

# Multilinearidade

Prof. Caio Azevedo

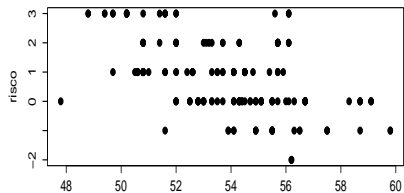
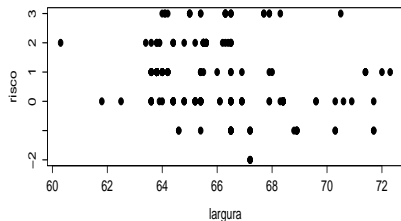
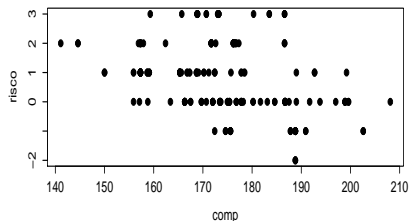
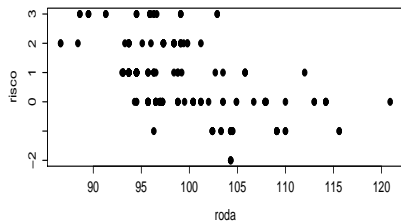
## Exemplo 9: risco de assegurar automóveis

- O conjunto de dados em questão diz respeito ao estudo do risco (para a seguradora) de assegurar determinado veículo, o qual varia no intervalo  $\{-3, -2, -1, 0, 1, 2, 3\}$ , em função de diversas características do veículo. Quanto maior o valor, maior o risco.
- Características de interesse: 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).

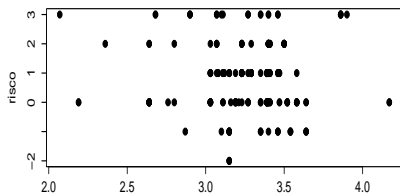
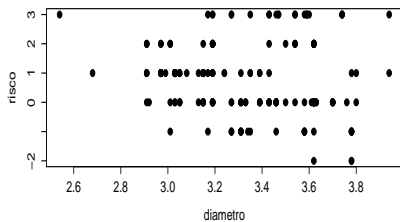
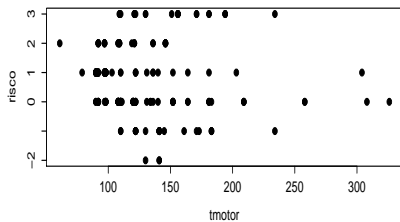
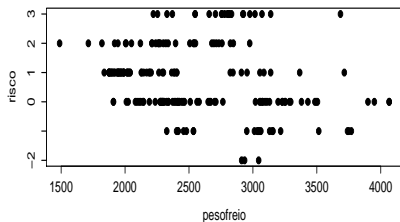
# Introdução

- 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 podemos 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?

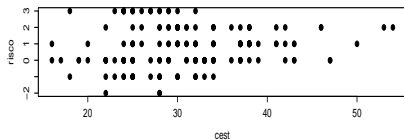
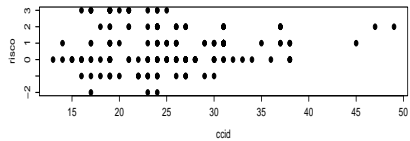
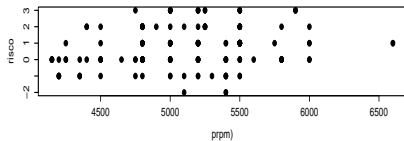
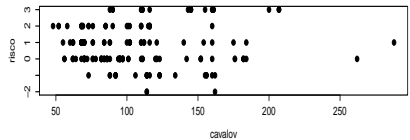
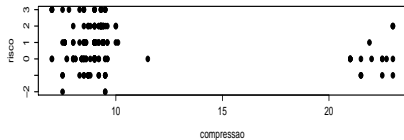
# Gráficos de dispersão



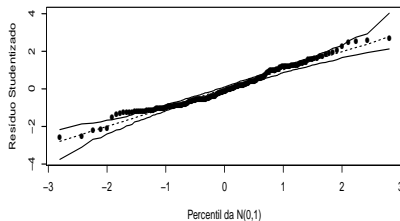
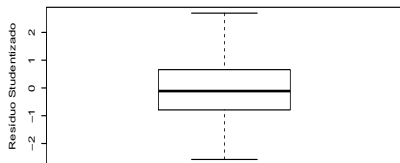
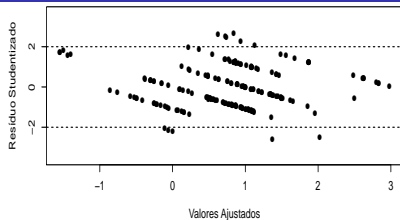
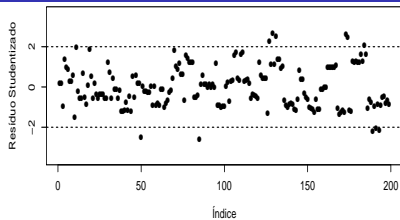
# Gráficos de dispersão



