Análise de Multicolinearidade

Prof. Caio Azevedo

Exemplo 10: risco de assegurar automóveis

- O conjunto de dados em questão diz respeito ao estudo do risco (para a seguradora) de assegurar determinado veículo, o qual varia no intervalo $\{-3, -2, -1, 0, 1, 2, 3\}$ (resposta), em função de diversas características do veículo (covariáveis). Quanto maior o valor, maior o risco.
- Podemos conjecturar, com bastante segurança que, independentemente de quaisquer outras questões, os MRNLH não são apropriados para analisar este problema, haja vista a natureza categorizada da variável resposta.
- Uma alternativa são os modelos de regressão para reposta multinomial (aqui). Veja também aqui e aqui.

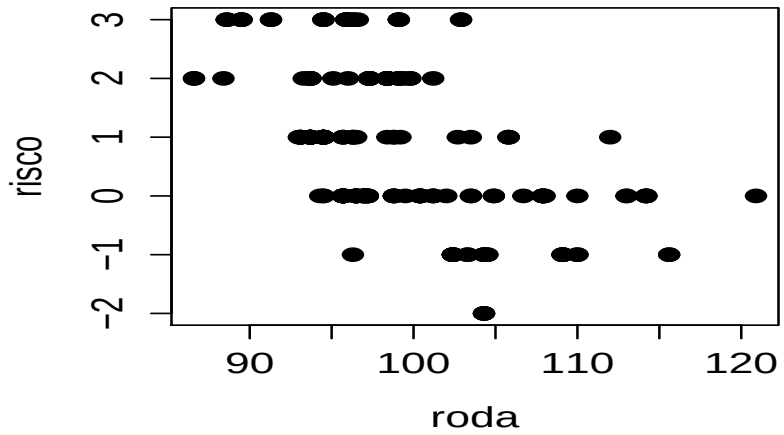
Exemplo 10 (Características de interesse: covariáveis)

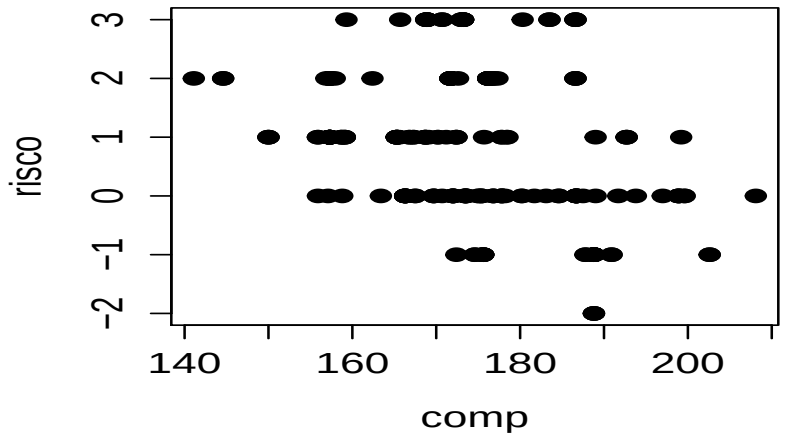
- base da roda (dimensão, em polegadas).
- comprimento do carro (em polegadas).
- largura do carro (em polegadas).
- altura do carro (em polegadas).
- peso do freio (em onças).
- tamanho do motor (em polegadas).
- diâmetro do carro (em polegadas).
- "stroke" (arranque).
- taxa de compressão (performance).
- cavalo-vapor (potência).
- pico-rpm (potência).
- consumo urbano (milhas por galão).
- consumo estrada (milhas por galão).

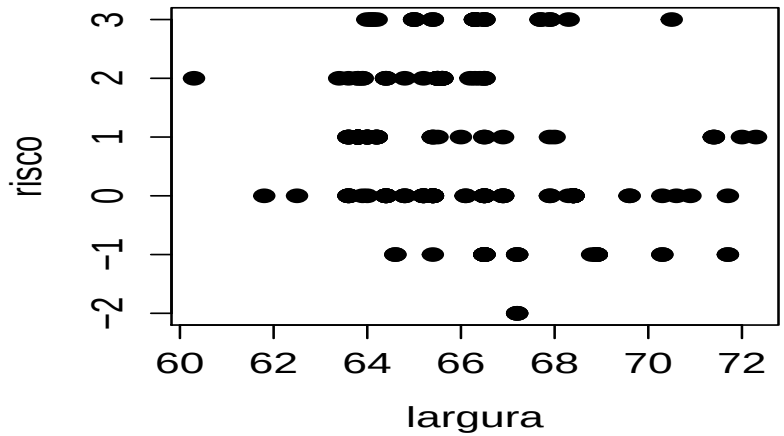
Introdução

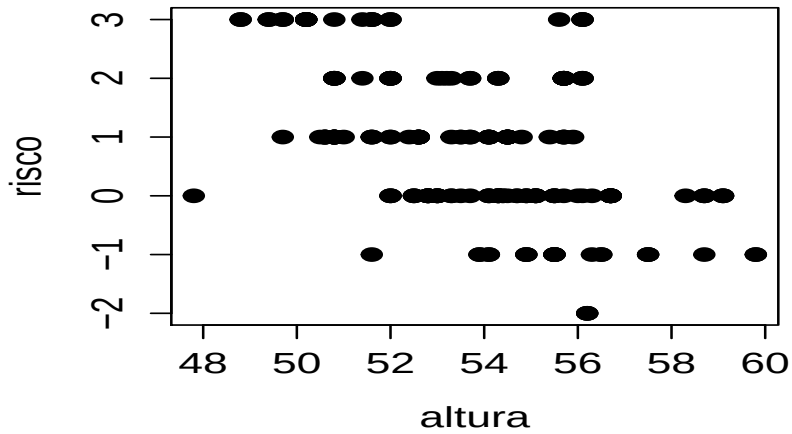
- Em geral, carros de marcas mais visadas e/ou caras, os quais terão riscos maiores, apresentarão certos padrões em relação às covariáveis em questão (tamanho, potência e consumo).
- No entanto, podemos notar que algumas variáveis partilham a mesma natureza e/ou podem estar (muito) correlacionadas:
  - Tamanho: comprimento, largura e altura
  - Potência: cavalo-vapor e pico-rpm.
  - Consumo: urbano e estrada.
- Faz sentido e/ou há problemas em considerar todas as covariáveis?

