Exemplo Multicolinearidade

Gilberto A. Paula

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

1º Semestre 2023

Sumário

Descrição dos Dados

Como ilustração para o tópico de multicolinearidade será analisado um conjunto de dados em que o calor (em calorias por grama) de $n = 13$ amostras de cimento é relacionado com as seguintes variáveis explicativas referentes a ingredientes usados na mistura do cimento:

- X_1: aluminato tricálcico
- X_2: silicato tricálcico
- X_3: aluminato-ferrita tetracálcico
- X_4: silicato dicálcico.

(Montgomery, Peck e Vining, 2021).

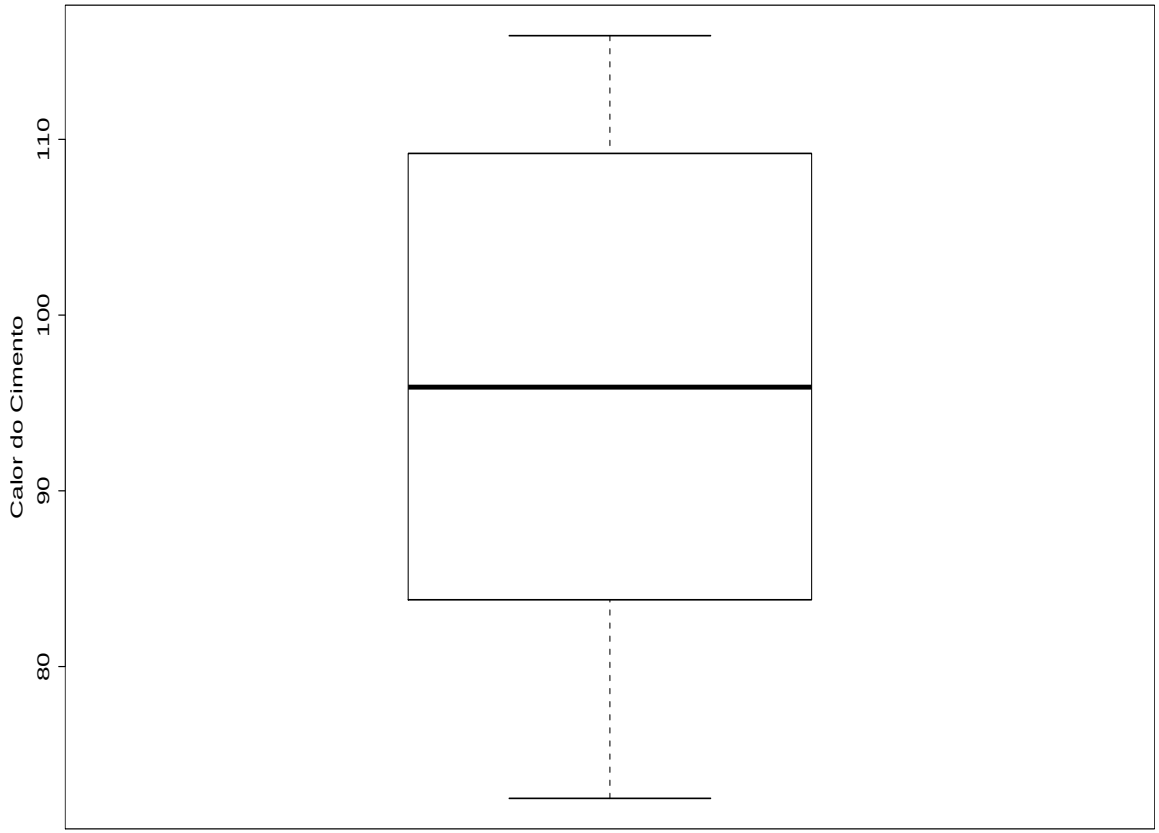
Sumário

Matriz de Correlações Amostrais

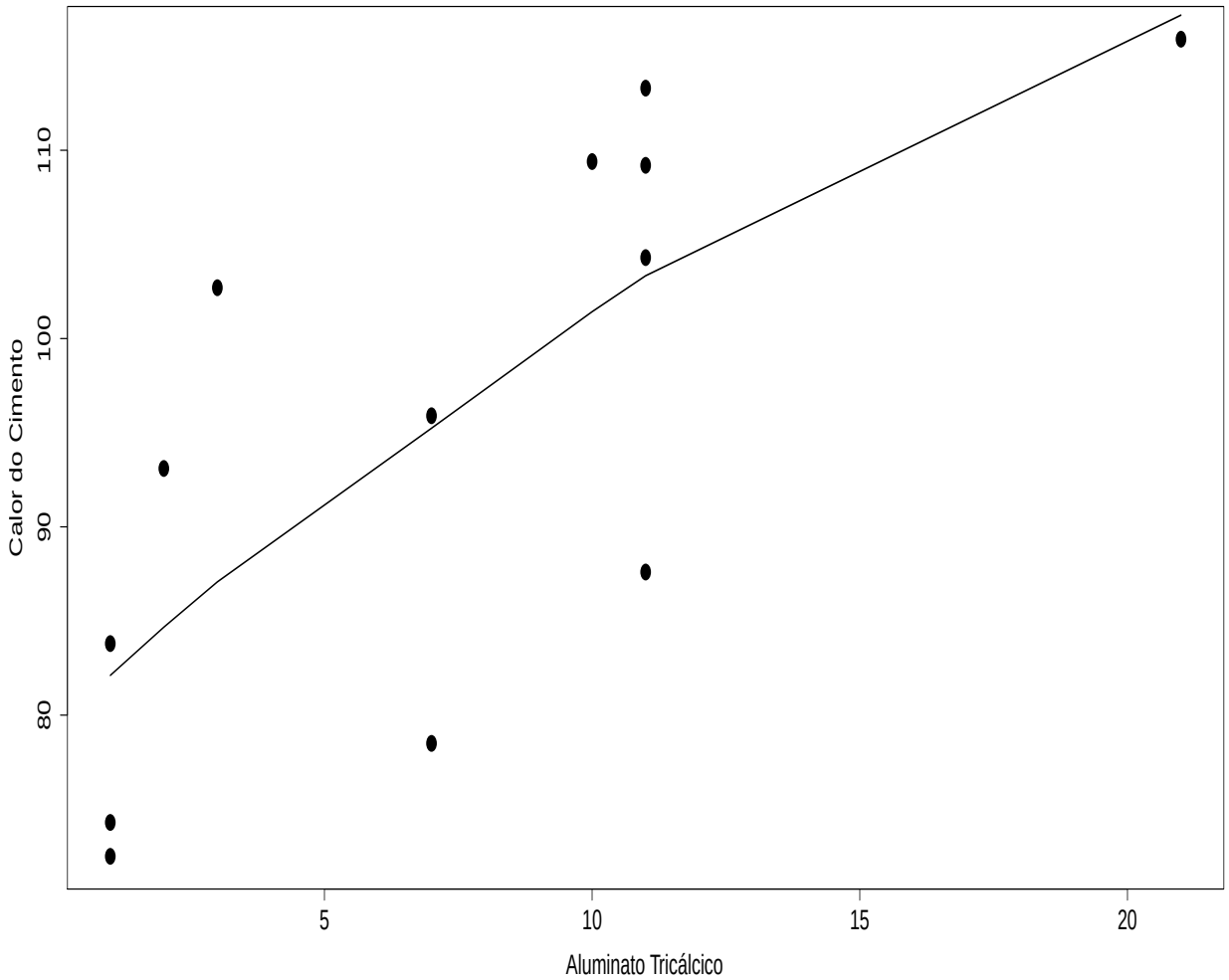
	Calor	X_1	X_2	X_3	X_4
Calor	1,0	0,731	0,816	-0,535	-0,821
X_1		1,0	0,229	-0,824	-0,245
X_2			1,0	-0,139	-0,973
X_3				1,0	0,029
X_4					1,0

Nota-se altas correlações amostrais entre as variáveis explicativas X_1 e X_3 e entre X_2 e X_4.

