kek

UFPB - Regressão I

Paulo Ricardo Seganfredo Campana Gabriel de Jesus Pereira

2 de novembro de 2023

Resumo

ESCREVER MERDA

Introdução

O concreto é um dos matérias de construção mais utilizados na engenharia civil devido à sua durabilidade, versatilidade e resistência. Ele é composto por agregados, água e principalmente cimento. Analisando o cenário e as necessidades da engenharia civil, foi proposto a modelagem da força compressiva do concreto para a possibilidade de expandir o conhecimento sobre concreto de alta performance na indústria. Para isso, utilizamos um banco com dados experimentais de diferentes combinações de substâncias que compõem o concreto, o tempo que a mistura foi deixada para secar e a força compressiva final.

Fazendo uso desses dados, da modelagem e análise estatística, foi possível chegar em modelos de regressão linear múltipla, focaremos em um modelo mais simples, com o intuito de poder saber o que leva um concreto a ser mais resistente do que outro e para previsão da força compressiva de certa mistura baseado nas variáveis de estudo.

Metodologia

As análises a seguir foram realizadas usando a linguagem de programação R (R Core Team 2023) com o *framework* de modelagem estatística *tidymodels* (Kuhn e Wickham 2020). Os códigos utilizados estão disponíveis no github (Campana e J. Pereira 2023) e os documentos do relatório e apresentação foram feitos com Quarto (Allaire et al. 2022), um sistema de escrita e publicação científica.

No conjunto de dados sobre concreto de alta performance (Yeh 2006) estudamos um modelo de regressão linear múltipla em que a `força_compressiva` do concreto é explicada a princípio pelas variáveis que achamos importantes no estudo: o `tempo` de secagem da mistura final (em dias) e os matérias que compõem a mistura: `cimento`, `escória_de_aço`, `cinzas_pulverizadas`, `água`, `superplastificante`, `agregado_graúdo` e `agregado_miúdo` (em quilogramas por metro cúbico). Algumas destas variáveis não estiveram presente no modelo final devido a não serem significantes no modelo linear.

Sendo assim, ajustamos dois modelos, um primeiro mais simples utilizando apenas 4 dos regressores e algumas transformações com o objetivo de estabelecer uma relação compreensível das substâncias que mais interferem na força compressiva do concreto. O segundo modelo é mais complexo, trazendo a interação entre as variáveis e composição das mesmas em novas medidas, este foca no poder preditivo da regressão. Porém a complexidade deste segundo modelo, mesmo que significativa, não se provou útil para a melhoria das métricas de performance, então decidimos não incluir nos resultados finais.

Fizemos uso da transformação Yeo-Johnson (Yeo e Johnson 2000), que de maneira similar a Box-Cox, é uma transformação feita para tornar a distribuição dos regressores mais normais e estabilizar a variância, com a vantagem de também funcionar para dados que contém valores 0 e números negativos. O parâmetro λ é estimado por máxima verossimilhança.

$$\psi(\lambda, x) = \begin{cases} [(1+x)^\lambda - 1]/\lambda & \lambda \neq 0,\ x \geqslant 0 \\ \ln(1+x) & \lambda = 0,\ x \geqslant 0 \\ [(1-x)^{2-\lambda} - 1]/(\lambda - 2) & \lambda \neq 2,\ x < 0 \\ -\ln(1-x) & \lambda = 2,\ x < 0 \end{cases}$$

A escolha de variáveis e transformações usadas foram julgadas através das métricas de performance do coeficiente de determinação (R^2) e raiz do erro quadrático médio (RMSE ou σ) porém mantendo todos os coeficientes do modelo significativos nos testes de hipótese individuais.

$$R^2 = 1 - \frac{SS_{\text{resid}}}{SS_{\text{total}}} \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\hat{y}_i - y_i)^2}$$

Resultados

Construção do modelo

Primeiramente, escolhemos as variáveis que foram mais importantes para o alcance dos objetivos citados acima: `cimento`, `escória_de_aço`, `água` e `tempo`. Realizamos uma

transformação de raiz quadrada na variável resposta, `força_compressiva` e transformações Yeo-Johnson em todos os regressores exceto `água`, onde a estimativa de λ foi muito próximo de 1, desse modo temos as seguintes variáveis transformadas:

Tabela 1: Transformações realizadas nas variáveis do modelo

Variável	λ	Transformação
`força_compressiva`		$y' = \sqrt{y}$
`cimento`	0.197	$x' = 5.065[(1+x)^{0.197} - 1]$
`escória_de_aço`	0.066	$x' = 15.16[(1+x)^{0.066} - 1]$
`tempo`	-0.006	$x' = \ln(1+x)$