Gráficos de dispersão

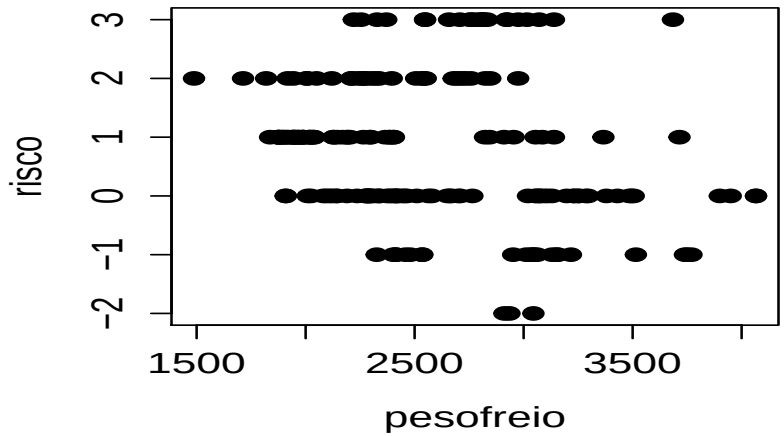


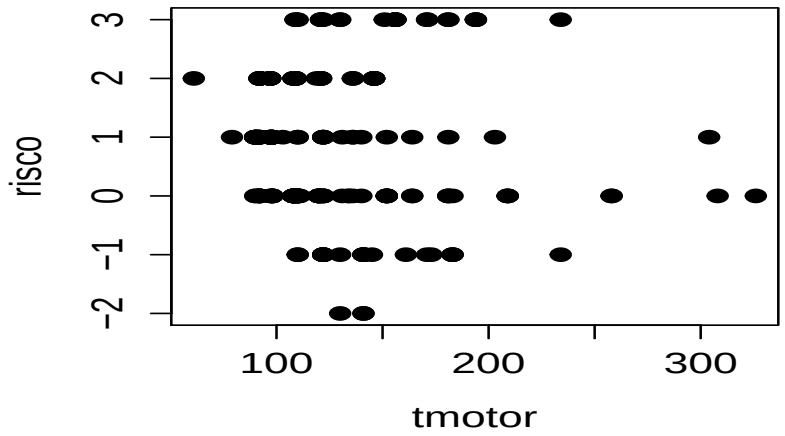


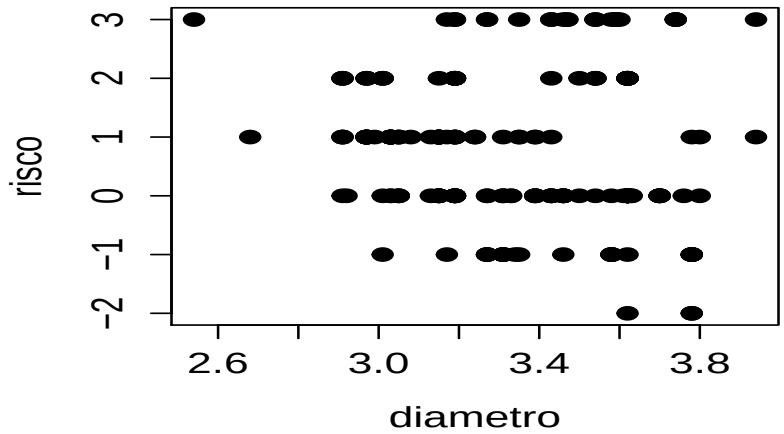


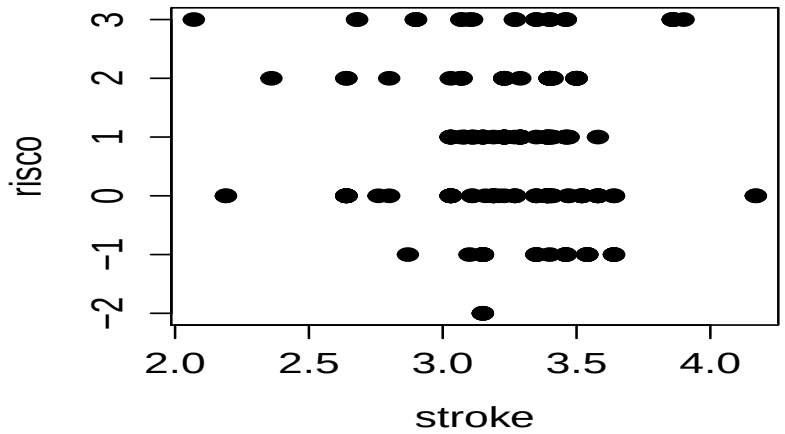


Gráficos de dispersão

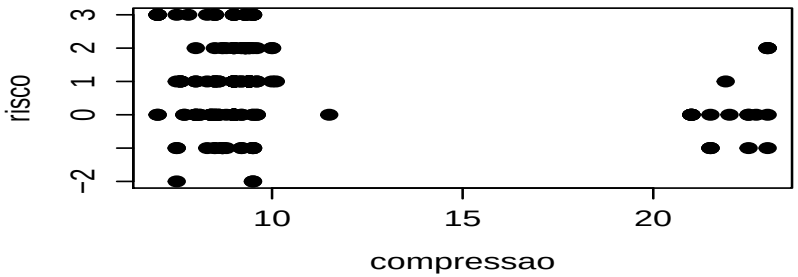


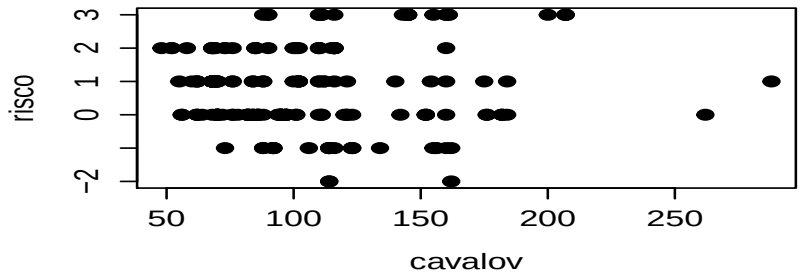


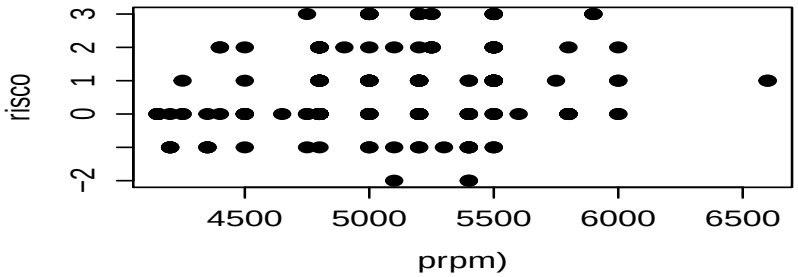


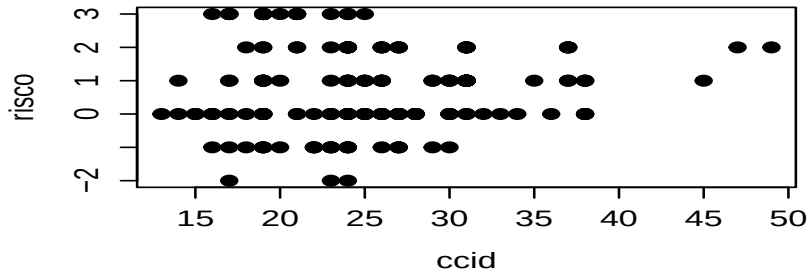


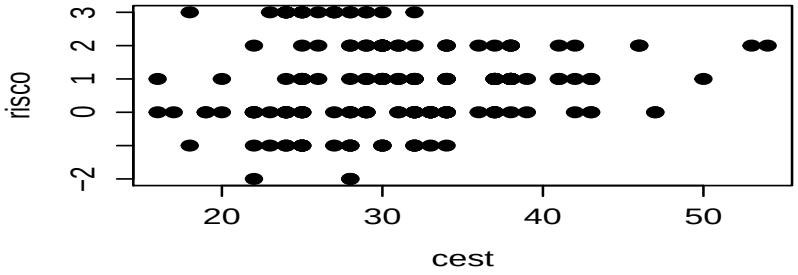
Gráficos de dispersão



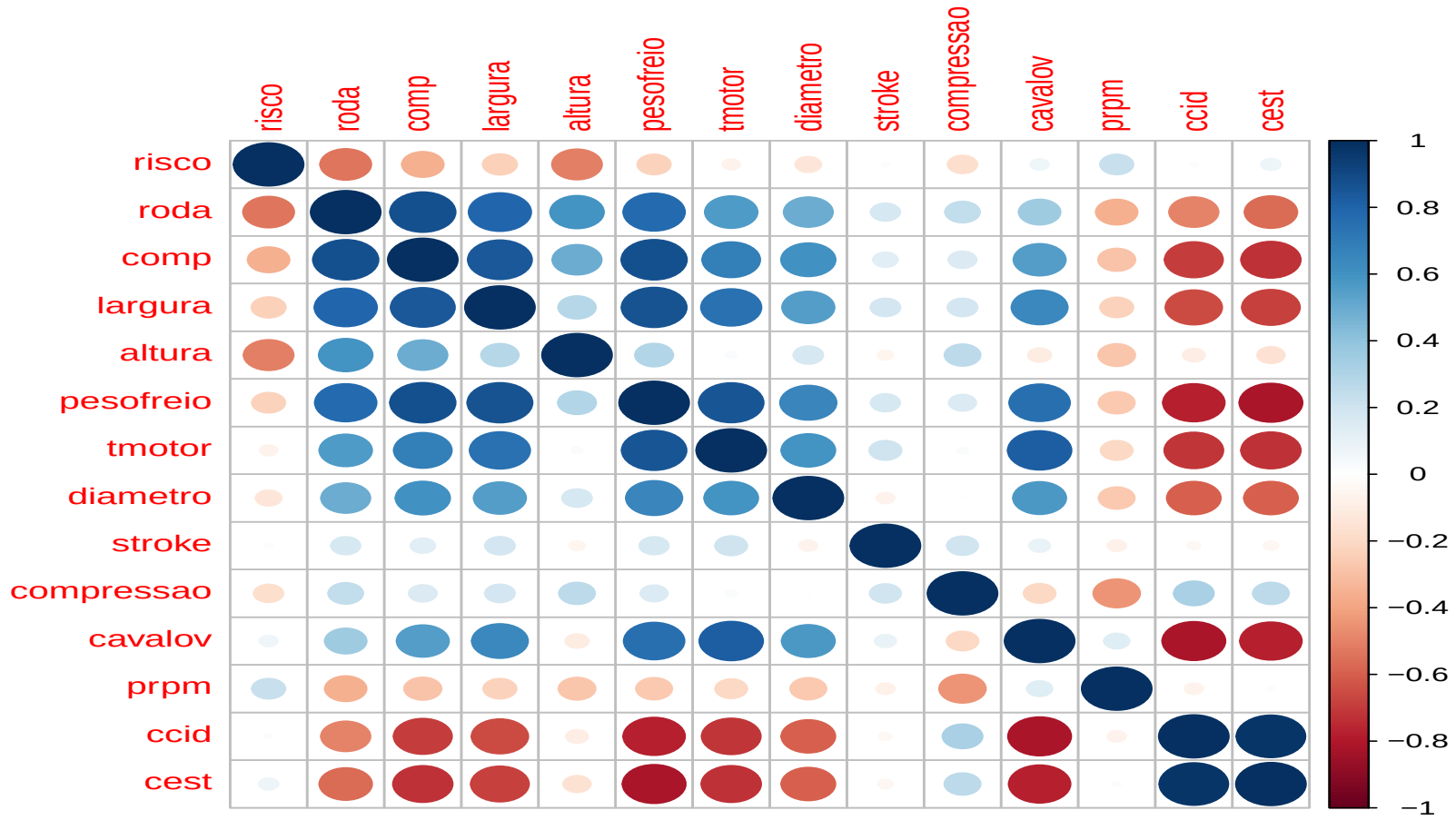




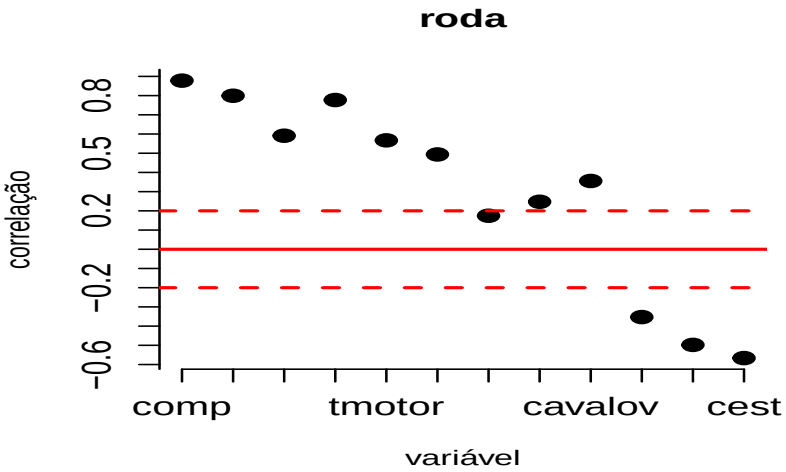


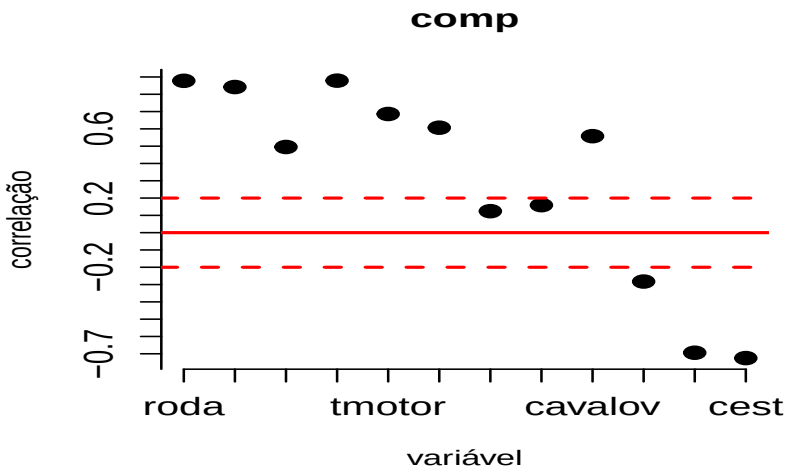


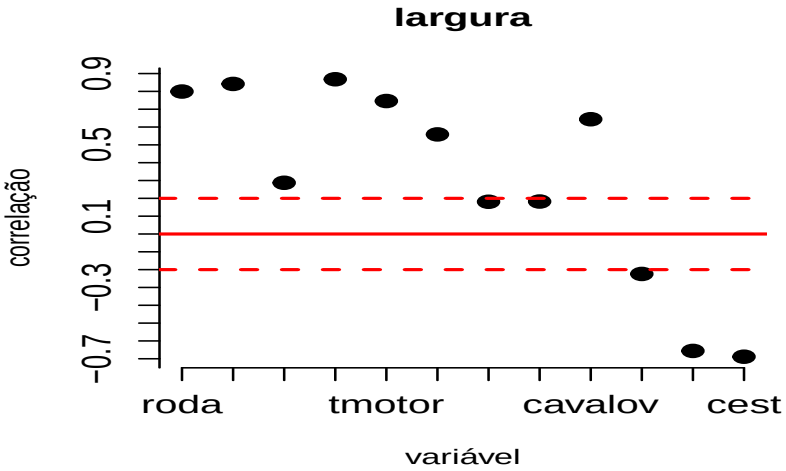
Mapa de calor das correlações

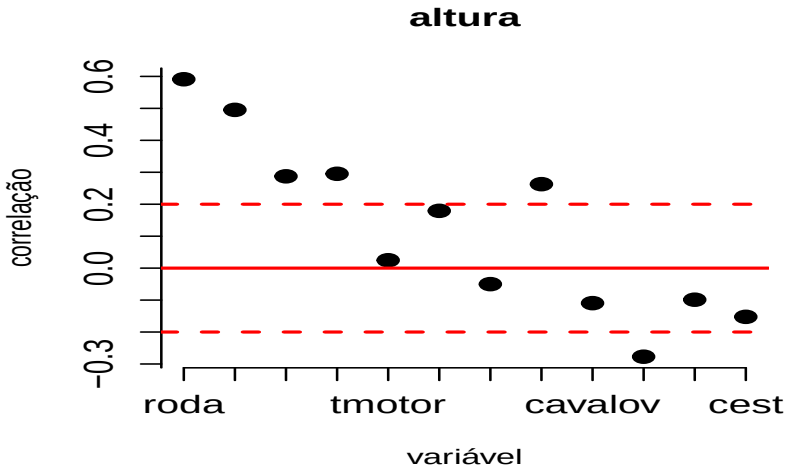


Correlação entre as covariáveis

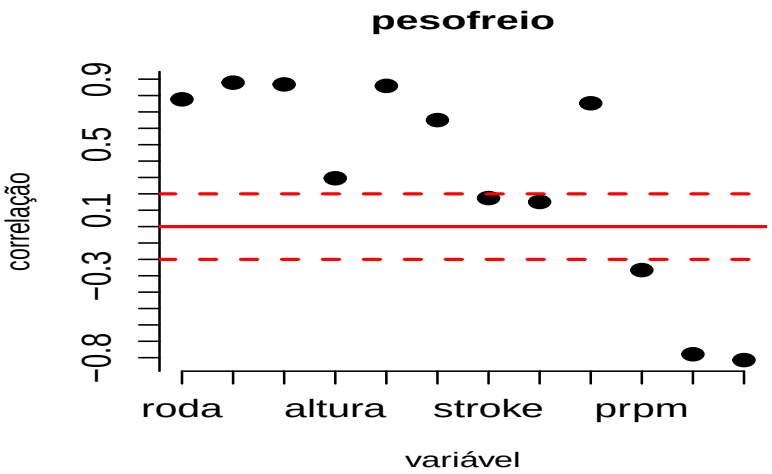


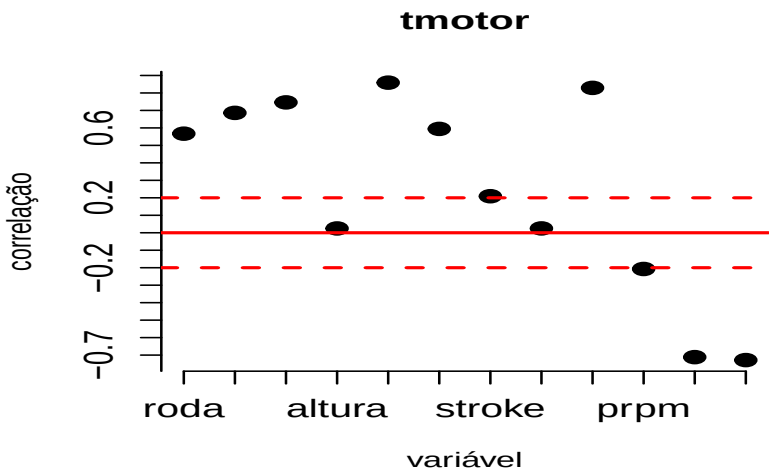


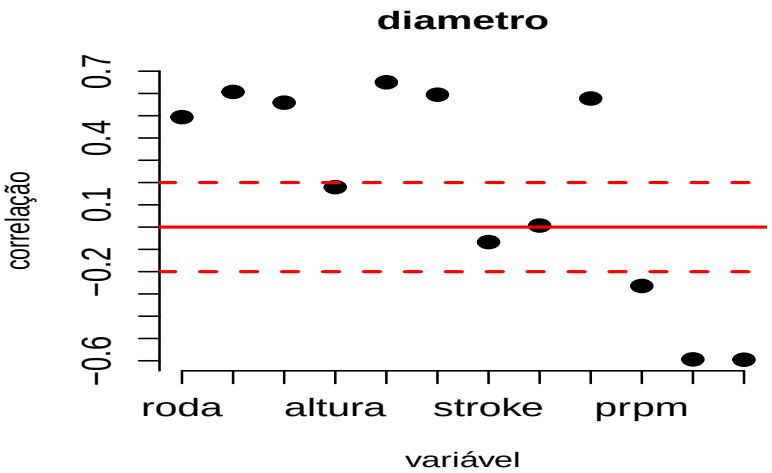


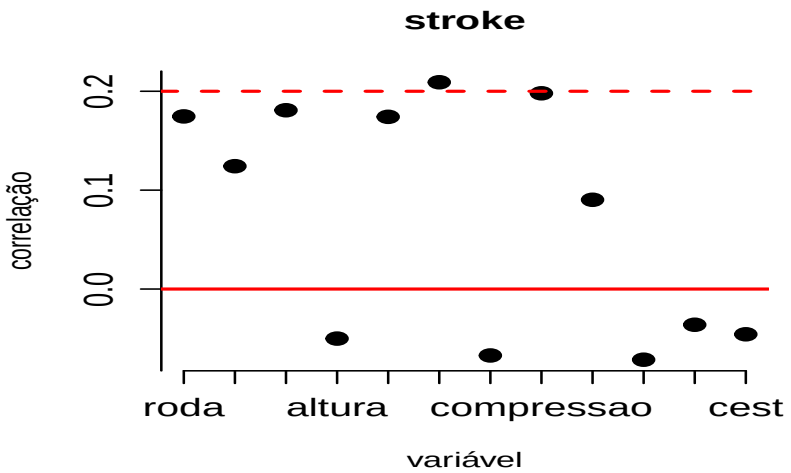


Correlação entre as covariáveis

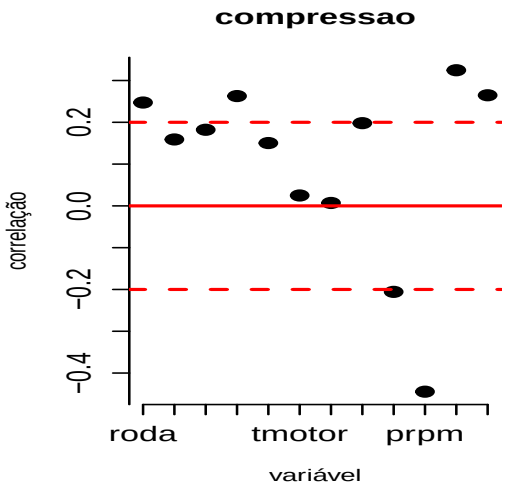


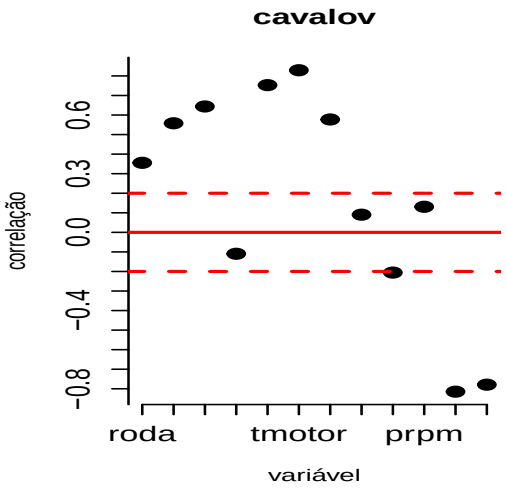


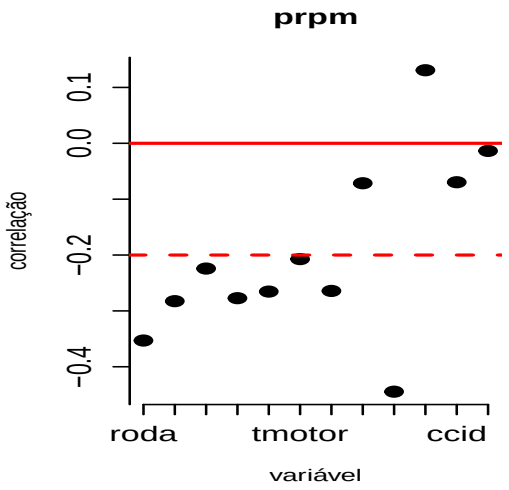


