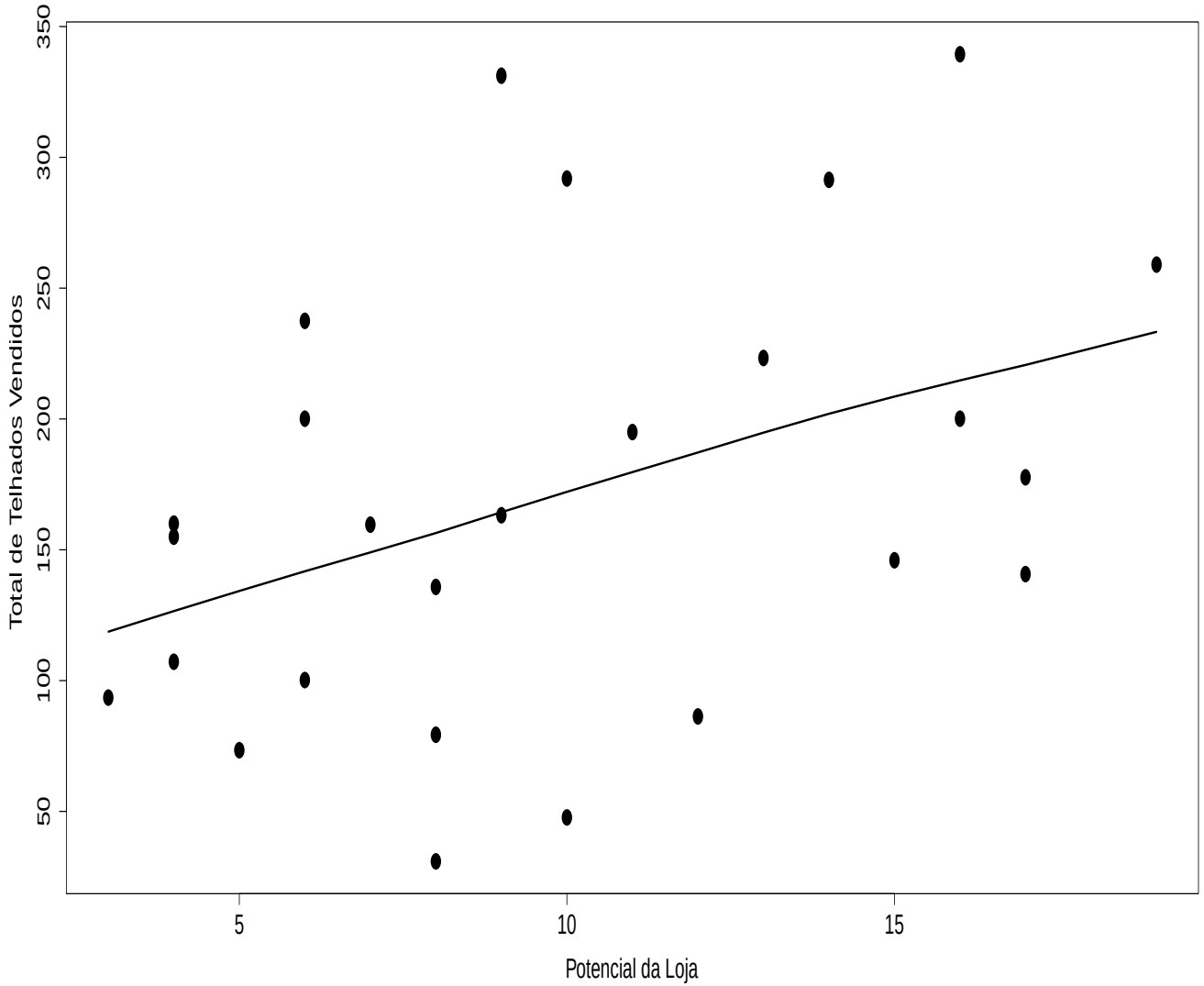
Dispersão Telhados versus Clientes



Dispersão Telhados versus Marcas



Dispersão Telhados versus Potencial



Correlações Lineares e Diagramas de Dispersão

- Nota-se uma tendência quadrática entre total de telhados vendidos e gastos com promoções, porém com correlação linear positiva baixa.
- Nota-se aumento do total de telhados vendidos com o aumento do número de clientes cadastrados e correlação linear positiva alta. Essa mesma tendência é observada com o potencial da loja, porém com correlação linear positiva moderada.
- Nota-se que o número de telhados vendidos decresce à medida que o número de marcas concorrentes cresce, com correlação linear negativa alta.
- As correlações lineares entre as variáveis explicativas são em geral pequenas. Observa-se correlação linear positiva moderada entre clientes e potencial.

Modelo Proposto

O seguinte modelo de regressão é inicialmente proposto:

$$y_i = \beta_1 + \beta_2 \times \text{Gastos} + \beta_3 \times \text{Clientes} + \beta_4 \times \text{Marcas} + \beta_5 \times \text{Potencial} + \epsilon_i,$$

em que y_i denota o total de telhados vendidos na i-ésima filial e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, 26$.

Sumário

Todas Regressões Possíveis

Descrição

Serão descritas na tabela a seguir as 16 regressões possíveis com as respectivas medidas resumo. A estatística PRESS será dividida pelo tamanho amostral n e a estimativa de σ^2 na estatística de Mallows será obtida da regressão selecionada pelo critério de Akaike.