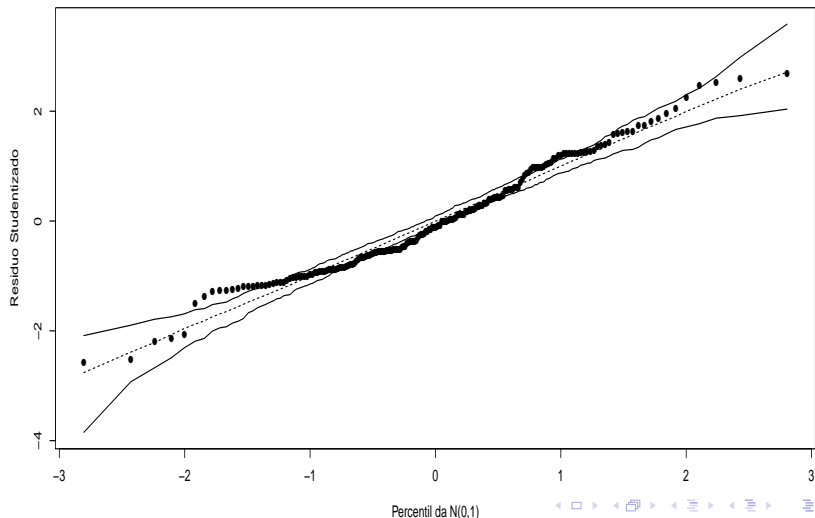
# Gráficos de dispersão



# Gráficos de resíduos



# Gráfico de envelopes para os resíduos

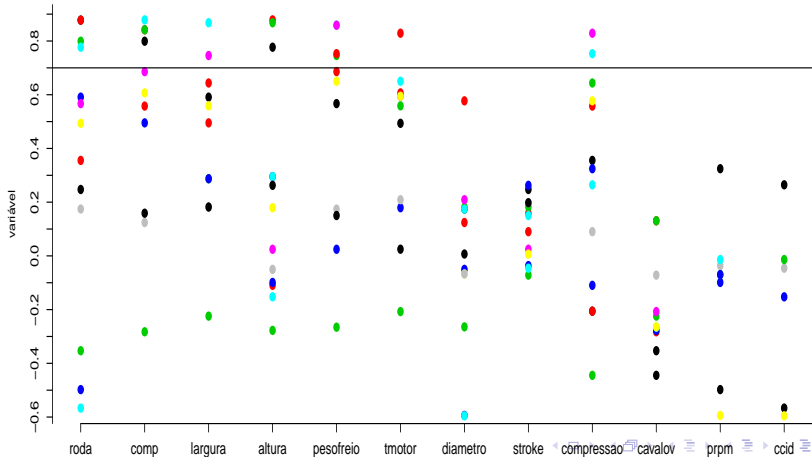




## Ajuste do modelo (MRNL) com as variáveis originais

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

# Correlação entre as covariáveis



# Problemas provocados pela multicolinearidade

- Se houver alguma estrutura de correlação 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  $\mathbf{X}'\mathbf{X}$  pode ser próximo de zero.
- A invertibilidade de  $\mathbf{X}'\mathbf{X}$  fica, assim, comprometida, e pode-se não conseguir obter as estimativas do vetor  $\beta$  ou ter-se um inflacionamento de  $Var(\hat{\beta}_j)$ , já que  $Var(\hat{\beta}_j) = \sigma^2\psi_j$  em que  $\psi_j$  é o  $j$ -ésimo elemento da diagonal principal de  $(\mathbf{X}'\mathbf{X})^{-1}$ .
- 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 contribuem de modo semelhante no modelo.
- O método de coleta dos dados (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).

# Identificação da multicolinearidade

- Perguntar ao pesquisador sobre a natureza das covariáveis e/ou ver a literatura a respeito.
- Estudar a matriz de correlação entre as covariáveis.
- Estudar a matriz  $\mathbf{X}'\mathbf{X}$ .
  - Calcular o fator de inflação da variância (estudar a magnitude de  $\psi_j$ ).
  - Estudar a magnitude de seus autovalores.