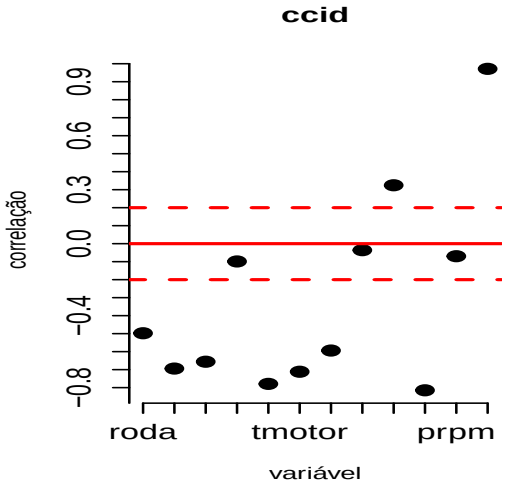


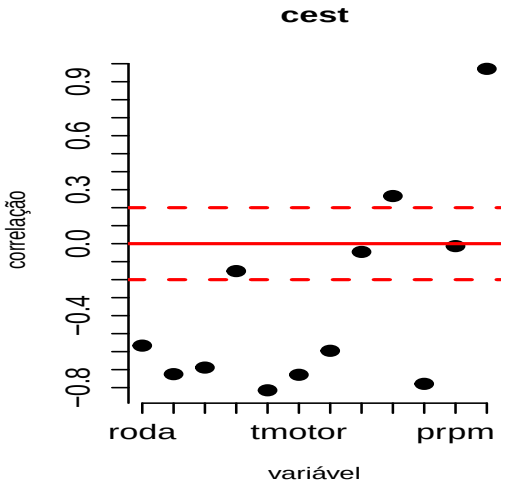
Correlação entre as covariáveis









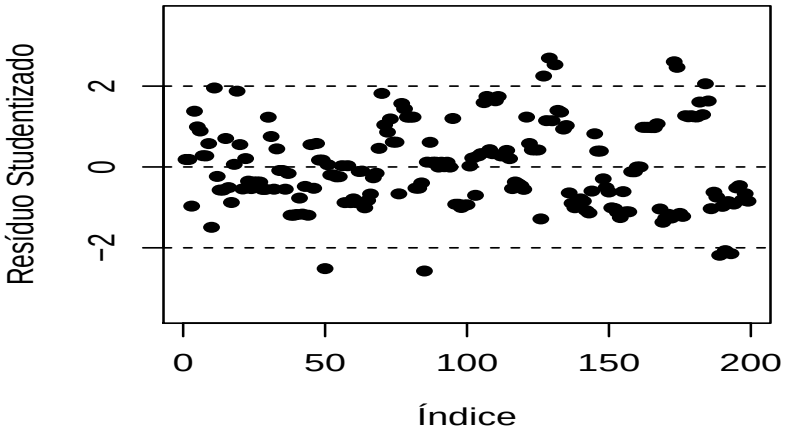


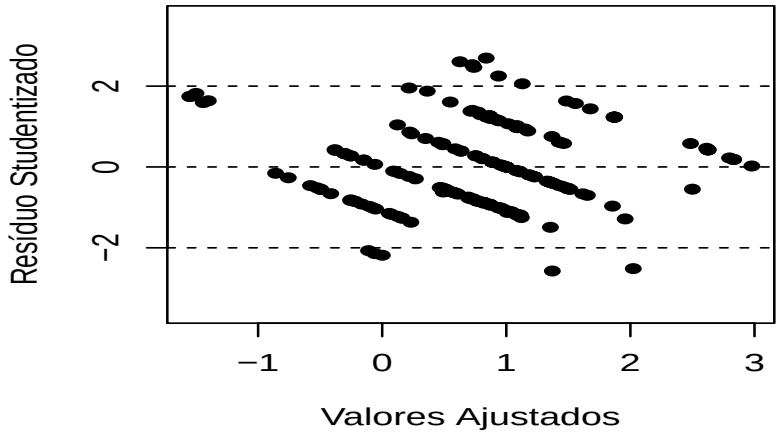
Modelo completo

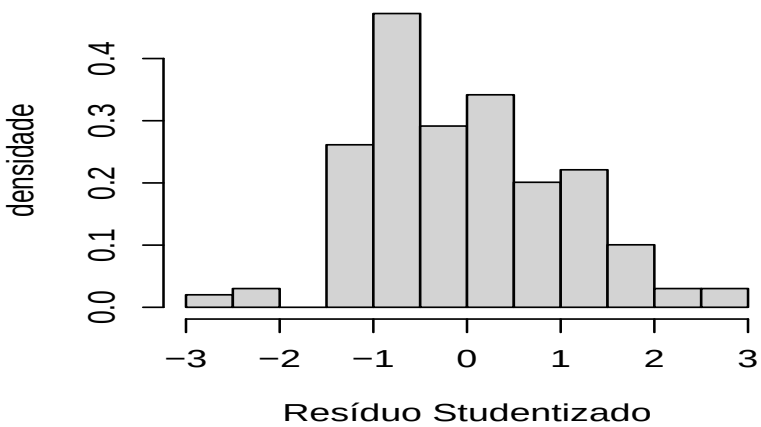
$$
\begin{aligned}
risco_i &= \beta_0 + roda_i\beta_1 + comp_i\beta_2 + largura_i\beta_3 + altura_i\beta_4 + pesofreio_i\beta_5 \\
&+ tmotor_i\beta_6 + diametro_i\beta_7 + stroke_i\beta_8 + compressao_i\beta_9 \\
&+ cavalov_i\beta_{10} + prpm_i\beta_{11} + ccid_i\beta_{12} + cest_i\beta_{13} + \xi_i
\end{aligned}
$$

$\xi \overset{i.i.d.}{\sim} N(0, \sigma^2)$, i=1,2,...,199 enquanto que os parâmetros $\beta_j, j = 1, 2, ..., 13$ seguem as interpretações usuais (escrevê-las). Por facilidade, não utilizaremos as covariáveis padronizadas (embora seja a abordagem mais apropriada). Exercício: interpretar os parâmetros do modelo também sob as covariáveis padronizadas.

Gráficos de resíduos







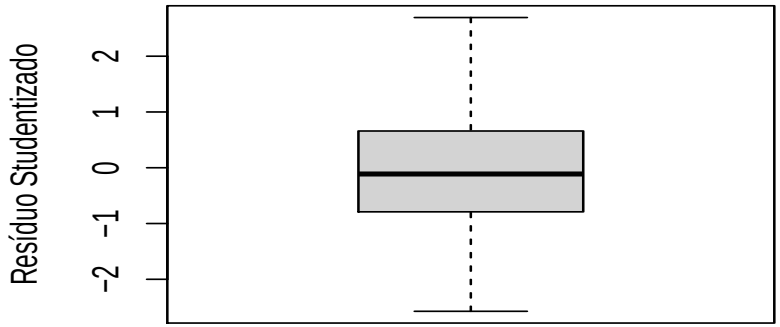
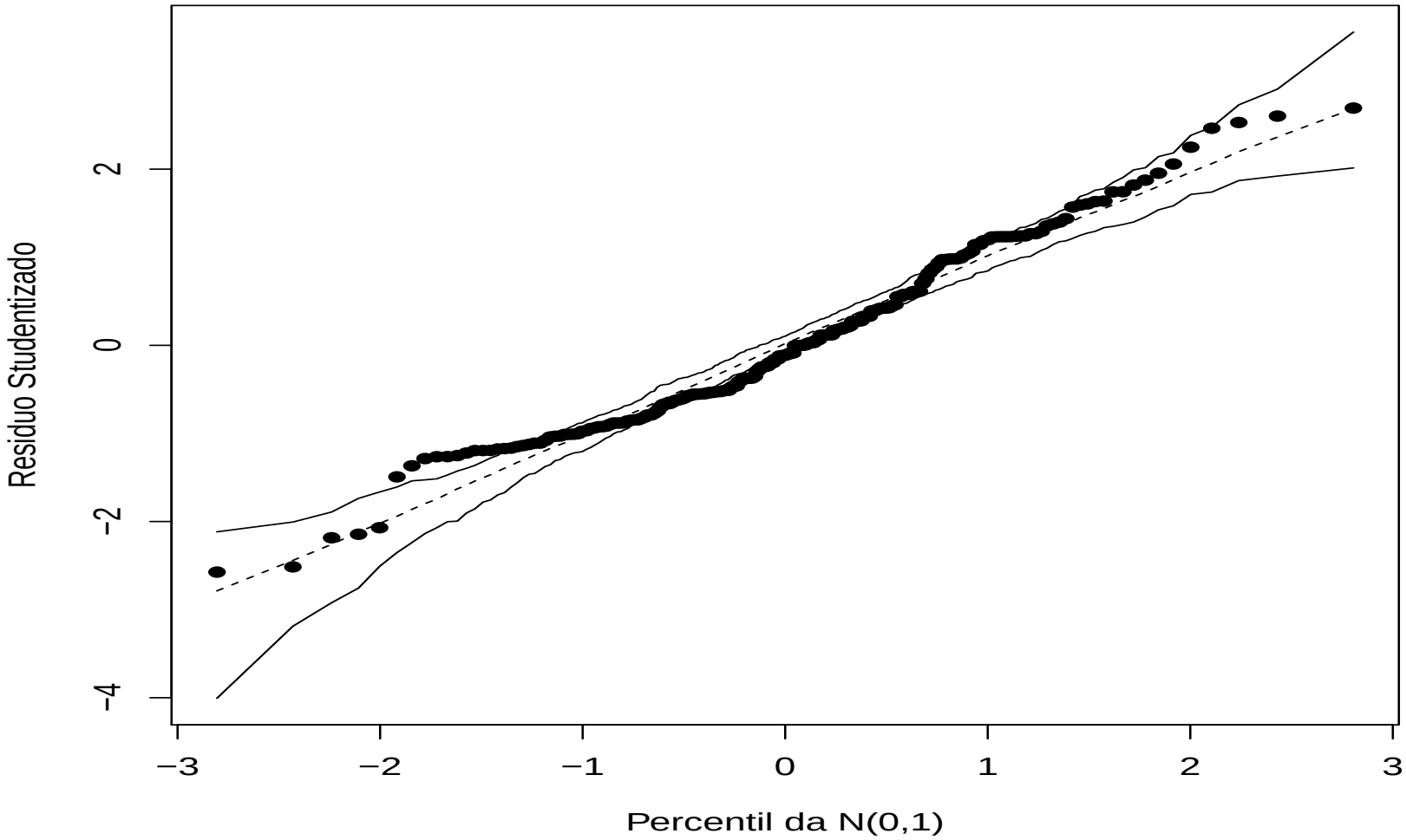