Descrição

Submodelo[a]	k-1	k	R_k^2	s_k	C_k	$Press_k$
1	0	1	0,00	84,6	1960,2	7434,5
1 + G	1	2	0,025	85,2	1912,1	7829,8
1 + C	1	2	0,613	53,7	746,2	3115,0
1 + M	1	2	0,694	47,8	585,4	2428,8
1 + P	1	2	0,166	78.8	1633,1	6522,2
1 + G + C	2	3	0,613	54,8	747,0	3508,8
1 + G + M	2	3	0,710	47,5	555,4	2543,8
1 + G + P	2	3	0,201	78,8	1564,9	6770,1

[a]T:Telhados, G:Gastos, C:Clientes, M:Marcas, P:Potencial

Descrição

Submodelo	k-1	k	R^2_k	s_k	C_k	$Press_k$
1 + C + M	2	3	0,988	9,8	4,5	113,6
1 + C + P	2	3	0,615	54,7	744,0	3330,4
1 + M + P	2	3	0,753	43,8	469,3	2166,2
1 + G + C + M	3	4	0,989	9,5	4,0	115,4
1 + G + C + P	3	4	0,616	55,9	743,9	3726,5
1 + G + P + M	3	4	0,775	42,6	428,4	2222,4
1 + C + P + M	3	4	0,988	10,0	6,4	120,8
1 + G + C + P + M	4	5	0,989	9,6	5,5	119,5

Todas Regressões Possíveis

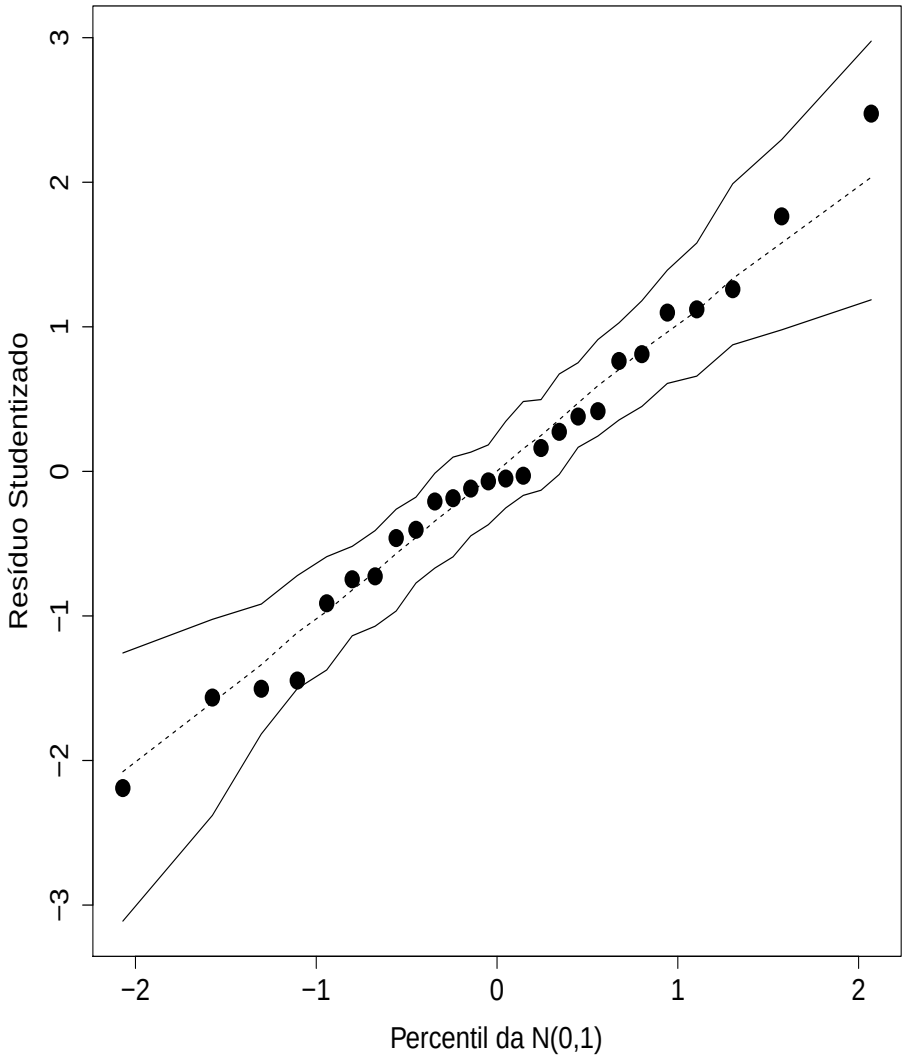
- Dois submodelos se destacam segundo os 4 critérios utilizados: 1 + Clientes + Marcas e 1 + Gastos + Cientes + Marcas.
- Levando-se em conta o número de variáveis explicativas o modelo 1 + Clientes + Marcas poderia ser escolhido.
- Todavia, deve-se fazer antes uma análise de diagnóstico com cada modelo.