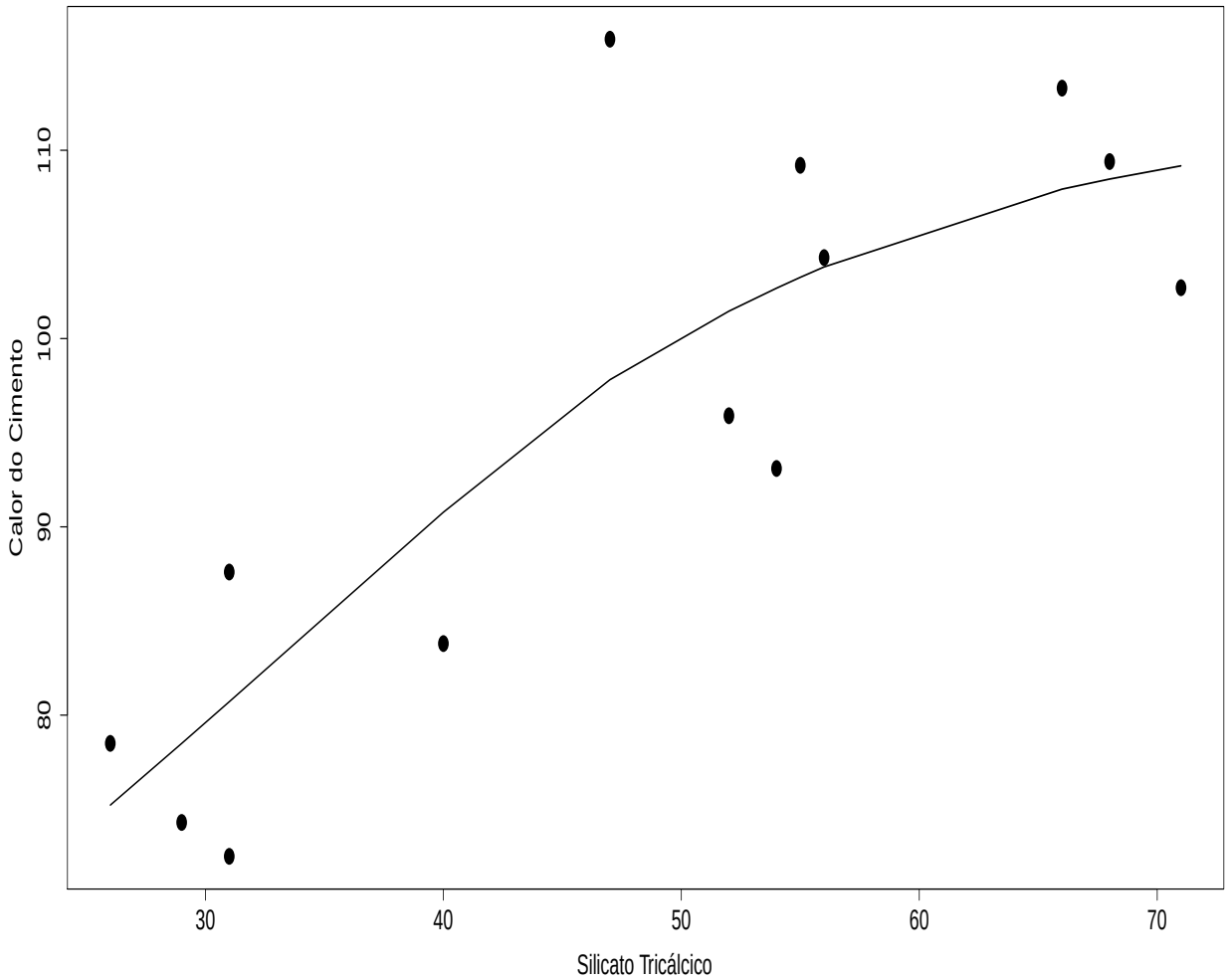
Boxplot Calor do Cimento



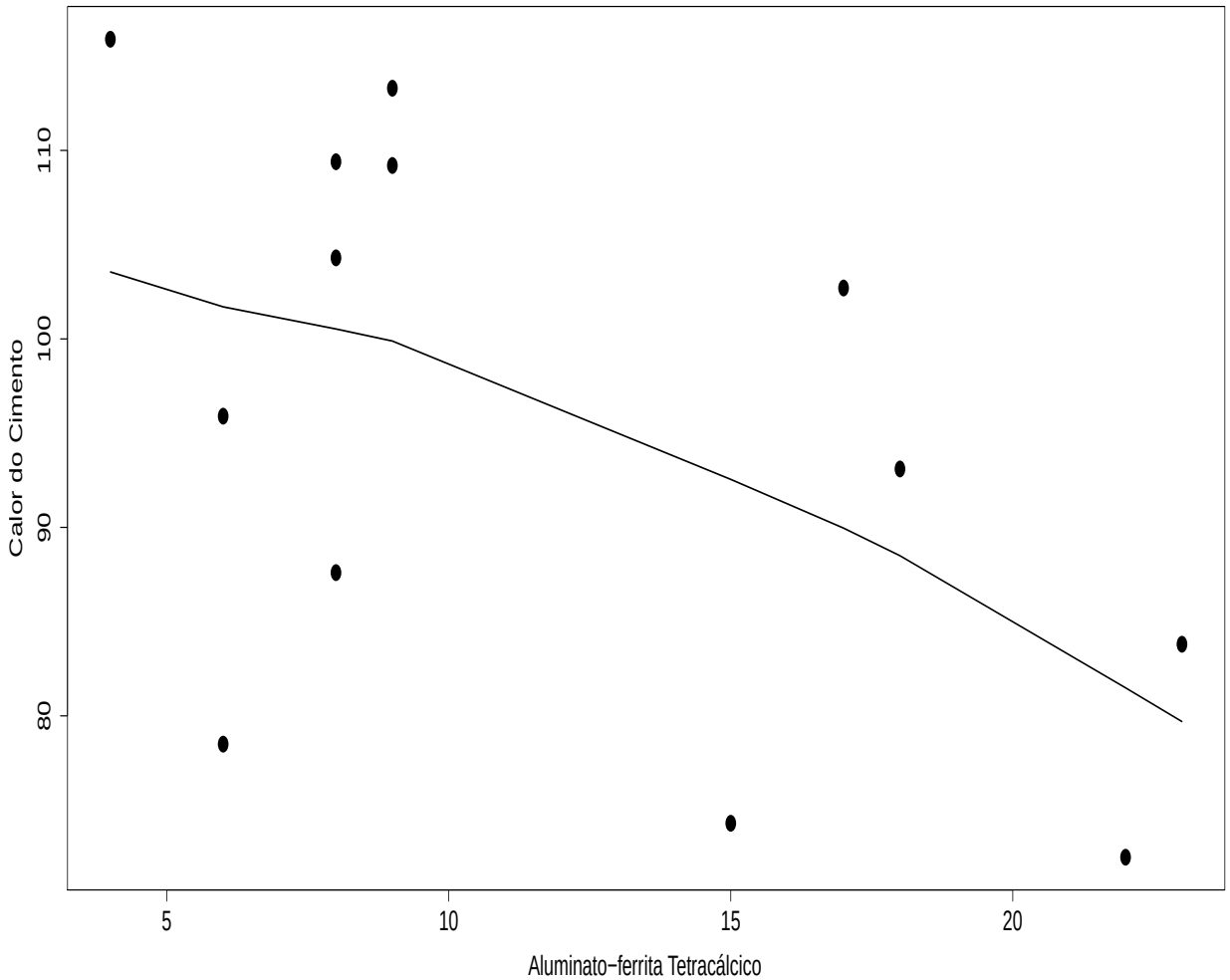
Dispersão Calor versus Aluminato Tricálcico