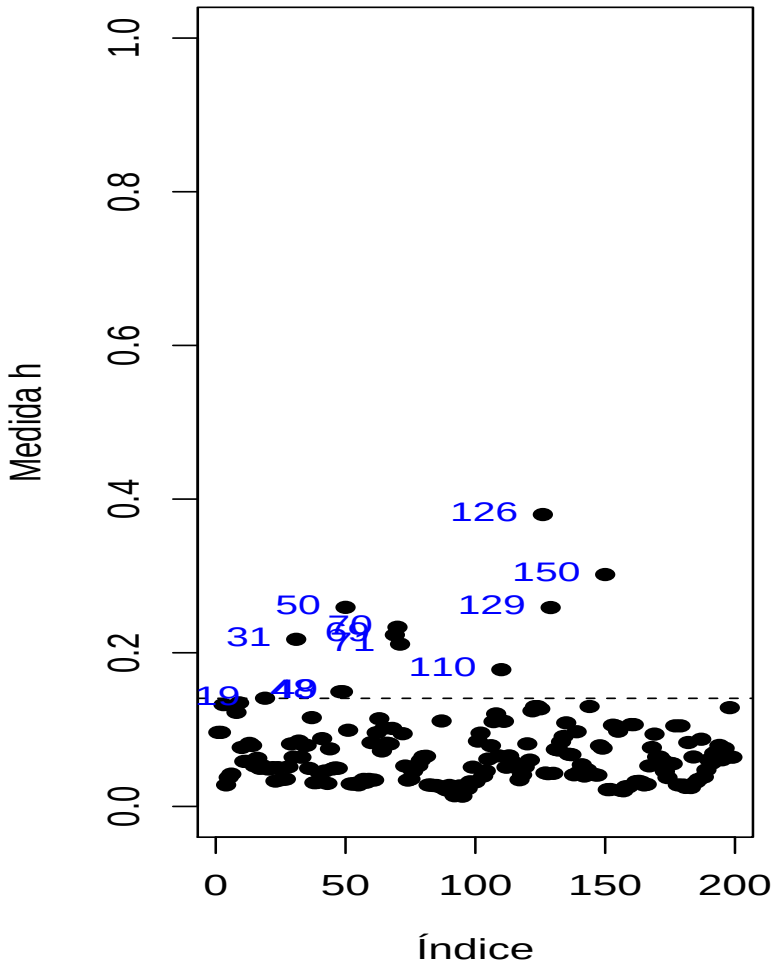
Gráfico de envelopes para os resíduos

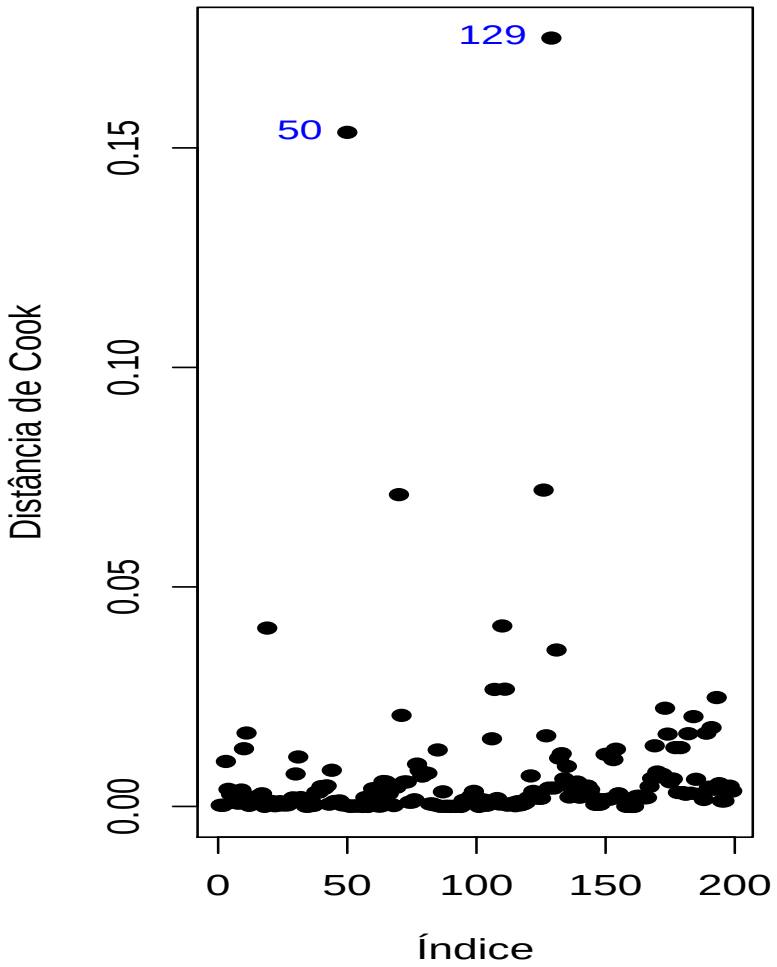


Ajuste do modelo (MRNL) com as variáveis originais

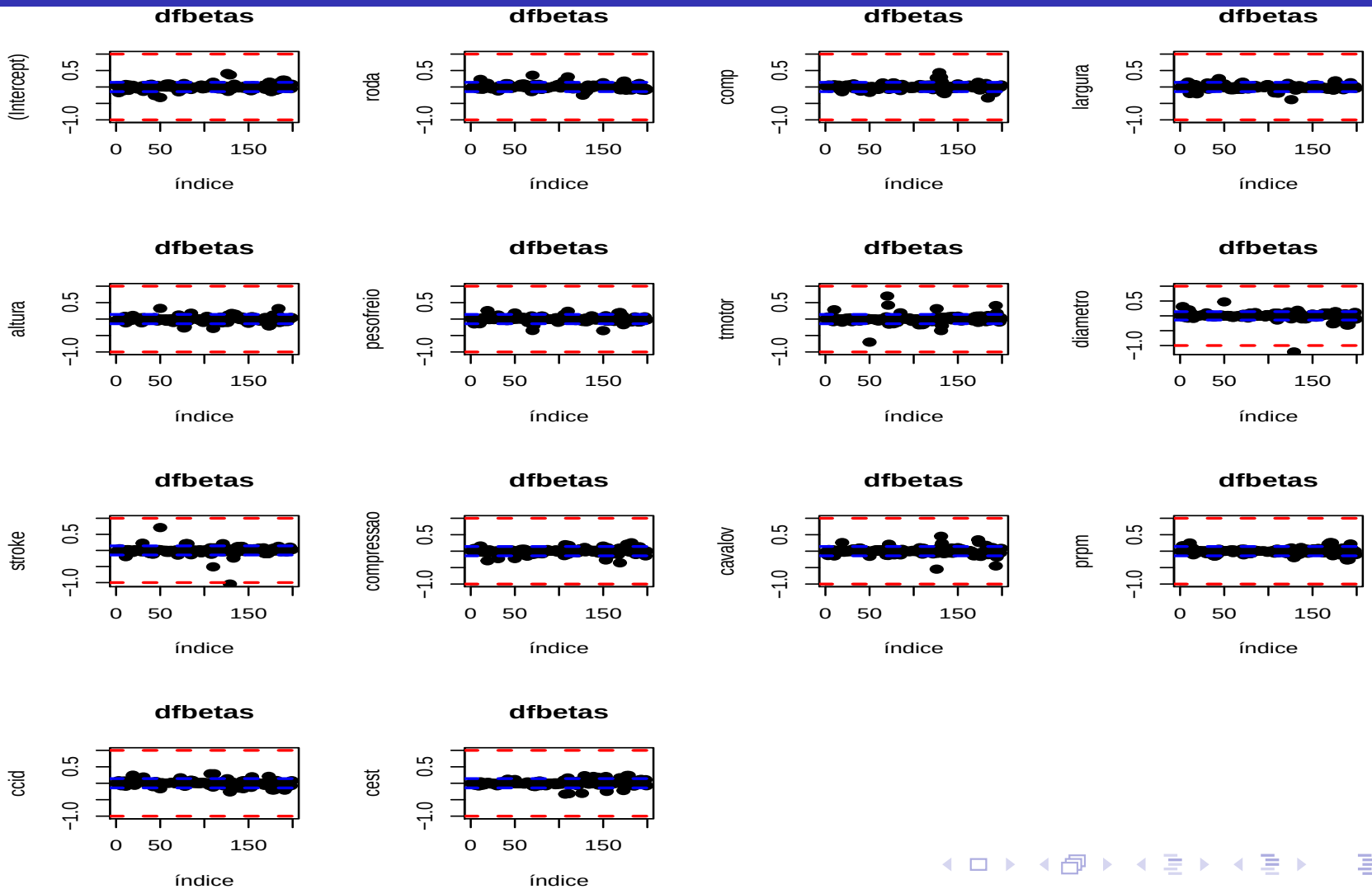
Variável	Estimativa	EP	IC(95%)	Estat. t	p-valor
Intercepto	12,446	4,630	[3,313 ; 21,580]	2,688	0,0078
roda	-0,173	0,031	[-0,234 ; -0,113]	-5,640	$<$0,0001
comp	0,005	0,017	[-0,028 ; 0,038]	0,323	0,7468
largura	0,146	0,074	[0,001 ; 0,291]	1,980	0,0491
altura	-0,084	0,042	[-0,167 ; 0,000]	-1,972	0,0501
pesofreio	0,000	0,001	[-0,001 ; 0,001]	0,057	0,9546
tmotor	0,004	0,004	[-0,004 ; 0,013]	1,037	0,3010
diametro	-0,041	0,359	[-0,748 ; 0,666]	-0,115	0,9087
stroke	0,142	0,238	[-0,328 ; 0,611]	0,595	0,5524
compressao	0,027	0,026	[-0,023 ; 0,078]	1,072	0,2849
cavalov	-0,007	0,005	[-0,017 ; 0,003]	-1,381	0,1689
prpm	$3{,}906 \times 10^{-5}$	$2{,}009 \times 10^{-4}$	$[< 0{,}001; 4{,}354 \times 10^{-4}]$	0,194	0,8461
ccid	-0,111	0,054	[-0,218 ; -0,005]	-2,068	0,0400
cest	0,049	0,048	[-0,045 ; 0,144]	1,029	0,3047

Análise de Influência





DFBETAS


dfbetas
(Intercept)
roda
comp
largura
altura
pesofreio
tmotor
diametro
stroke
compressao
cavalov
prpm
ccid
cest
índice
0
50
150
-1.0
0.5

Problemas provocados pela multicolinearidade

- Se houver alguma estrutura de correlação (significativa) entre as covariáveis, significa que uma ou mais pode(m) ser escrita(s) como uma combinação linear de uma ou outras covariáveis.
- Assim, o determinante da matriz $\boldsymbol{X}'\boldsymbol{X}$ pode ser próximo de zero.
- A inversibilidade de $\boldsymbol{X}'\boldsymbol{X}$ fica, assim, comprometida, e pode-se não conseguir obter as estimativas do vetor $\boldsymbol{\beta}$ e/ou ter-se um inflacionamento de $Var(\widehat{\beta}_j)$, já que $Var(\widehat{\beta}_j) = \sigma^2\psi_j$ em que ψ_j é o j-ésimo elemento da diagonal principal de $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
- A interpretação dos parâmetros também pode ficar comprometida.

Fontes de multicolinearidade

- Covariáveis partilham a "mesma natureza" e/ou possuem uma relação funcional (peso, altura e IMC ; consumo de combustível no perímetro urbano e consumo de combustível na estrada). Assim, elas tendem a contribuir de modo semelhante no modelo.
- O método de coleta dos dados (p.e. somente carros "populares" compõem a amostra).
- Estudo de um subconjunto específico da população (por exemplo, carros de luxo).
- Especificação equivocada do modelo (por exemplo, a mesma covariável é considerada de modelo semelhante no modelo, mais de uma vez).

Identificação da multicolinearidade

- Perguntar ao pesquisador sobre a natureza das covariáveis e/ou ver a literatura, relativa ao problem em análise.
- Estudar a matriz de correlação entre as covariáveis.
- Estudar a matriz $\boldsymbol{X}'\boldsymbol{X}$.
  - Calcular o fator de inflação da variância (estudar a magnitude de ψ_j).
  - Estudar a magnitude de seus autovalores.