Submodelo: 1 + Clientes + Marcas

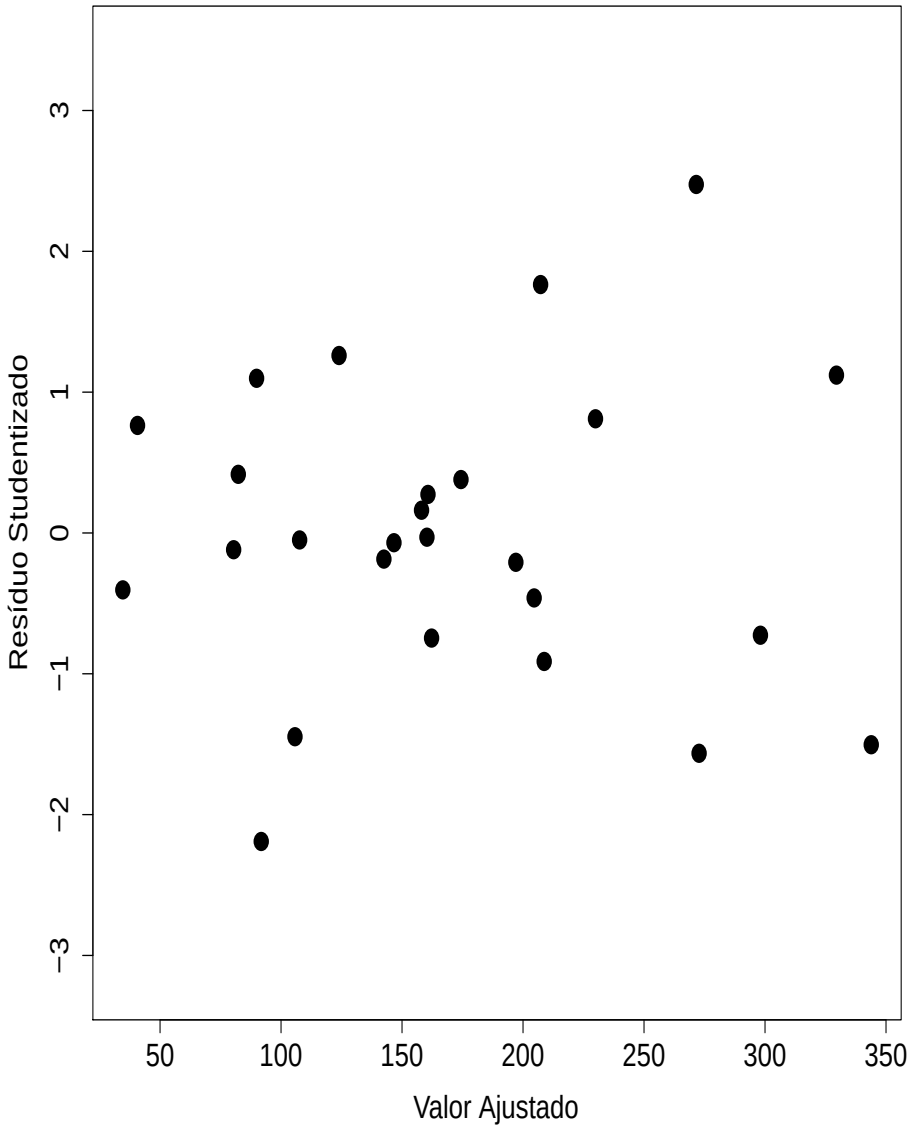
As estimativas dos parâmetros são dadas abaixo.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	186,694	12,259	15,23	0,00
Clientes	3,408	0,146	23,37	0,00
Marcas	-21,193	0,803	-26,40	0,00
s	9,803			
R^2	0,988			
$\overline{R}^2$	0,987			

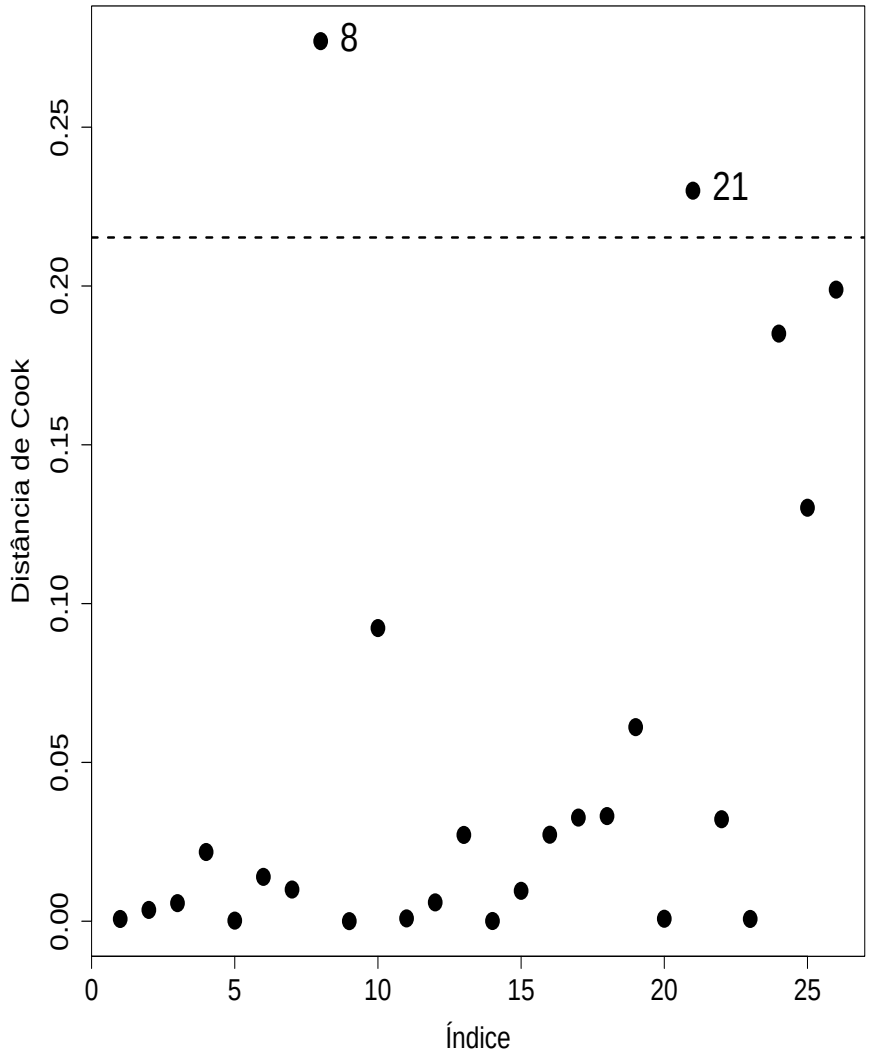
Análise de Resíduos



Análise de Resíduos



Distância de Cook



Análise Confirmatória Submodelo: 1 + Clientes + Marcas

Estimativa e valor-P com todos os pontos e variação percentual e valor-P elimanando-se cada observação destacada.