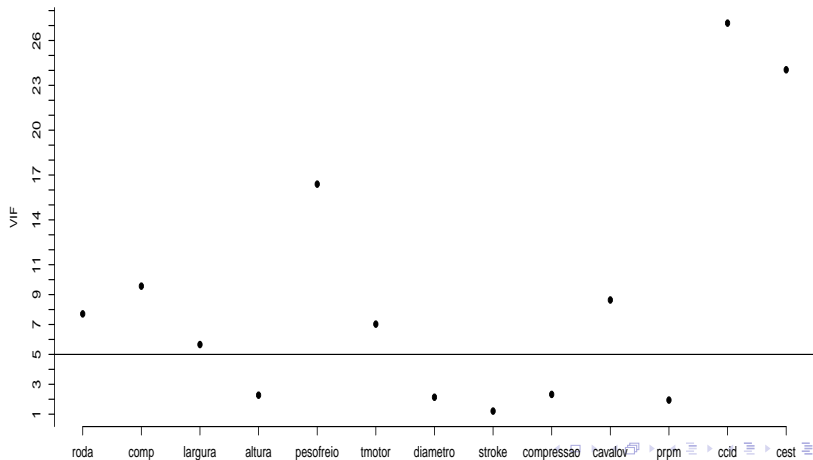
# Como tratar a Multicolinearidade

- Eliminar algumas covariáveis do modelo.
- Regressão “ridge”.
- Utilizar 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.

## Fator de inflação da variância

- Pode-se provar que  $\psi_j = (1 - R_j^2)^{-1}$ , 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  $\psi_j$ .
- Em geral, valores maiores do que 5 ou 10 são uma indicação de que o coeficiente associado à covariável  $x_j$  será mal estimado.

# 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  $\mathbf{X}'\mathbf{X}$  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 multicolineariedade. Em geral, se tal valor for maior do que 1000, há indícios da existência de multicolineariedade.
- No nosso caso  $\kappa = \frac{\lambda_{max}}{\lambda_{min}} = 156230462720$ . Assim, há indícios da existência de multicolineariedade.

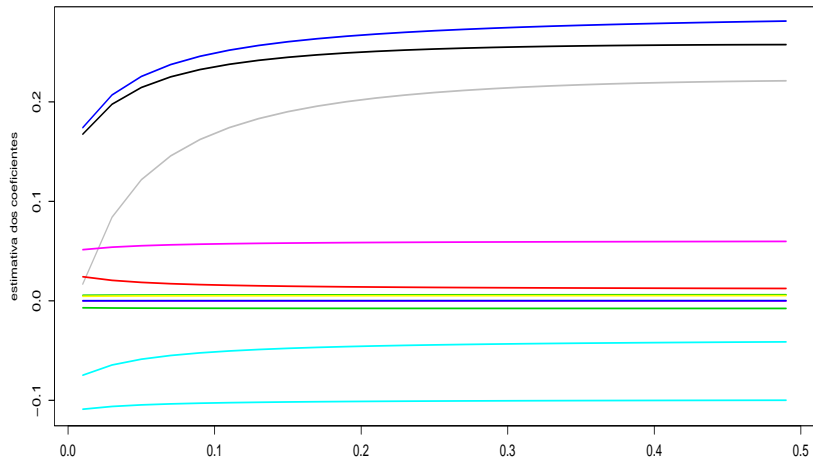
# Regressão ridge

- Consiste em utilizar o estimador  $\hat{\beta}_R = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}'\mathbf{Y}$  ao invés do estimador usar  $\hat{\beta}_R = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$ .
- A constante  $\lambda$  pode ser:
  - Escolhida de modo adhoc.
  - Escolhida a partir de um conjunto pré-definido de valores (a escolha pode ser feita através de alguma mecanismo de comparação de modelos ou graficamente). Em sendo a escolha feita de modo gráfico, o valor  $\lambda$  será escolhido até se obter uma estabilidade nas estimativas.
  - Estimada em conjunto com os parâmetros  $\beta$ .

# Regressão ridge

- Problemas: não é fácil escolher um  $k$  apropriado. Novas expressões para as variâncias do estimador  $\hat{\beta}_R$  têm de ser obtidas. Se  $\lambda$  for estimado (ao invés de escolhido), os resultados, em termos de intervalos de confiança e testes de hipótese, terão de ser adaptados do que foi visto para o estimador  $\hat{\beta}$ .