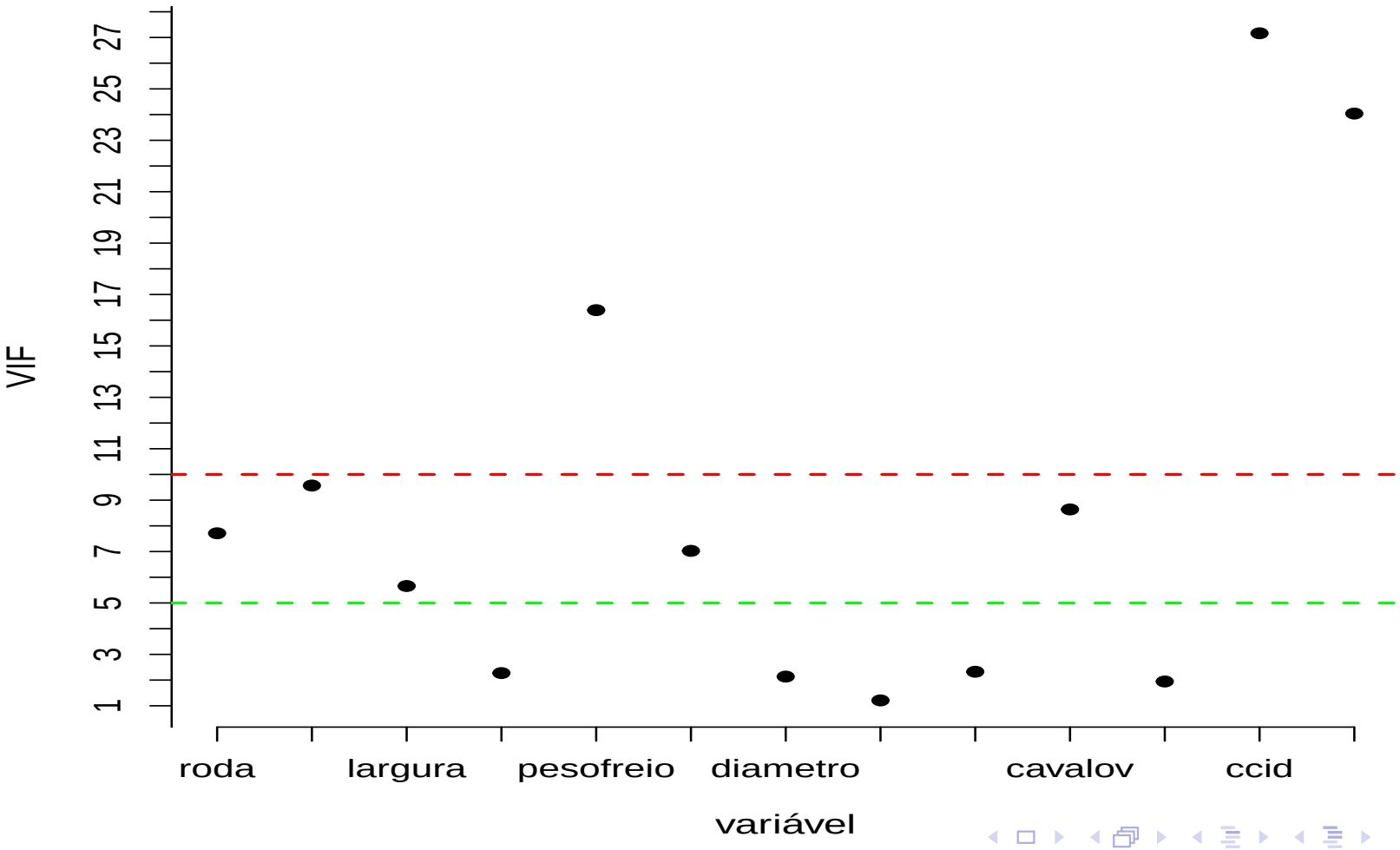
Como tratar a Multicolinearidade

- Eliminar algumas covariáveis do modelo.
- Regressão "ridge".
- Regressão por mínimos quadrados parciais.
- Utillizar funções das covariáveis originais no lugar delas através da:
  - Criação de índices, por exemplo, combinações lineares, que envolvam as covariáveis causadores da multicolinearidade.
  - Redução do número de covariáveis utilizando componentes principais.
- Utilizar decomposições matriciais da matriz $\boldsymbol{X}'\boldsymbol{X}$, por exemplo a decomposição QR ou a LU.
- Mais recentemente: LASSO, LASSO generalizado, spike and slab, spike and slab LASSO. .
- Veja também aqui.

Fator de inflação da variância (VIF: variance inflation factor)

- Pode-se provar (exercício) que (veja slide 18) $\psi_j = (1 - R_j^2)^{-1}$, em que R_j^2 é o coeficiente de determinação ($R^2 = 1 - \frac{SQR}{SQT}$) da regressão da covariável x_j em função das demais covariáveis.
- Assim, quanto menor for a correlação entre x_j e as demais, mais próximo de 1 será o valor de ψ_j.
- Em geral, valores maiores do que 5 ou 10 são uma indicação de que o coeficiente associado à covariável x_j será mal estimado (vício, erro-padrão eqm).
- Veja também aqui.

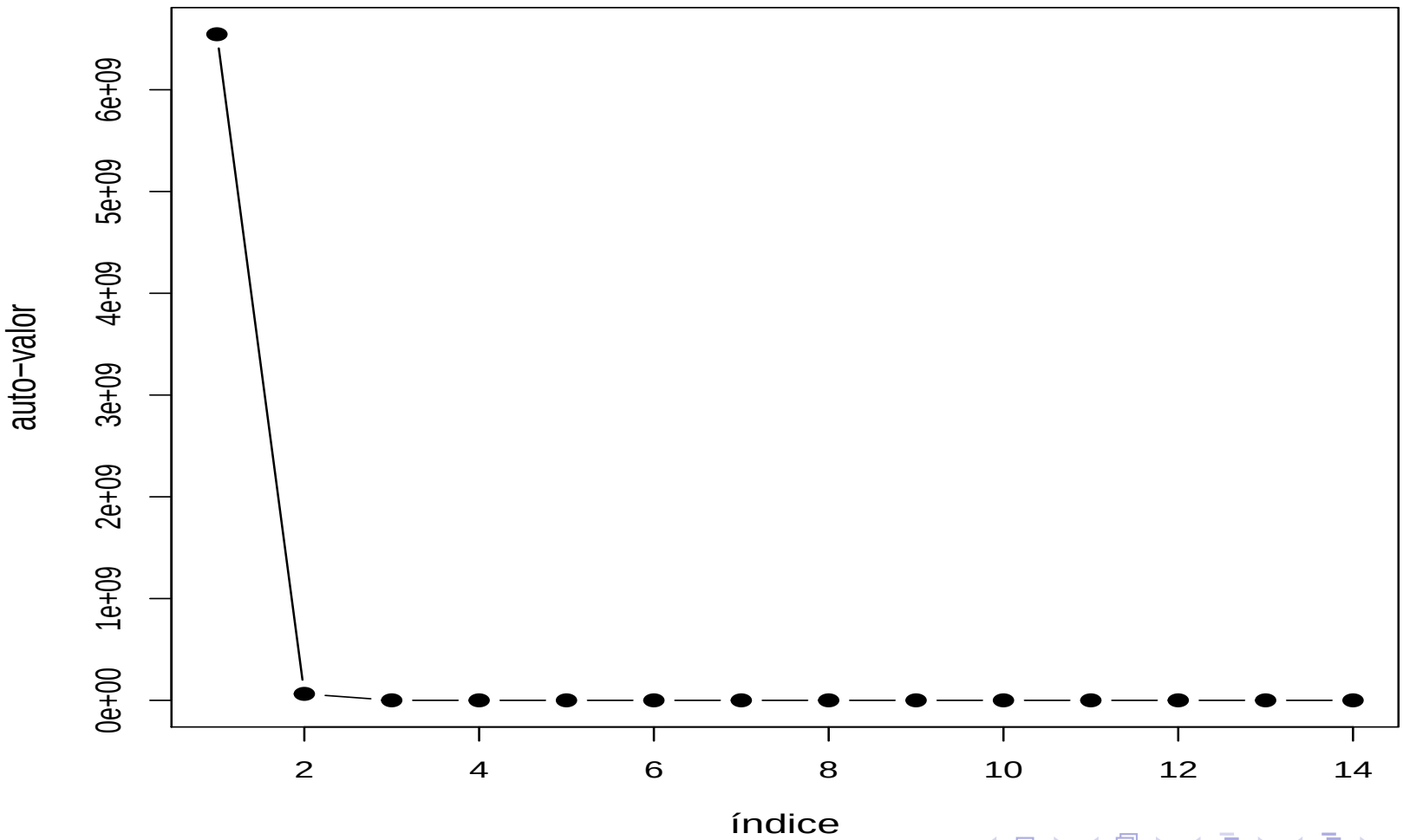
Fator de inflação da variâncias



Autovalores

- Sabe-se (análise multivariada) que quanto mais correlacionadas forem as covariáveis, maior será a magnitude do maior autovalor de $\boldsymbol{X}'\boldsymbol{X}$ em relação à seu menor autovalor.
- Assim, a razão entre o maior autovalor e o menor (o chamado índice de condição) fornece uma idéia sobre a existência de multicolinearidade. Em geral, se tal valor for maior do que 1000, há indícios da existência de multicolinearidade.
- No nosso caso $\kappa = \dfrac{\lambda_{max}}{\lambda_{min}} = 156.230.462.720$. Assim, há indícios da existência de multicolinearidade.

Auto-valores da matriz $X'X$



Regressão ridge

- Consiste em utilizar o estimador $\widehat{\boldsymbol{\beta}}_R = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}'\boldsymbol{Y}$ ao invés do estimador $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Y}$. Exercício, considerando λ conhecido, encontre a distribuição exata de $\widehat{\boldsymbol{\beta}}_R$. Compare os dois estimadores de forma adequada (veja também aqui, slides de 30 a 38).
- O estimador $\widehat{\boldsymbol{\beta}}_R$ é obtido pela minimização de :

$$Q^*(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda (\boldsymbol{\beta}'\boldsymbol{\beta}) = Q(\boldsymbol{\beta}) + g(\lambda, \boldsymbol{\beta})$$

considerando λ conhecido (em relação à $Q(\boldsymbol{\beta})$ veja aqui).

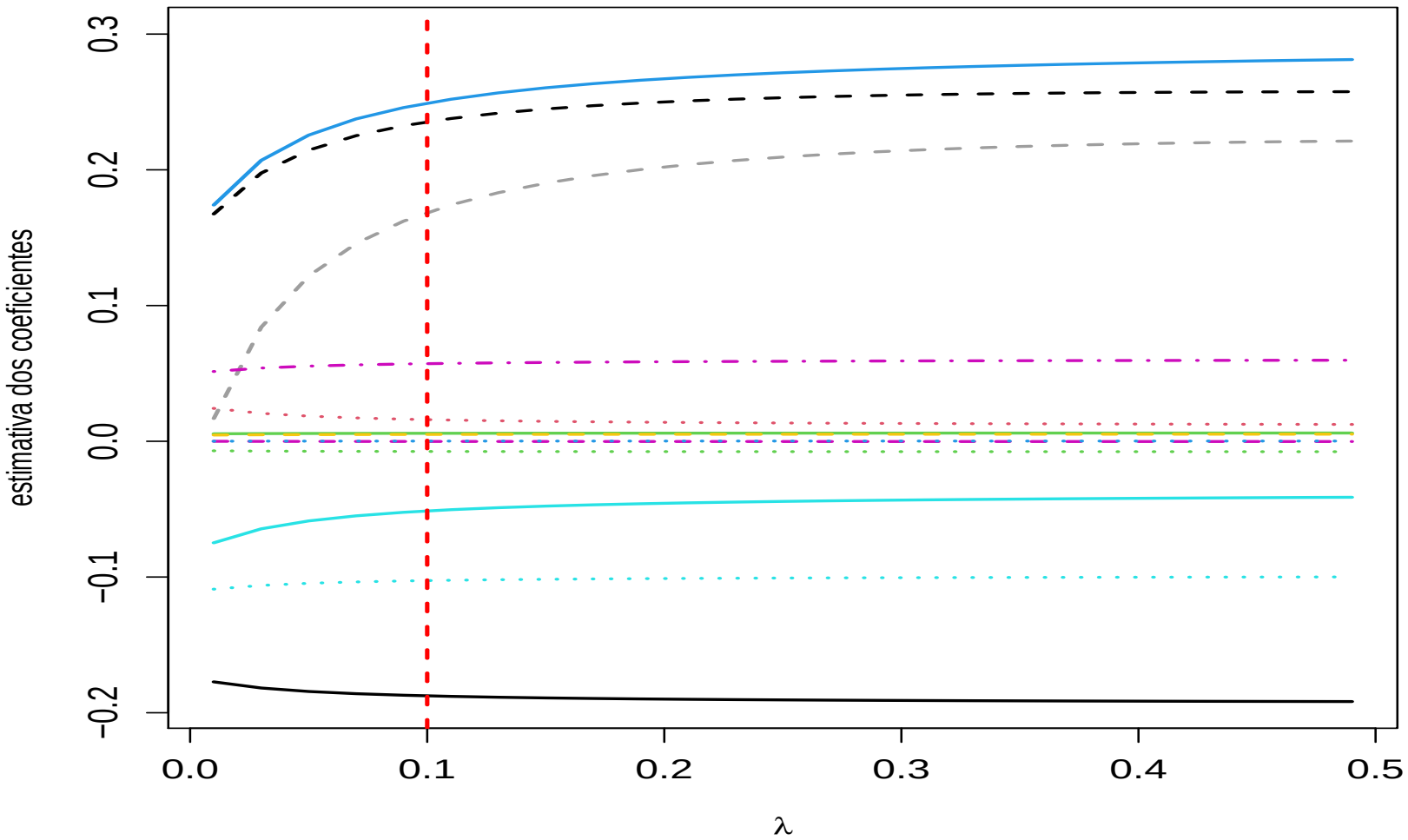
Regressão ridge

- A constante λ pode ser:
  - Escolhida de modo adhoc (segundo algum critério apropriado para modelagem do problema em análise).
  - Escolhida a partir de um conjunto pré-definido (e apropriado) de valores. A escolha pode ser feita através de alguma mecanismo de:
    - Comparação de modelos: critérios de informação, validação cruzada (aqui e aqui), porder de preditivo (raiz quadrade do erro quadrático médio), entre outros.
    - Graficamente: o valor de λ será escolhido (por exemplo) de modo a se obter uma "certa estabilidade" nas estimativas.
  - Estimada em conjunto com os parâmetros $\boldsymbol{\beta}$ (estimação penalizada).