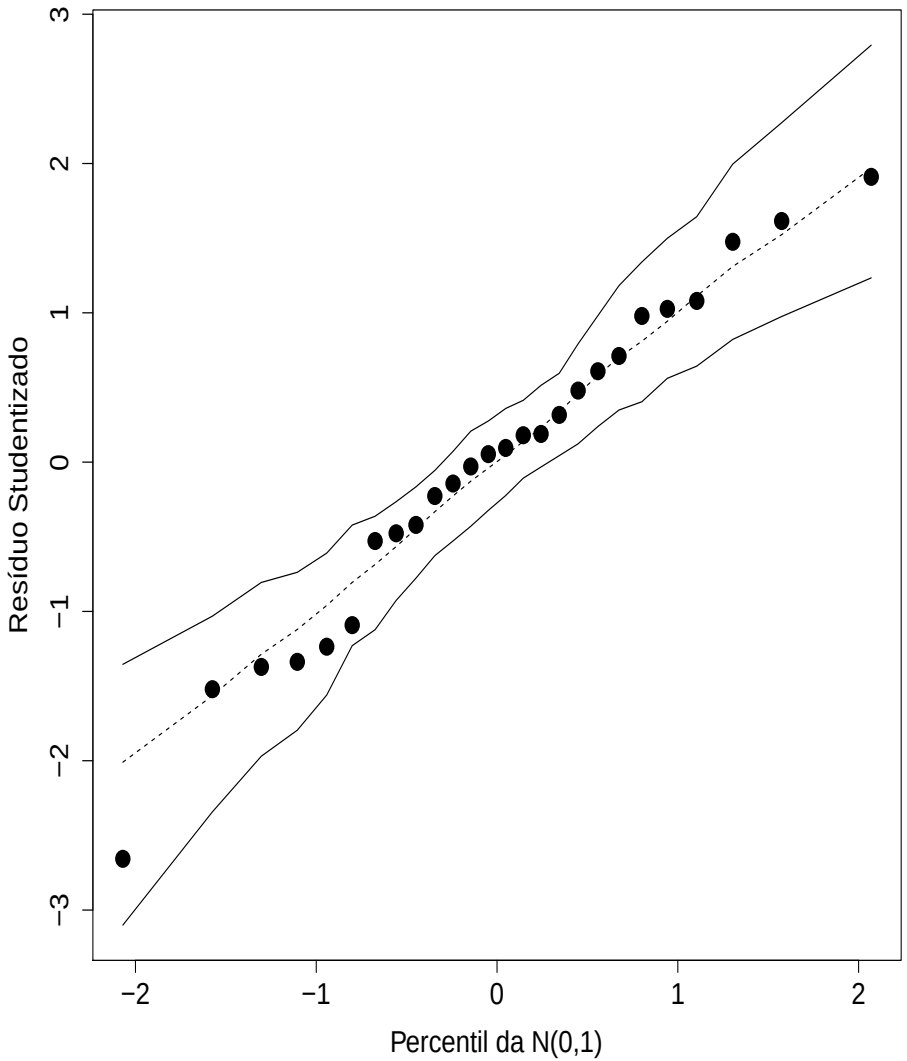
Efeito	Todos Ptos	Excluindo #8	Excluindo #21
Constante	186,694	3,95%	3,41%
	(0,00)	(0,00)	(0,00)
Clientes	3,408	0,49%	1,13%
	(0,00)	(0,00)	(0,00)
Marcas	-21,193	2,90%	2,69%
	(0,00)	(0,00)	(0,00)