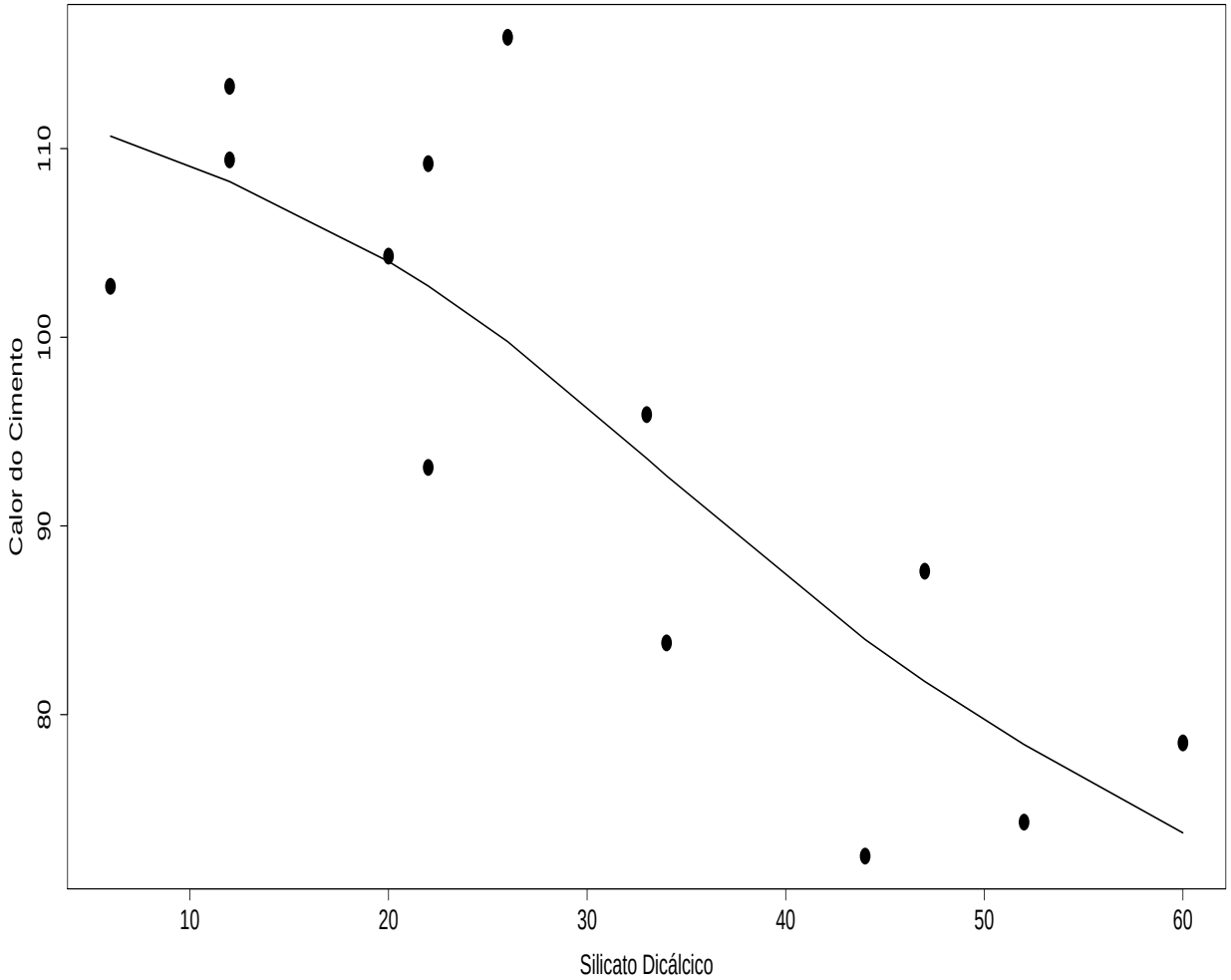
Dispersão Calor versus Silicato Tricálcico