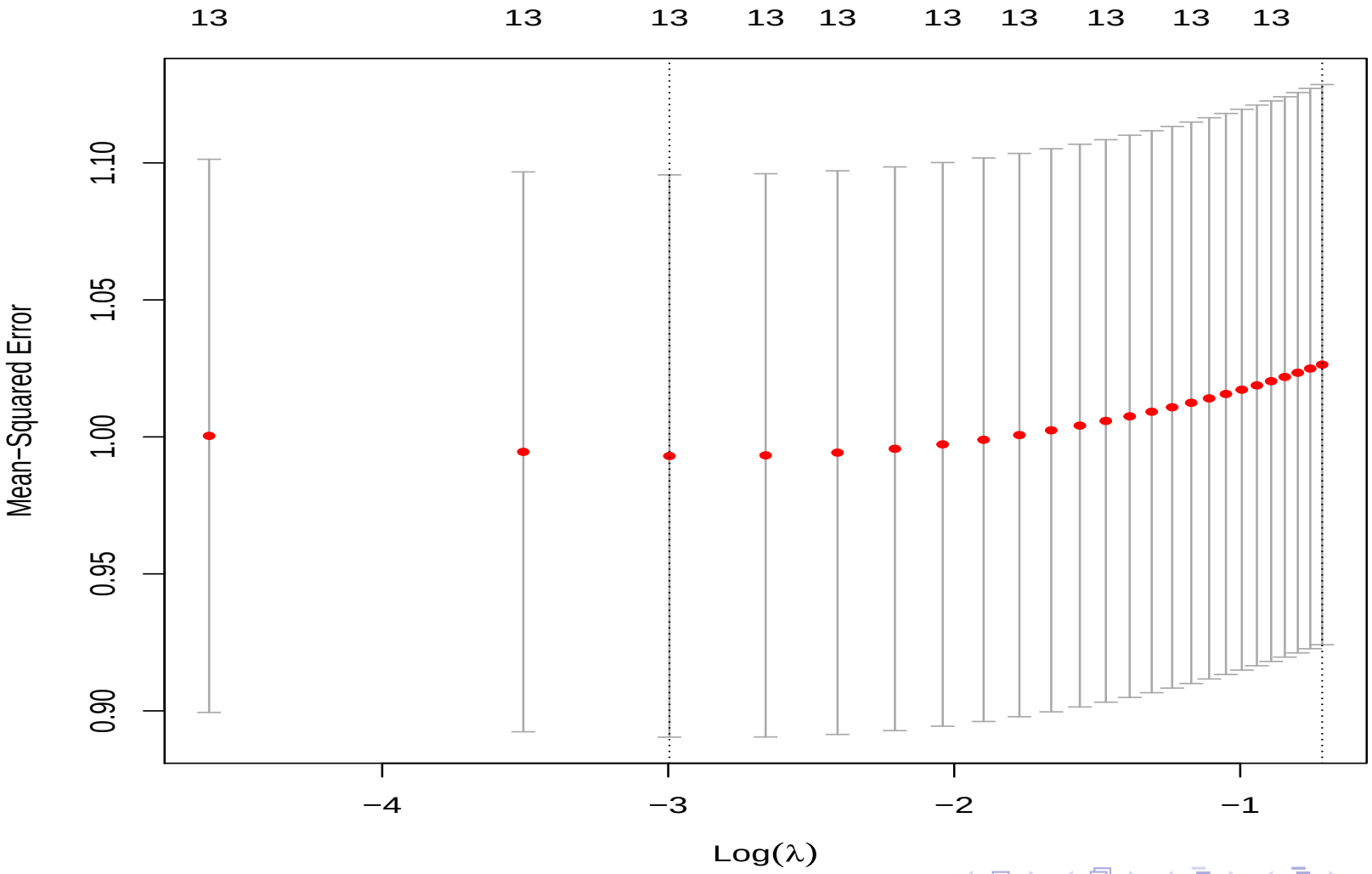
Regressão ridge

- Problemas: não é fácil escolher um λ apropriado. Novas expressões para as variâncias do estimador $\widehat{\boldsymbol{\beta}}_R$ têm de ser obtidas.
- Se λ for estimado (ao invés de escolhido), os resultados vistos, em termos de intervalos de confiança e testes de hipótese, terão de ser adaptados.
- Exercício: pesquisar a respeito.
- O pacote glmnet apresenta algumas formas para a determinação do valor de λ (além de uma série de outras ferramentas ligadas à seleção de covariáveis, estimação penalizada, LASSO, dentre outras, para diferentes modelos de regressão)

Escolha (gráfica) do parâmetro λ



Escolha (validação cruzada) do parâmetro $\lambda = 0,05$



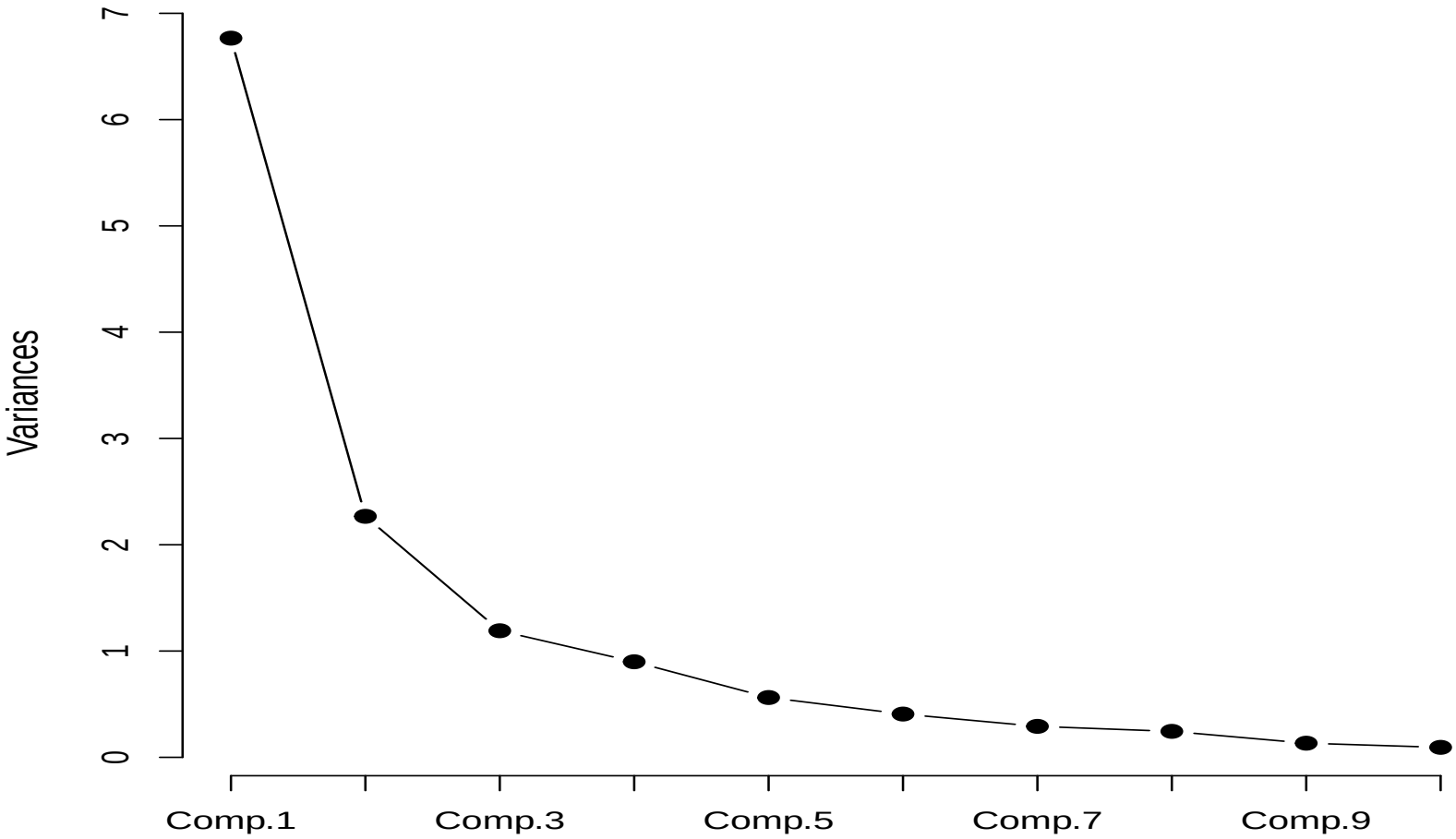
Análise de componentes principais (ACP)

- Essencialmente, consiste em gerar, a partir da matriz de dados de interesse (no caso a matriz de planejamento, $\boldsymbol{X}_{(n\times p)}$) uma combinação linear de variáveis, digamos $\boldsymbol{W}_{(n\times p)}$, de sorte que estas sejam não correlacionadas e retenham a maior parte da estrutura de variabilidade dos dados originais.
- Realizamos uma análise de componentes principais nas covariáveis, utilizando a matriz de correlações.
- O objetivo é utilizar um número menor de componentes principais (que são não correlacionadas) ao invés das covariáveis originais.

Análise de componentes principais (ACP)

- Critério de escolha: percentual acumulado mínimo de variância explicada e interpretabilidade das componentes principais.
- Três componentes contribuem para explicar 78,65% da variabilidade dos dados.
- Para mais detalhes sobre ACP veja aqui.

Gráfico de autovalores (matriz de correlações de $\boldsymbol{X}$)



Coeficientes (correlações entre parênteses) das CP's

Variável	Componente 1	Componente 2	Componente 3
roda	**0,31 (0,80)**	0,29 (0,43)	-0,11 (-0,12)
comp	**0,35 (0,91)**	0,16 (0,23)	-0,11 (-0,12)
largura	**0,34 (0,89)**	0,10 (0,14)	0,08 (0,08)
altura	0,12 (0,31)	**0,40 (0,61)**	-0,47 (-0,51)
pesofreio	**0,37 (0,97)**	0,05 (0,07)	0,07 (0,08)
tmotor	**0,33 (0,87)**	-0,09 (-0,13)	0,25 (0,27)
diametro	**0,28 (0,72)**	-0,02 (-0,03)	-0,15 (-0,16)
stroke	0,06 (0,15)	0,11 (0,17)	**0,74 (0,81)**
compressao	0,01 (0,04)	**0,52 (0,78)**	0,27 (0,30)
cavalov	**0,30 (0,79)**	-0,31 (-0,47)	0,15 (0,17)
prpm	-0,09 (-0,23)	**-0,45 (-0,68)**	-0,05 (-0,05)
ccid	**-0,33 (-0,85)**	0,28 (0,42)	0,09 (0,10)
cest	**-0,34 (-0,88)**	0,22 (0,34)	0,10 (0,10)

Equações de regressão da Componente Principal 1

"Completa"

$$\begin{aligned}\widetilde{CP}_1 &= 0,30 roda_i^* + 0,35 comp_i^* + 0,34 largura_i^* + 0,12 altura_i^* \\ &+ 0,37 pesofreio_i^* + 0,33 tmotor_i^* + 0,28 diametro_i^* + 0,06 stroke_i^* \\ &+ 0,01 compressao_i^* + 0,30 cavalov_i^* - 0,09 prpm_i^* - 0,33 ccid_i^* \\ &- 0,34 cest^*\end{aligned}$$

"Reduzida"

$$\begin{aligned}\widetilde{CP}_1 &= 0,30 roda_i^* + 0,35 comp_i^* + 0,34 largura_i^* \\ &+ 0,37 pesofreio_i^* + 0,33 tmoto_i r^* + 0,28 diametro_i^* \\ &+ 0,01 compressao_i^* + 0,30 cavalov_i^* - 0,33 ccid_i^* - 0,34 cest_i^*\end{aligned}$$

em que $*$ representa a respectiva variável padronizada.

Equações de regressão da Componente Principal 2

"Completa"

$$\begin{aligned}\widetilde{CP}_2 &= 0,29roda_i^* + 0,16comp_i^* + 0,19largura_i^* + 0,40altura_i^* \\ &+ 0,05pesofreio_i^* - 0,09tmoto_ir^* - 0,02diametro_i^* + 0,11stroke_i^* \\ &+ 0,52compressao_i^* - 0,31cavalov_i^* - 0,45prpm_i^* - 0,28ccid_i^* \\ &+ 0,22cest_i^*\end{aligned}$$

"Reduzida"

$$\widetilde{CP}_2 = 0,40altura_i^* + 0,52compressao^* - 0,45prpm^*$$

em que $*$ representa a respectiva variável padronizada.

Equações da Componentes Principal 3

"Completa"

$$\begin{aligned}\widetilde{CP}_3 &= -0,11 roda_i^* - 0,11 comp_i^* + 0,08 largura_i^* - 0,47 altura_i^* \\ &+ 0,07 pesofreio_i^* + 0,25 tmotor_i^* - 0,15 diametro_i^* + 0,74 stroke_i^* \\ &+ 0,27 compressao^* + 0,15 cavalov_i^* - 0,05 prpm_i^* + 0,09 ccid_i^* \\ &+ 0,10 cest_i^*\end{aligned}$$

"Reduzida"

$$\widetilde{CP}_3 = 0,74 stroke_i^*$$

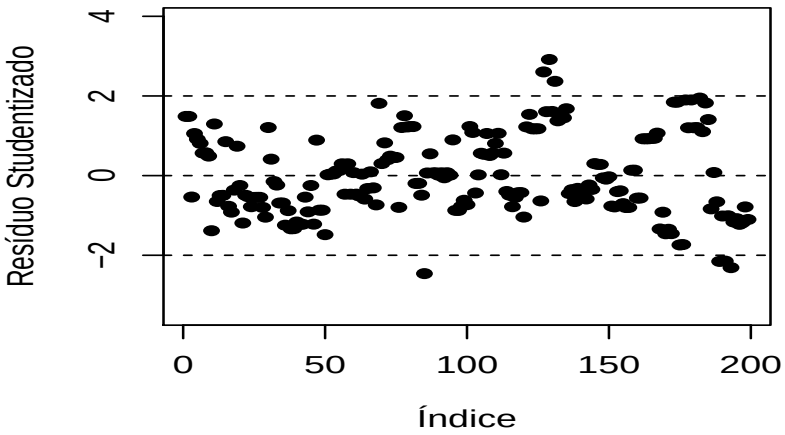
em que $*$ representa a respectiva variável padronizada.