Submodelo: 1 + Gastos + Clientes + Marcas

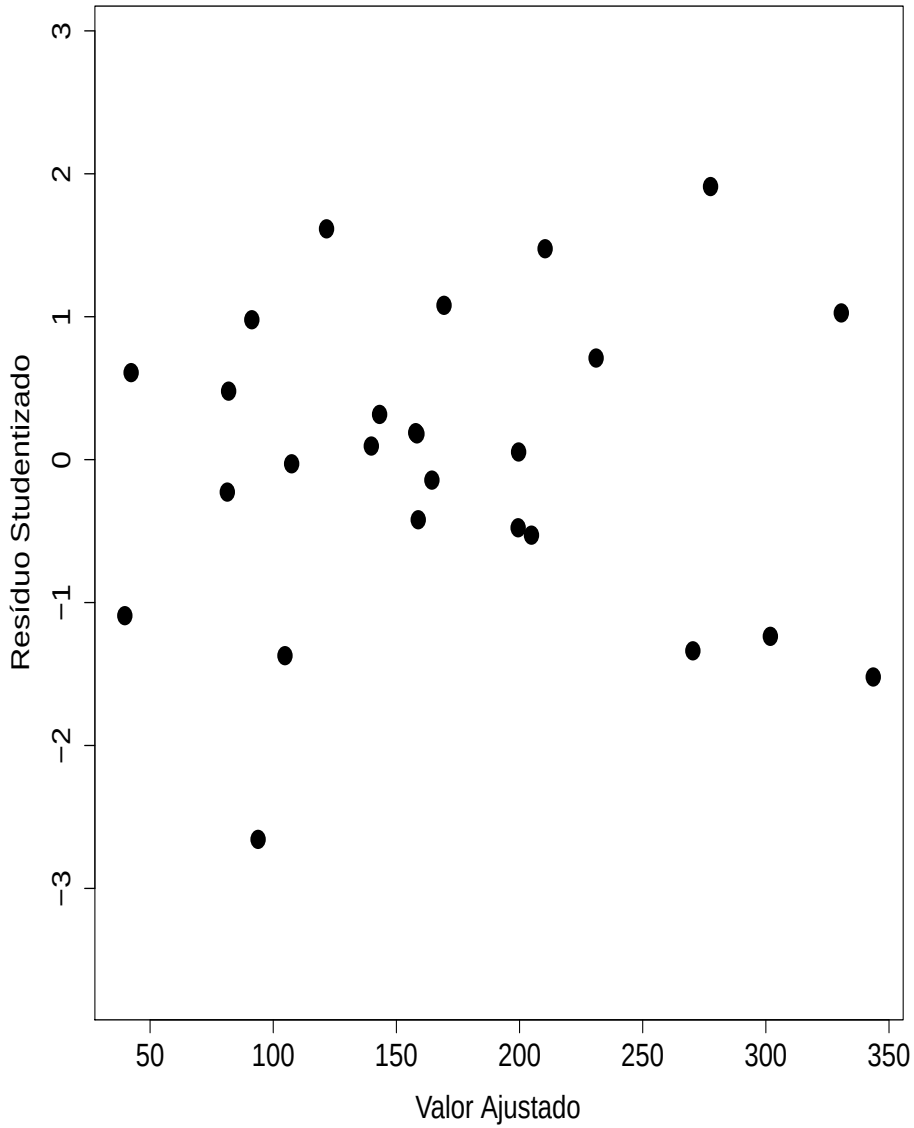
As estimativas dos parâmetros são dadas abaixo.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	179,844	12,621	14,25	0,00
Gastos	1,677	1,052	1,59	0,12
Clientes	3,369	0,143	23,52	0,00
Marcas	-21,217	0,773	-27,30	0,00
s	9,491			
R^2	0,989			
$\overline{R}^2$	0,987			

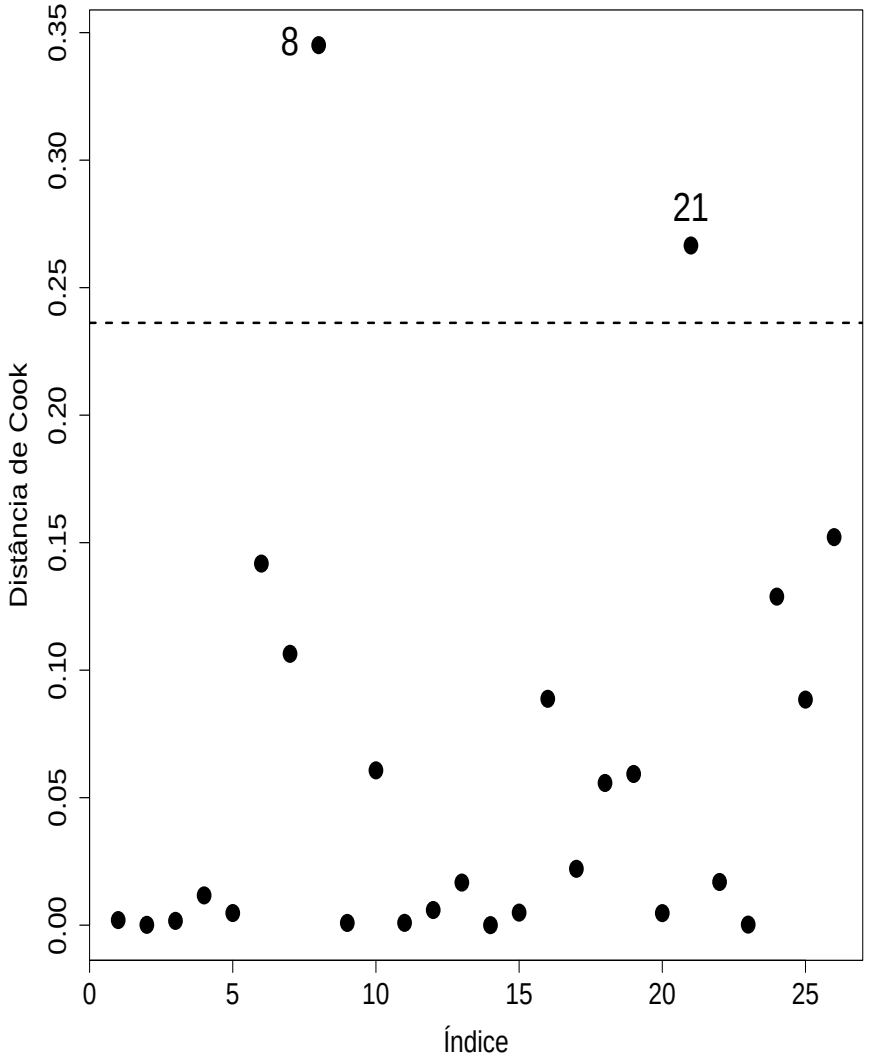
Análise de Resíduos



Análise de Resíduos



Distância de Cook



Análise Confirmatória Submodelo: 1 + Gastos + Clientes + Marcas

Estimativa e valor-P com todos os pontos e variação percentual e valor-P elimanando-se cada observação destacada.