Dispersão Calor versus Aluminato-ferrita Tetracálcico



Dispersão Calor versus Silicato Dicálcico



Comentários

- Nota-se que à medida que aumenta o calor do cimento aumentam o aluminato tricálcico e o silicato tricálcico.
- Por outro lado, à medida que aumenta o calor do cimento diminui o aluminato-ferrita tetracálcico e o silicato dicálcico.

Sumário

Regressão Linear Múltipla

Com base nos diagramas de dispersão o seguinte modelo é proposto:

$$cy_i = \beta_1 cx_{i1} + \beta_2 cx_{i2} + \beta_3 cx_{i3} + \beta_4 cx_{i4} + \epsilon_i,$$

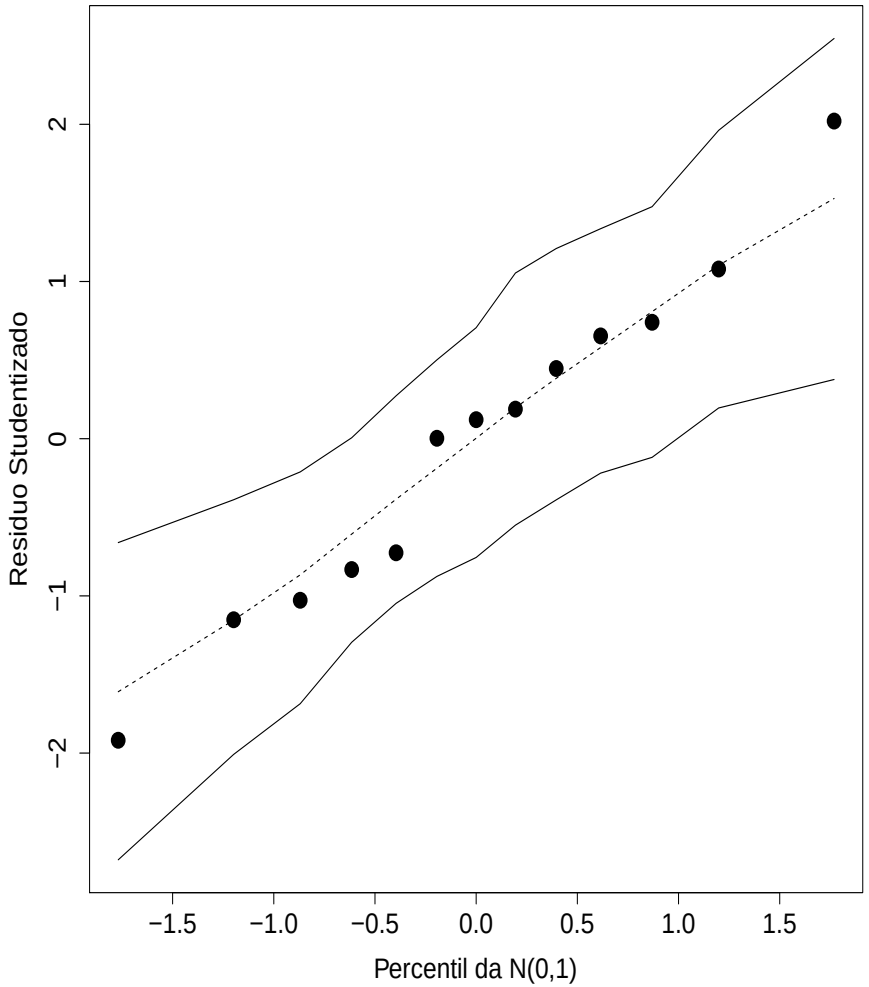
em que cy_i denota o calor da i-ésima amostra de cimento centralizada (subtraído da média amostral), bem como os valores das variáveis explicativas e $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, para $i = 1, \ldots, 13$. Dessa forma, não é necessário incluir o intercepto.