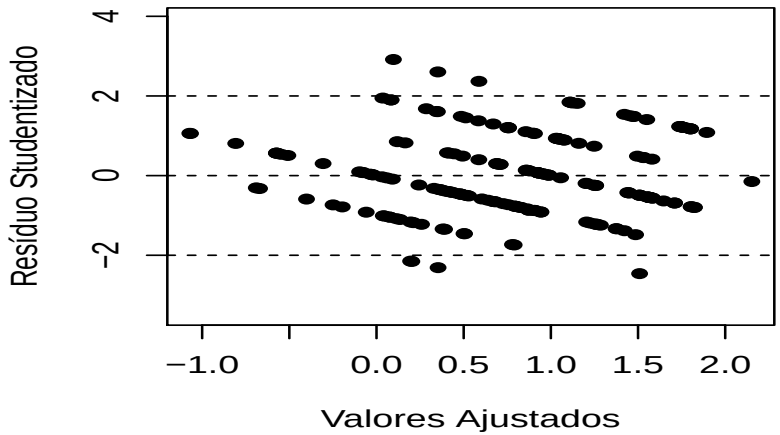
Exemplo 9: continuação

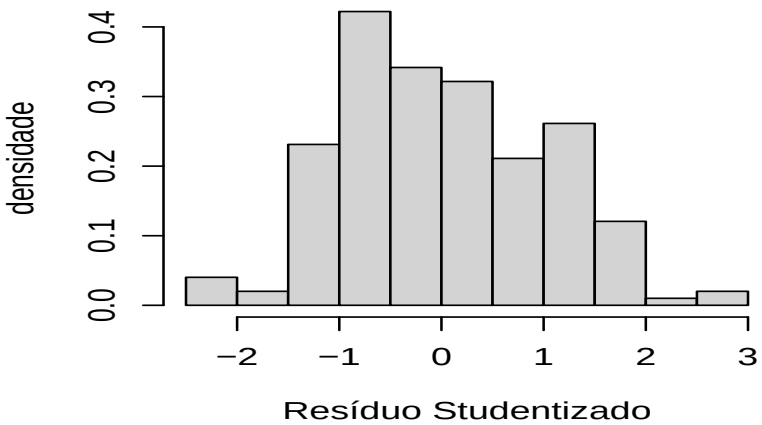
$$Y_i = \beta_0 + \beta_1 comp_{1i} + \beta_2 comp_{2i} + \beta_3 comp_{3i} + \xi_i, i = 1, ..., 199$$

- $\xi \overset{i.i.d}{\sim} N(0, \sigma^2)$.
- β_0 : risco esperado de assegurar carro com valor nulo para todas as componentes.
- β_j : incremento (positivo ou negativo) no risco de assegurar esperado, para o aumento em uma unidade na componente $j, j = 1, 2, 3$.

Gráficos de resíduos







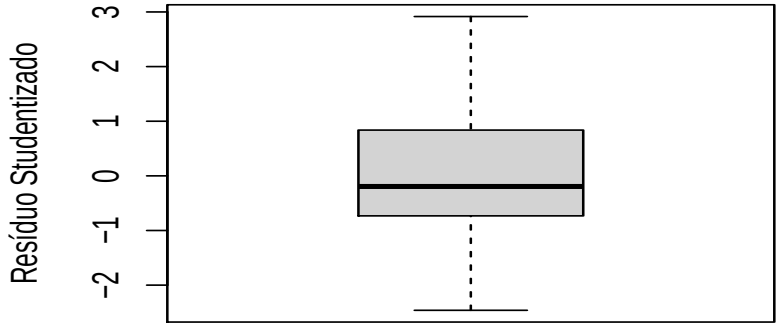
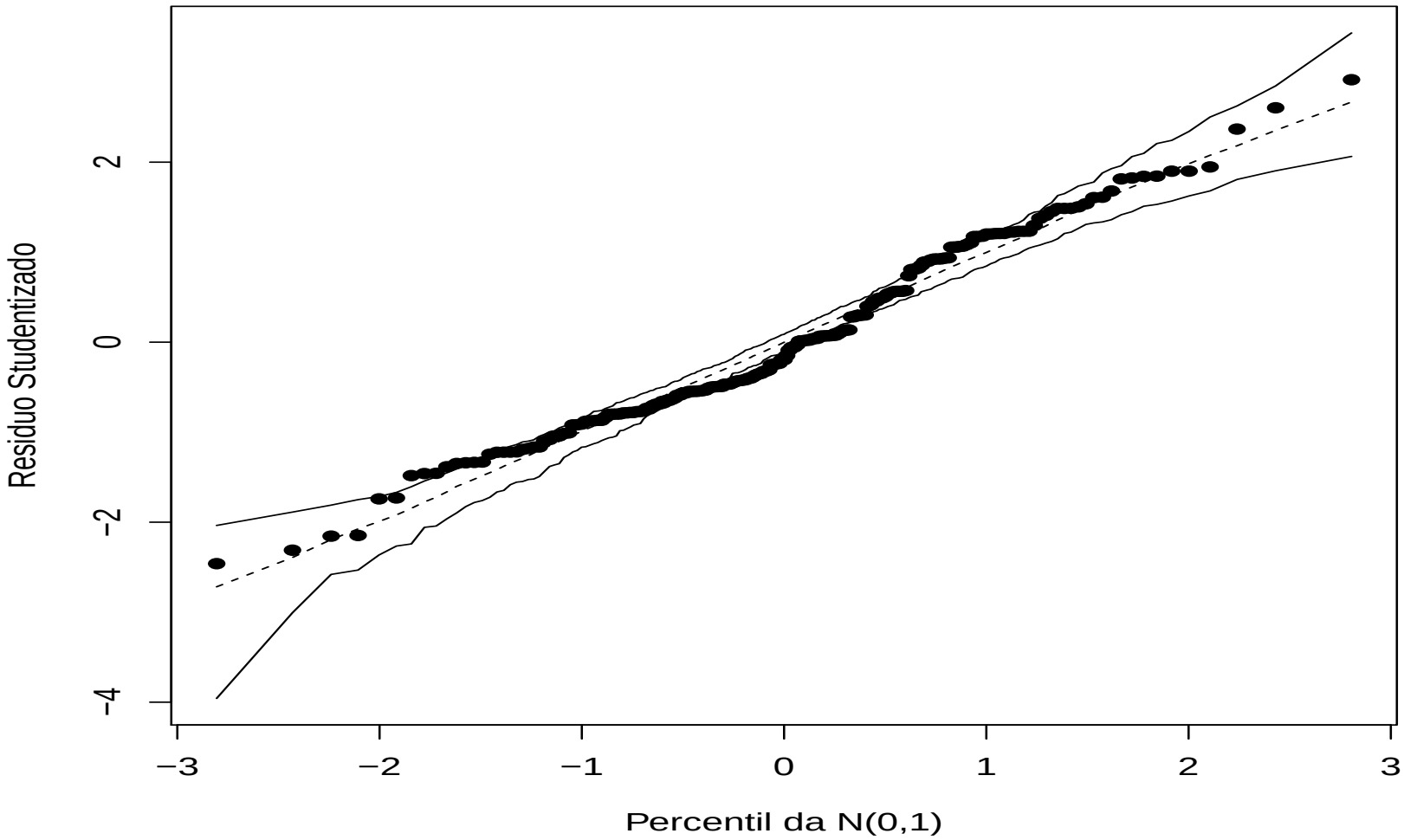


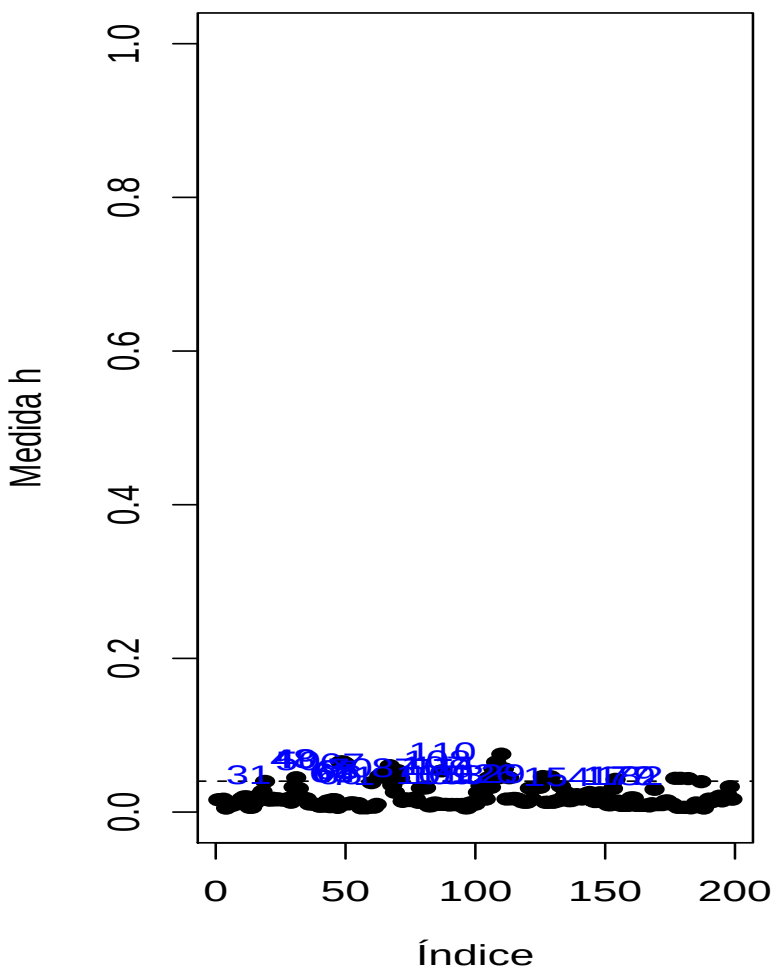
Gráfico de envelopes para os resíduos

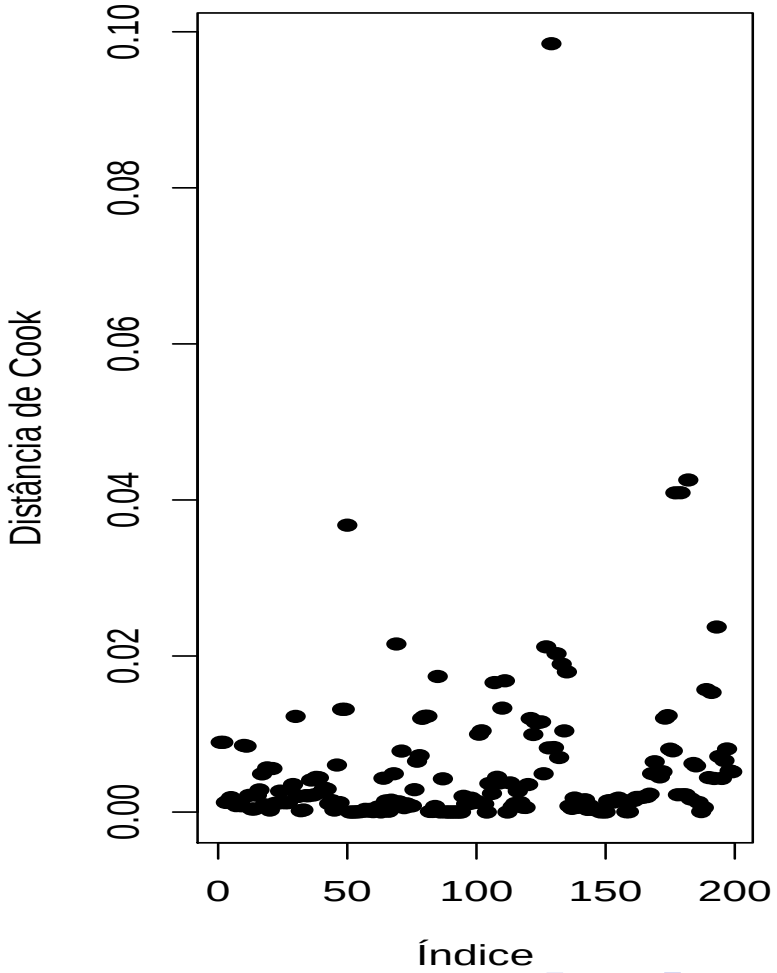


Ajuste do modelo considerando as componentes principais

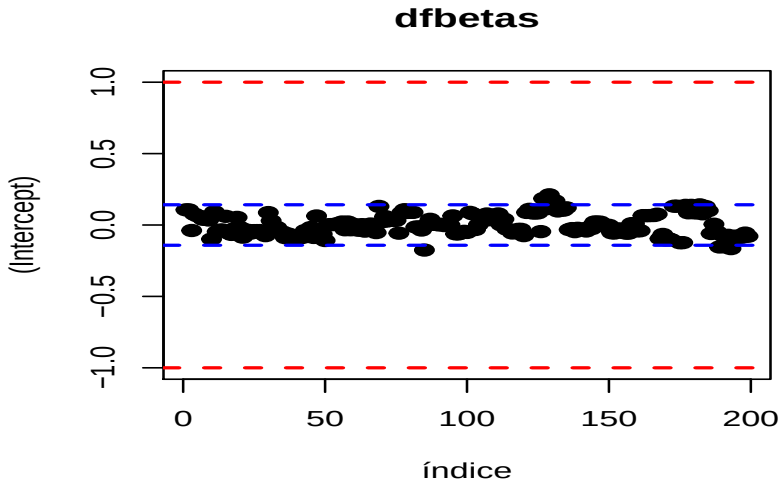
Variável	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	0,789	0,074	[0,644 ; 0,934]	10,715	<0,0001
β_1	0,109	0,028	[0,053 ; 0,165]	3,848	0,0002
β_2	0,341	0,049	[0,244 ; 0,437]	6,972	<0,0001
β_3	-0,263	0,067	[-0,396 ; -0,129]	-3,891	0,0001