Efeito	Todos Ptos	Excluindo #8	Excluindo #21
Constante	179,844	1,51%	4,84%
	(0,00)	(0,00)	(0,00)
Gastos	1,647	53,23%	21,64%
	(0,12)	(0,483)	(0,042)
Clientes	3,369	1,04%	1,05%
	(0,00)	(0,00)	(0,00)
Marcas	-21,217	2,56%	3,02%
	(0,00)	(0,00)	(0,00)

Todas Regressões Possíveis

- Os dois submodelos selecionados 1 + Clientes + Marcas e 1 + Gastos + Cientes + Marcas apresentaram excelentes ajustes, porém Gastos aparece marginalmente não significativa no $2^{\underline{o}}$ submodelo.
- Ambos os modelos destacam os mesmos pontos potencialmente influentes. A eliminação da observação #21 deixa Gastos significativa ao nível de 5%. Essa observação está mascarando o efeito de Gastos.
- Assim, deve-se escolher o submodelo 1 + Gastos + Clientes + Marcas.

Gráfico da Variável Adicionada de Clientes

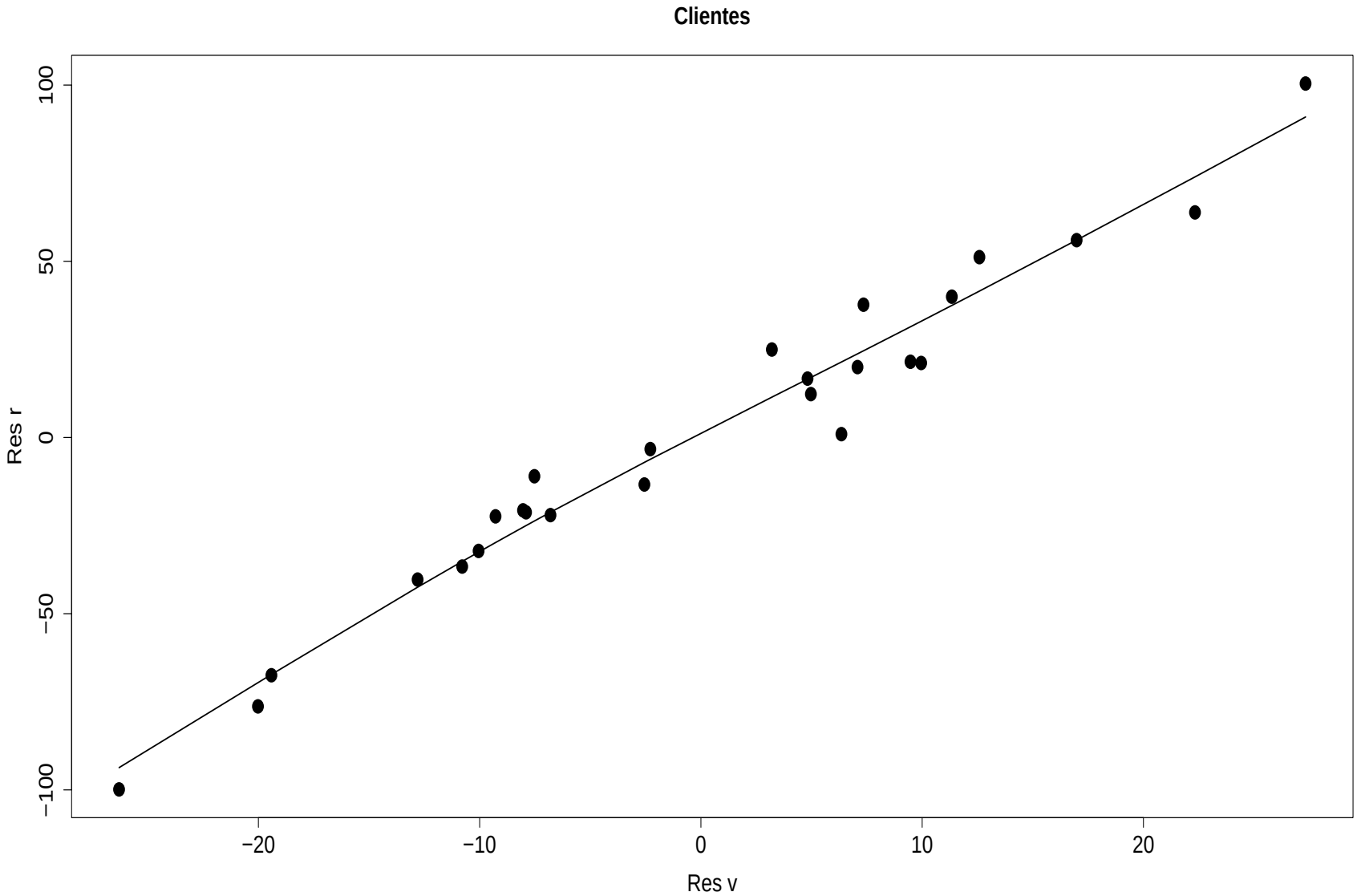


Gráfico da Variável Adicionada de Marcas

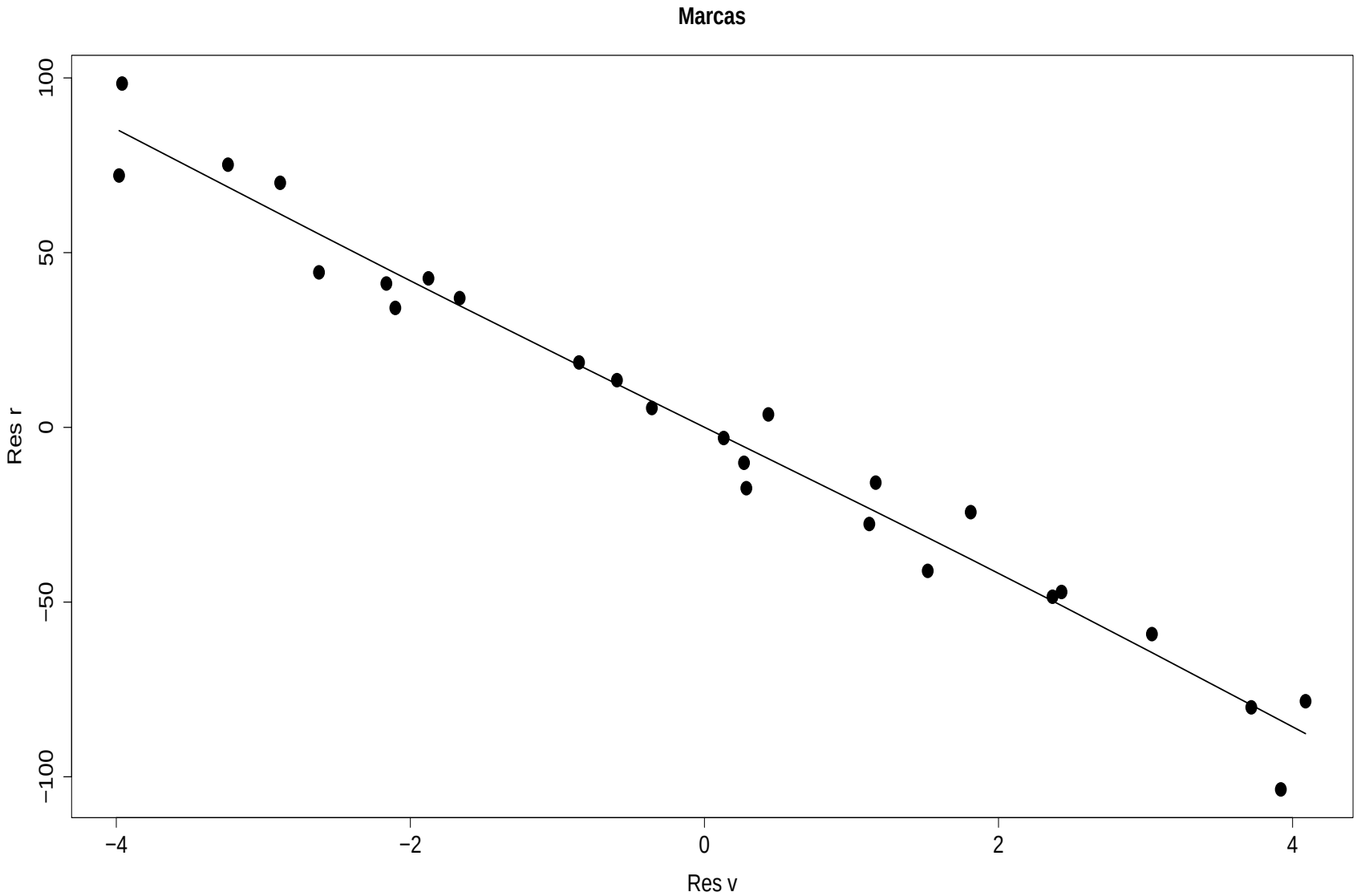


Gráfico da Variável Adicionada de Gastos

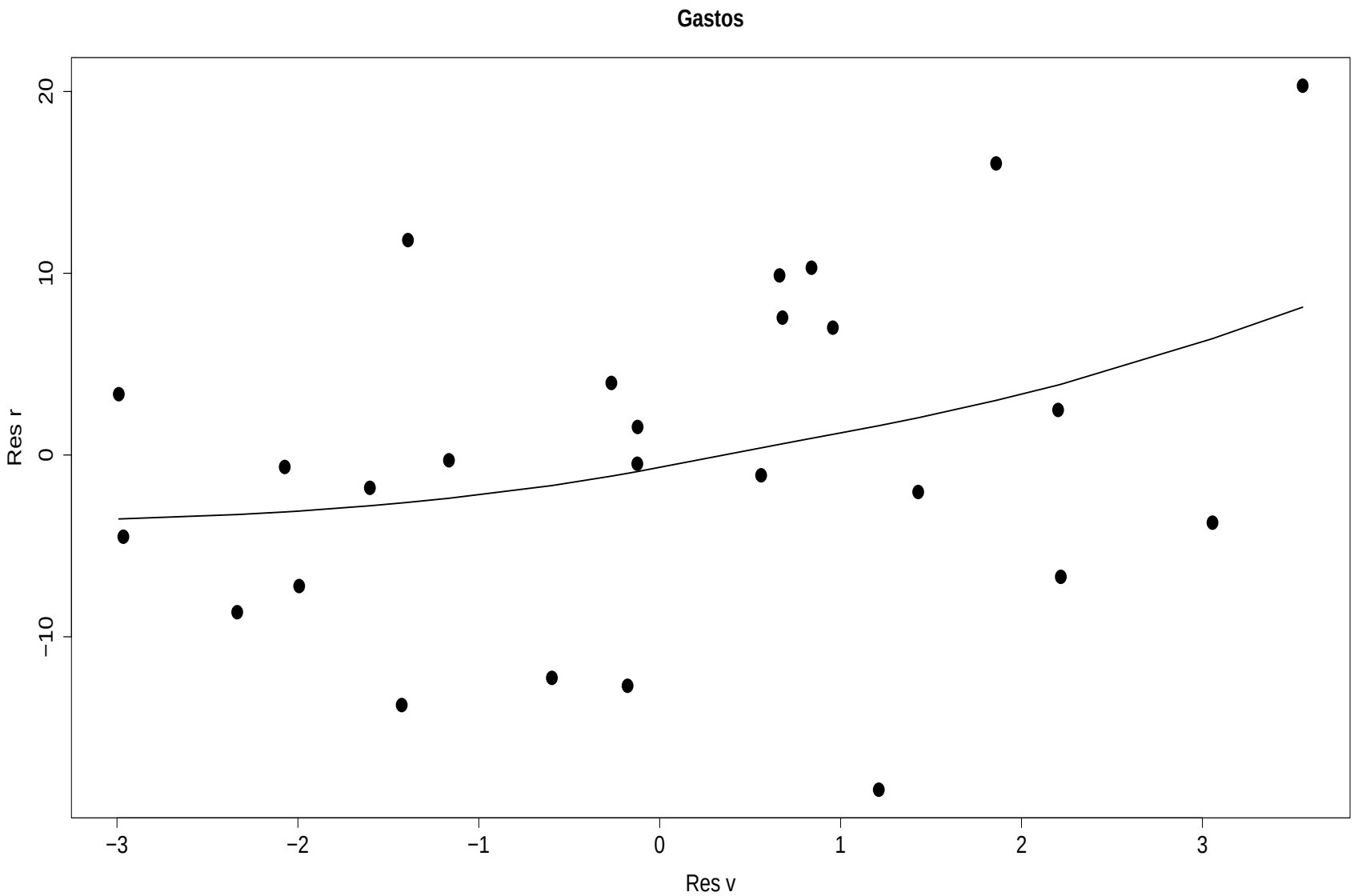


Gráfico da Variável Adicionada

- Para o submodelo selecionado 1 + Gastos + Clientes + Marcas os gráficos da variável adicionada referentes a Clientes e Marcas indicam que essas duas variáveis entraram de forma correta no modelo.
- O gráfico da variável adicionada para a variáveis Gastos indica que um termo quadrático poderia ser adicionado. Porém, a inclusão desse termo quadrático, dado que o termo linear está no modelo, mostrou-se não significativo.

Sumário

Descrição dos Passos supondo PE=PS=0,15