Descrição das Estimativas

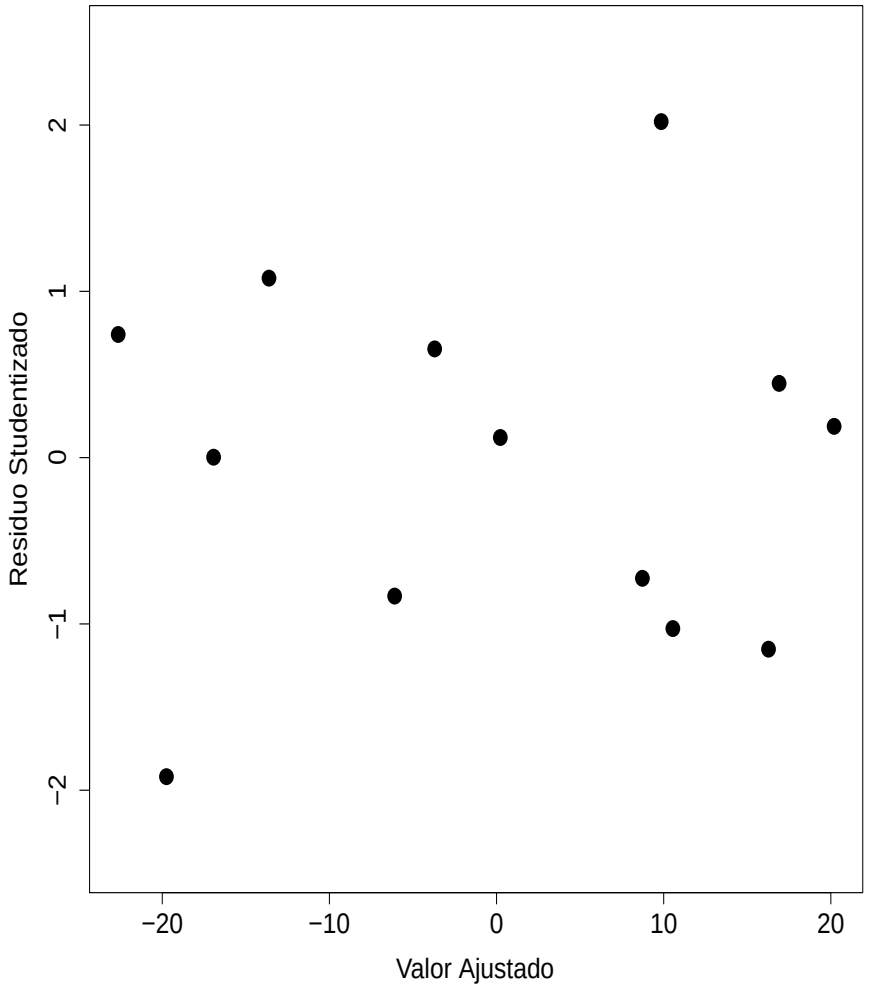
Efeito	Estimativa	E.Padrão	valor-t	valor-P
cx_1	1,5511	0,7022	2,209	0,055
cx_2	0,5102	0,6024	0,748	0,474
cx_3	0,1019	0,7115	0,143	0,889
cx_4	-0,1441	0,6685	-0,215	0,834
s	2,306			
R^2	0,982			
R^2-ajustado	0,975			

Apenas a variável X_1 é marginalmente significativa.