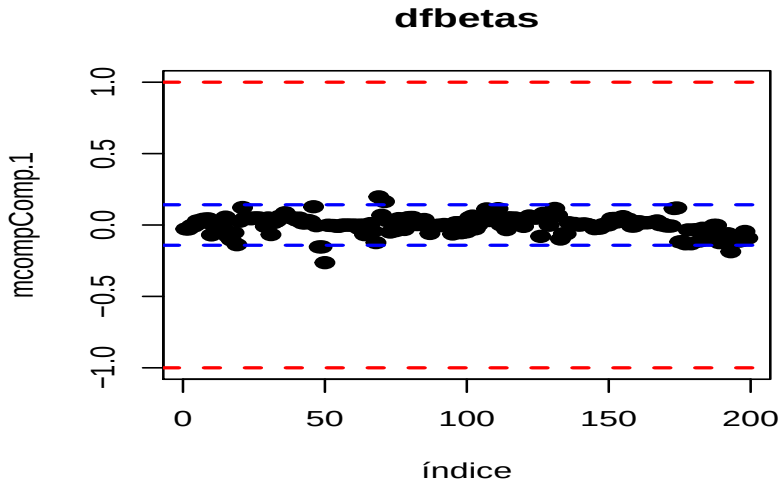
Análise de Influência

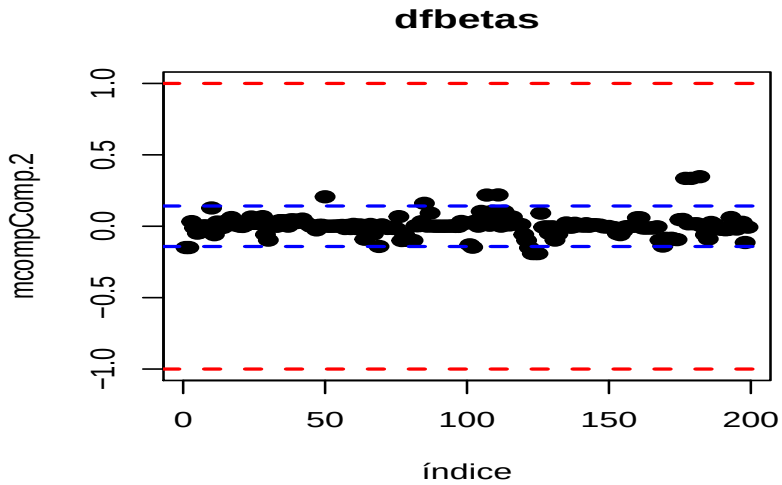


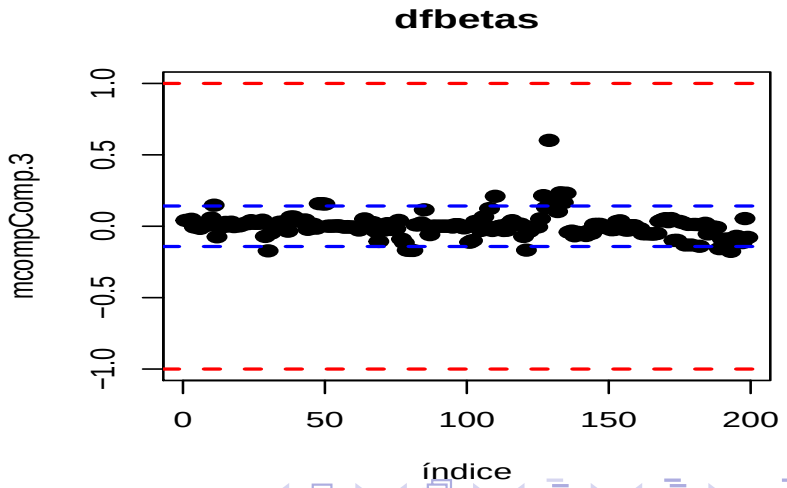


Análise de Influência









LASSO

- Um dos objetivos do algoritmo LASSO é selecionar covariáveis (de forma apropriada) frente à situações "adversas" (e.g., muitas covariáveis, multicolinearidade, ter de considerar significância, também, do ponto de vista do problema dentre outras).
- Consiste em minimizar a seguinte forma quadrática (com relação à $\boldsymbol{\beta}$):

$$Q^{**}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| = Q(\boldsymbol{\beta}) + g(\lambda, \boldsymbol{\beta})$$

LASSO

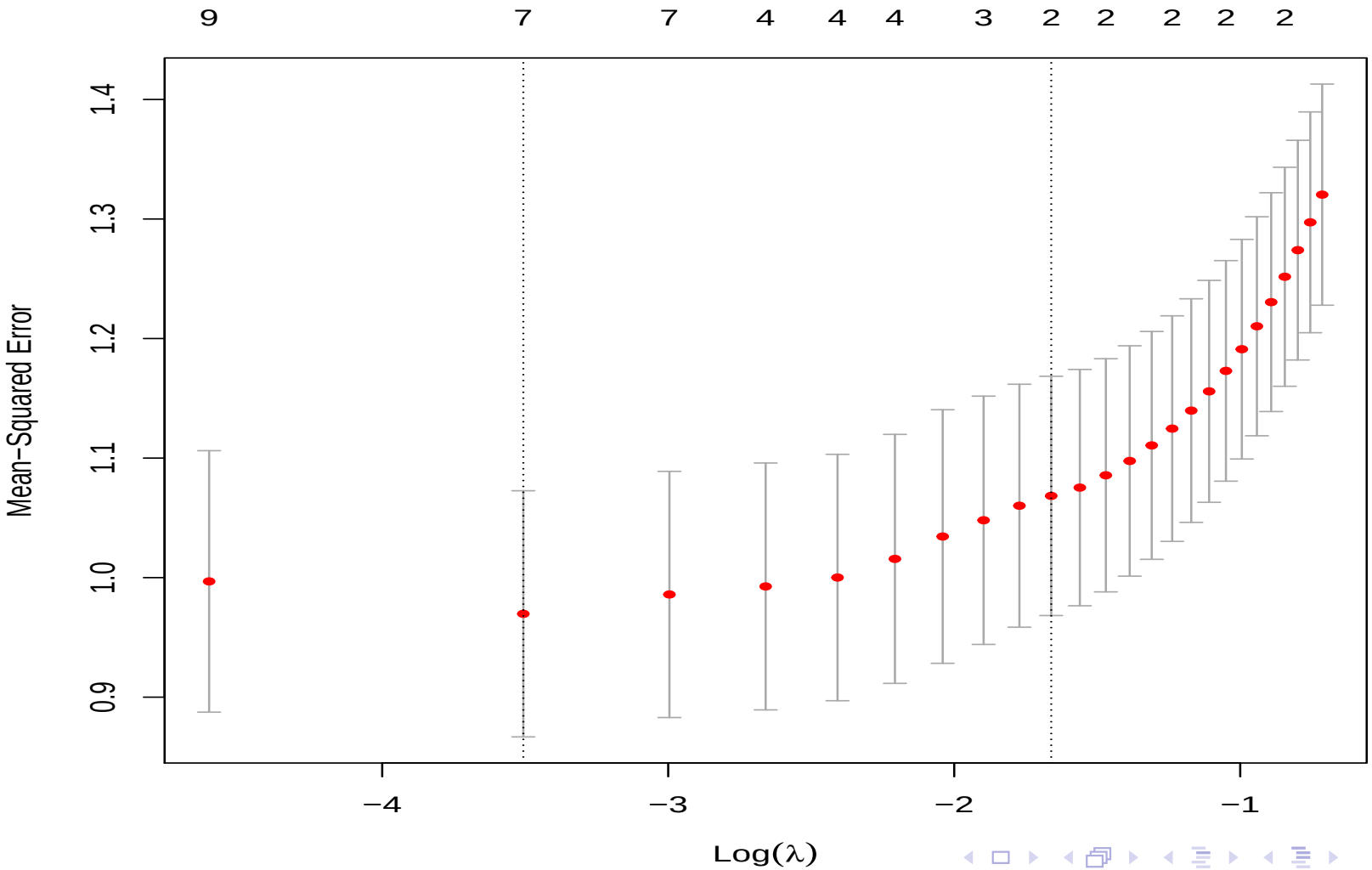
- Equivalentemente, minimizar

$$Q^{**}(\beta) = (\boldsymbol{Y} - \boldsymbol{X}\beta)' (\boldsymbol{Y} - \boldsymbol{X}\beta)$$

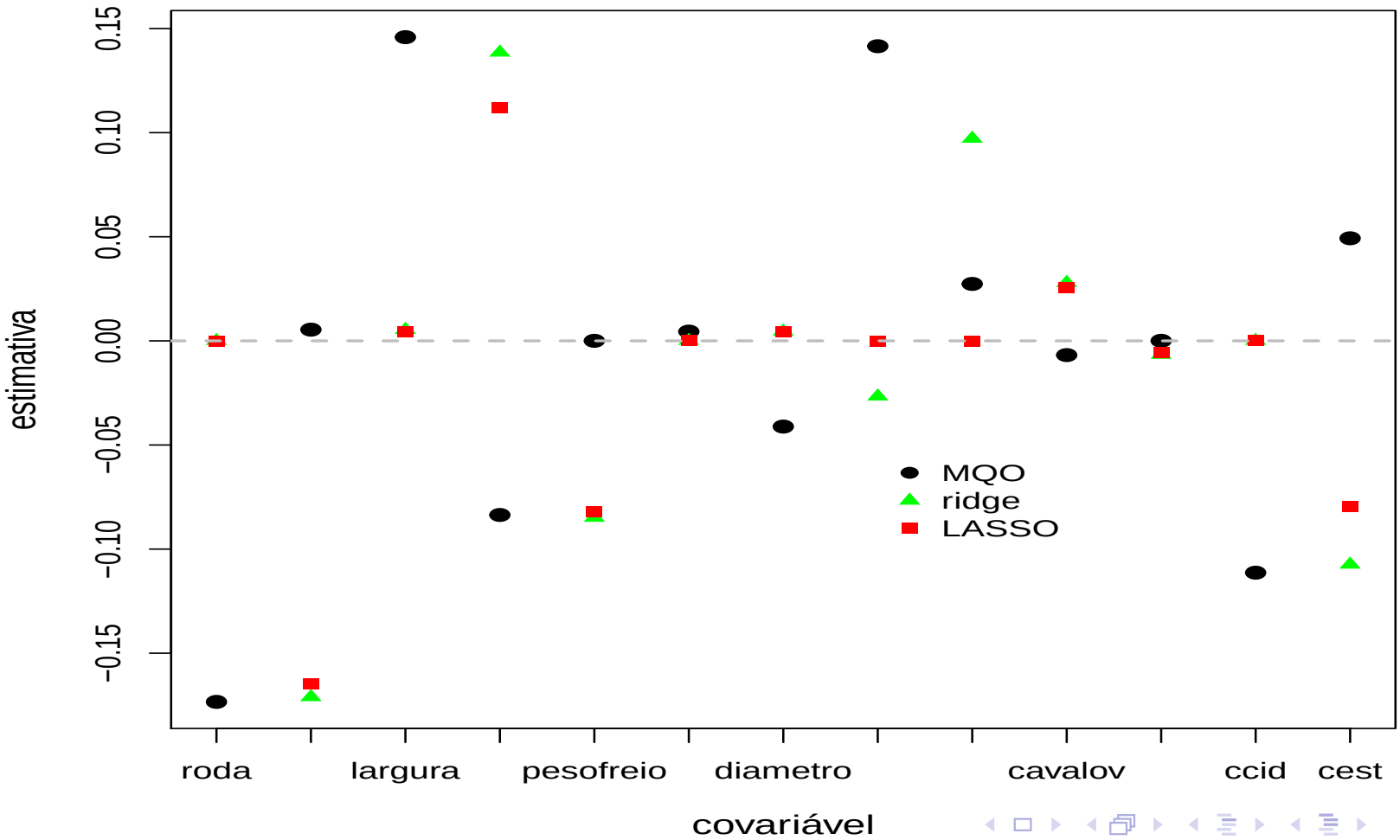
  sujeito a $\sum_{j=1}^{p} |\beta_j| \leq t, t > 0$

- É um problema não trivial que requer o uso de métodos de otimização apropriados. Veja as referências mencionadas anteriormente

Escolha (validação cruzada) do parâmetro $\lambda = 0,03$



Comparação das estimativas de β (menos o intercepto)



Comentários

- A multicolinariedade pode afetar o processo de inferência bem como as conclusões a serem depreendidas.
- Deve ser devidamente tratada, desde um ponto de vista do problema a ser modelado, bem como de um ponto de vista estatístico.
- A melhor forma depende dos objetivos, das característcas dos dados e dos recursos (temporais e computacionais) disponíveis.
- Tutoriais sobre o pacote glmnet: aqui, aqui e aqui.