- Fazer todos os ajustes com apenas uma variável explicativa. Selecionar o ajuste com menor valor-P para o teste F. Se for menor do que PE entra no modelo a variável correpondente.

	Gastos	Clientes	Marcas	Potencial
Passo 1	0,4382	0,0000	0,0000	0,0389

- Fazer todos os ajustes com Marcas mais uma variável explicativa. Selecionar a variável com menor valor-P para o teste F. Se for menor do que PE essa variável entra no modelo.

	Gastos	Clientes	Potencial
Passo 2	0,2693	0,0000	0,0274

Descrição dos Passos supondo PE=PS=0,15

- Tentar retirar Marcas dado que Clientes está no modelo. Comparar o valor-P do teste F com PS. Se for maior do que PS a variável sai do modelo.

	Marcas
Passo 3	0,0000

- Fazer todos os ajustes com Marcas e Clientes mais uma variável explicativa. Selecionar a variável com menor valor-P para o teste F. Se for menor do que PE essa variável entra no modelo.

	Gastos	Potencial
Passo 4	0,1252	0,6968

Descrição dos Passos supondo PE=PS=0,15

- Tentar retirar Marcas e Clientes do modelo (uma de cada vez) dado que Gastos entrou no modelo. Comparar o maior valor-P do teste F com PS. Se for maior do que PS a variável sai do modelo.

	Marcas	Clientes
Passo 5	0,0000	0,0000

- Fazer o ajuste com Marcas, Clientes e Gastos no modelos mais o Potencial. Comparar o valor-P do teste F com PE. Se for menor do que PE essa variável entra no modelo, caso contrário encerra-se o processo.

	Potencial
Passo 6	0,4854

Resumo dos Passos

Passo	Gastos	Clientes	Marcas	Potencial
Passo 1	0,4382	0,0000	0,0000	0,0389
Passo 2	0,2693	0,0000	-	0,0274
Passo 3	-	-	0,0000	-
Passo 4	0,1252	-	-	0,6968
Passo 5	-	0,0000	0,0000	-