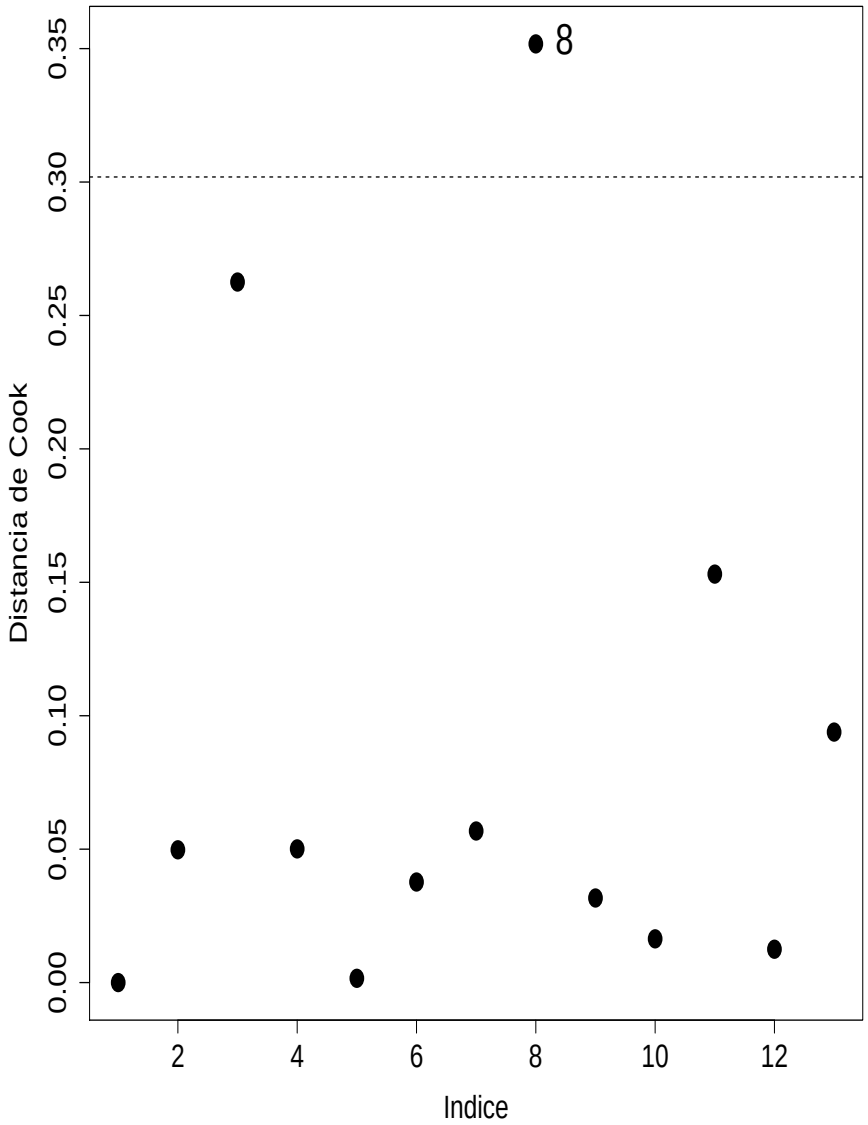
Análise de Resíduos



Análise de Resíduos



Distância de Cook



Comentários

- Os gráficos de resíduos mostram-se adequados, não há indícios de afastamentos da normalidade, de presença de observações aberrantes e de variância não constante.
- A observação #8 aparece como possivelmente influente. Quando essa observação não é considerada na regressção o valor-P correspondente à estimativa do coeficiente da variável X_1 reduz para 0,020, os demais coeficientes continuam não significativos e todos com sinal positivo.

Fator de Inflação da Variância

Valores de VIF

Variável	VIF
cx_1	38,49
cx_2	254,42
cx_3	46,87
cx_4	282,51

Portanto, pelos critérios adotados as variáveis explicativas têm VIF ≥ 10 indicando indícios de multicolinearidade.