# Escolha do parâmetro $\lambda$



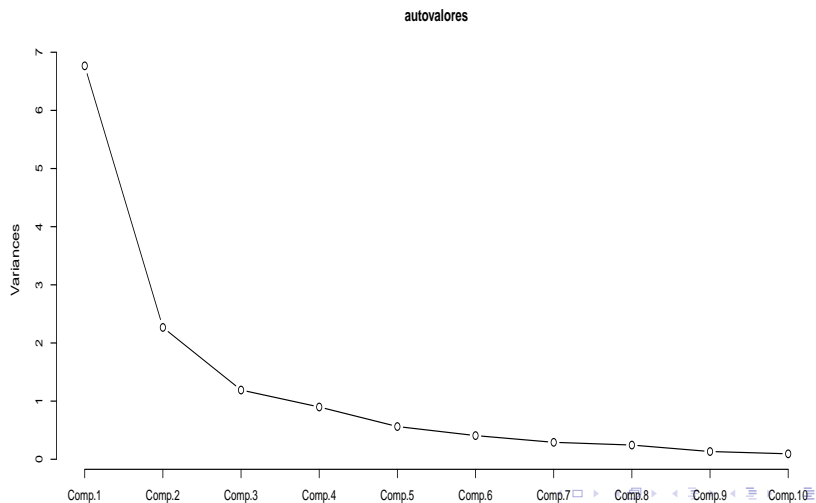
$\lambda$



# Análise de componentes principais

- 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.
- 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.

# Gráfico de autovalores



## Coeficientes (correlações entre parênteses) das CP's

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

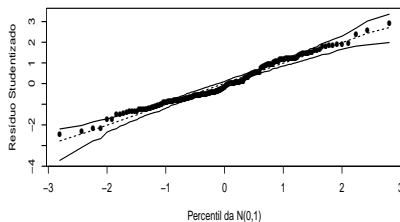
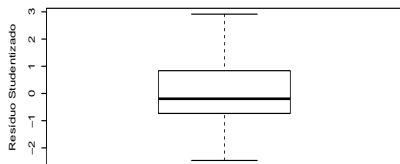
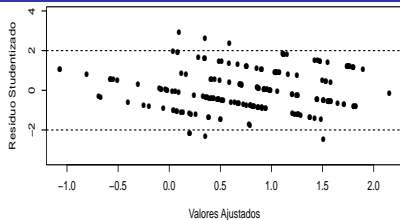
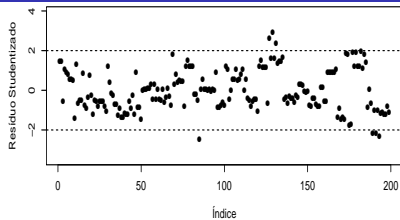
## Exemplo 1: considerando as etiologias cardíacas

$$Y_i = \beta_0 + \beta_1 \text{comp}_{1i} + \beta_2 \text{comp}_{2i} + \beta_3 \text{comp}_{3i} + \xi_i, i = 1, \dots, 199$$

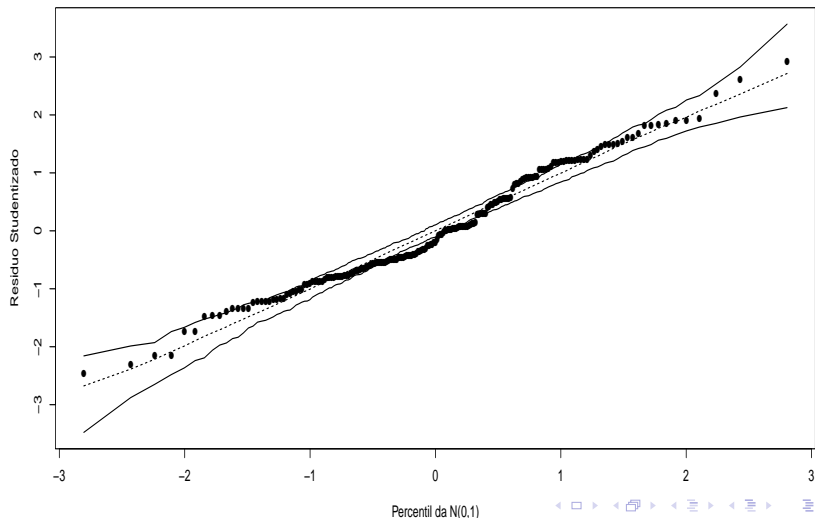
- $\xi \stackrel{i.i.d}{\sim} N(0, \sigma^2)$ .
- $\beta_0$  : risco esperado de assegurar carro com valor nulo para todas as componentes.
- $\beta_j$  : incremento (positivo ou negativo) no risco de assegurar esperado, para o aumento em uma unidade na componente  $j, j = 1, 2, 3$ .



# 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
$\beta_0$	0,789	0,074	[0,644 ; 0,934]	10,715	<0,0000
$\beta_1$	0,109	0,028	[0,053 ; 0,165]	3,848	0,0002
$\beta_2$	0,341	0,049	[0,244 ; 0,437]	6,972	<0,0000
$\beta_3$	-0,263	0,067	[-0,396 ; -0,129]	-3,891	0,0001