Passo 6	-	-	-	0,4854

Regressão Selecionada

Similarmente ao procedimento com todas as regressões possíveis, pelo critério stepwise com PE=PS=0,15 a regressão selecionada contém as variáveis explicativas Marcas, Clientes e Gastos.

Critério de Akaike

Seja $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$. Pelo critério de Akaike deve-se escolher o modelo com k coeficientes tal que AIC seja mínimo

$$\text{AIC} = -2\text{L}(\hat{\boldsymbol{\theta}}) + 2k.$$

A regressão selecionada contém as variáveis explicativas Marcas, Clientes e Gastos.

Conclusões

- Pelo critério de todas regressões possíveis as variáveis explicativas Marcas, Clientes e Gastos entraram no modelo com o auxílio de procedimentos de diagnóstico.
- Essas mesmas variáveis explicativas entraram no modelo pelo procedimento stepwise com PE=PS=0,15.
- O critério de Akaike também seleciona as mesmas variáveis explicativas.

Submodelo: 1 + Gastos + Clientes + Marcas

As estimativas dos parâmetros com as variáveis explicativas padronizadas são dadas abaixo.

Efeito	Estimativa	E.Padrão	valor-t	valor-P
Constante	170,208	1,861	91,45	0,00
Gastos	3,073	1,927	1,59	0,12
Clientes	47,895	2,036	23,52	0,00
Marcas	-54,779	2,007	-27,30	0,00
s	9,491			
R^2	0,989			
$\overline{R}^2$	0,987			

Sumário

Referências

Referência

- Montgomery, D. C.; Peck, E. A. e Vining, G. G. (2021, Capítulo 10). *Introduction to Linear Regression Analysis, 6th Edition*. Hoboken: Wiley.
- Neter, J.; Kutner, M. H.; Nachtsheim, C. J. e Wasserman, W.(1996). *Applied Linear Regression Models*, 3rd Edition. Irwin, Illinois.