Número da Condição

Autovalor	Valor
λ_1	6213,5625
λ_2	809,9572
λ_3	148,8652
λ_4	2,8458

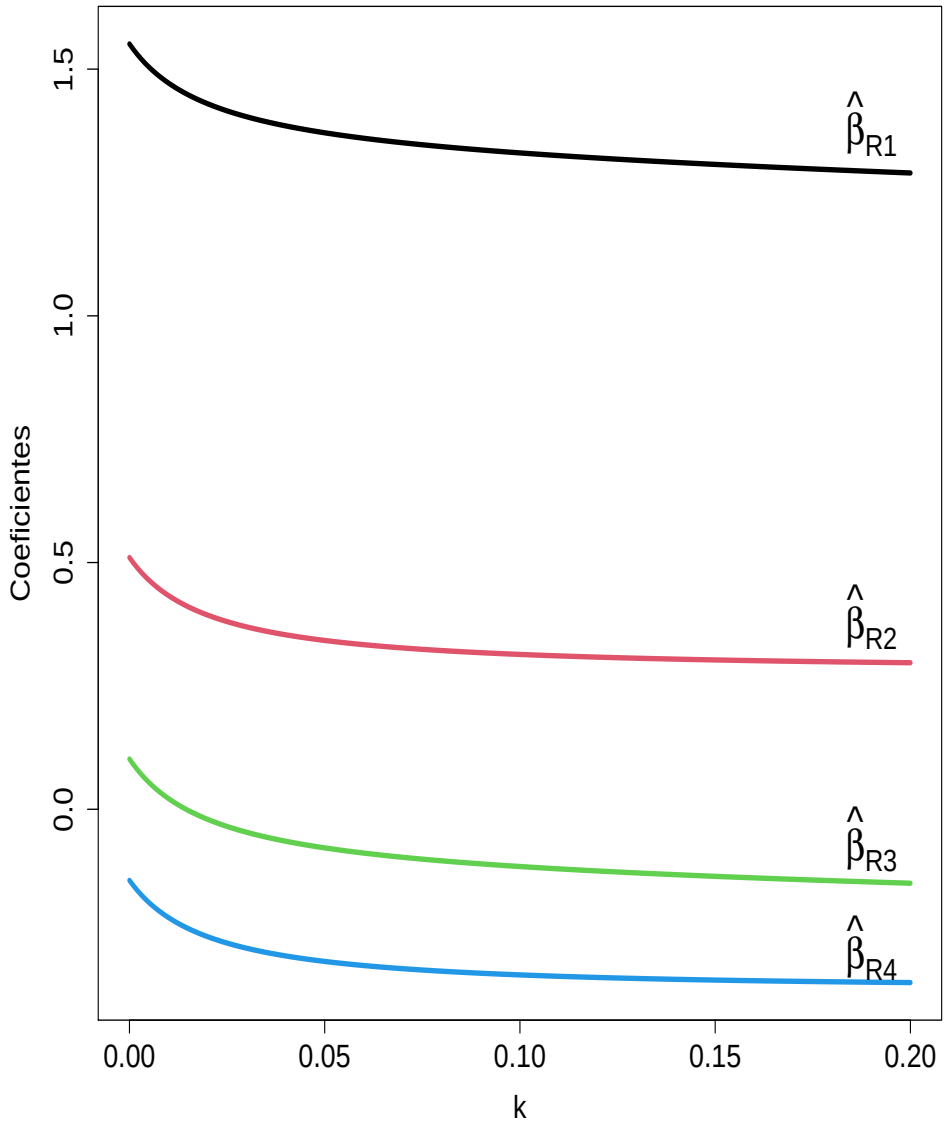
Portanto, o número da condição fica dado por $k = \frac{\lambda_{max}}{\lambda_{min}} = \frac{6213,5625}{2,8458} = 2183,42$, indicando fortes indícios de multicolinearidade.

Índices da Condição

Índice	Valor
$k_1 = \frac{\lambda_{max}}{\lambda_1}$	1,00
$k_2 = \frac{\lambda_{max}}{\lambda_2}$	7,67
$k_3 = \frac{\lambda_{max}}{\lambda_3}$	41,74
$k_4 = \frac{\lambda_{max}}{\lambda_4}$	2183,42

Logo, temos $k_4 > 1000$ indicando indícios de multicolinearidade.

Regressão Ridge



Modelo Ridge Ajustado

Estimativas para $k = 0,076$

Efeito	Estimativa	Erro padrão	valor-z
cx_1	1,3460	0,6844	1,967
cx_2	0,3236	0,6651	0,486
cx_3	-0,1018	0,6934	-0,147
cx_4	-0,3263	0,6514	-0,501

Apenas a variável X_1 é marginalmente significativa.

Autovalores

Autovalor	Valor	Explicação
λ_1	6213,5625	86,60%
λ_2	809,9572	11,29%
λ_3	148,8652	2,07%
λ_4	2,8458	0,04%

Autovetores

$\mathbf{T}_1$	$\mathbf{T}_2$	$\mathbf{T}_3$	$\mathbf{T}_4$
-0,067800	0,646018	-0,567315	0,506180
-0,678516	0,019993	0,543969	0.493268
0,029021	-0,755310	-0,403554	0,515567
0,730874	0,108480	0,468398	0,484416

Escolha de Componentes

Considerando apenas o primeiro componente principal, que explica 86,60%, tem-se que

$$z_1 = -0,067800cx_1 - 0,678516cx_2 + 0,029021cx_3 + 0,730874cx_4.$$

Então o modelo na forma canônica fica dado por

$$cy_i = z_{i1}\alpha + \epsilon_i$$

em que cy_i denota o calor da i-ésima amostra de cimento centralizado e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, 13$. Desse ajuste obtém-se $\widehat{\alpha} = -0,5537(0,1043)$.

Sumário

Considerações Finais

- Este é um exemplo em que há fortes indícios de multicolinearidade com as 4 variáveis explicativas ajustadas conjuntamente.
- Através da regressão ridge com $k = 0,076$ foi possível estabilizar as estimativas, porém não houve mudança na significância dos coeficientes.
- A variável explicativa X_1 continua sendo a única marginalmente significativa.
- Apenas o 1^{o} componente principal é suficiente para explicar as 4 variáveis explicativas.

Sumário

Referências

Referência

- Montgomery, D. C.; Peck, E. A. e Vining, G. G. (2021). *Introduction to Linear Regression Analysis, 6th Edition*. Hoboken: Wiley.