Devido a relação entre a força compressiva e o tempo de secagem ser não linear, criamos duas variáveis com o tempo transformado: tempo ao quadrado e tempo ao cubo. Ajustando um modelo sem intercepto com essas variáveis temos a seguinte relação entre a força compressiva do concreto (y'), a quantidade de cimento (x'_1), água (x'_2), escória de aço (x'_3) e o tempo de secagem (x'_t):

$$\hat{y}' = 0.769x'_1 + 0.186x'_2 - 0.023x'_3 + 0.396x'^2_t - 0.050x'^3_t$$

Verificação das suposições do modelo linear

Normalidade dos resíduos

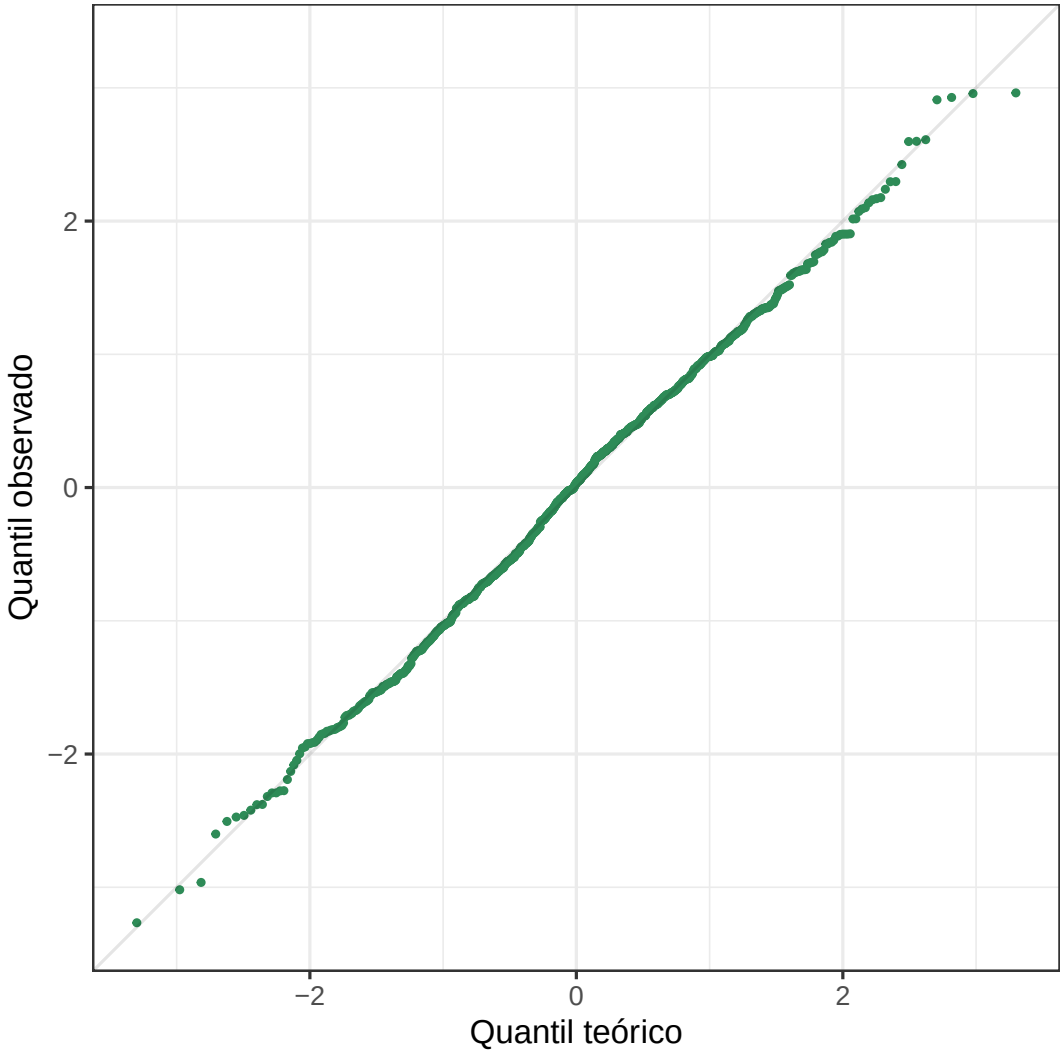
É uma suposição importante para a construção de intervalos de confiança e predição para a força compressiva, para os coeficientes e para o cálculo da estatística F do modelo, porém pequenos desvios da normalidade não afetam o modelo.

Tabela 2: Resultado dos testes para normalidade

Teste	Estatística	p-valor
Anderson-Darling	$A = 0.812$	0.035
Cramer-von Mises	$W = 0.155$	0.020
Lilliefors	$D = 0.034$	0.007
Pearson	$P = 29.97$	0.467
Shapiro-Francia	$W = 0.998$	0.263
Shapiro-Wilk	$W = 0.998$	0.217
Jarque-Bera	$JB = 1.578$	0.454

Mais da metade dos testes da Tabela 2 não rejeitam a hipótese de normalidade dos resíduos, e gráficamente pelo Q-Q plot da Figura 1 os resíduos parecem sim ter distribuição aproximadamente normal pois se assemelham aos quantis teóricos da distribuição normal, isso já cumpre as necessidades para a correta utilização dos intervalos de confiança que se baseiam na normalidade.

Figura 1: Q-Q plot dos resíduos padronizados do modelo



Linearidade

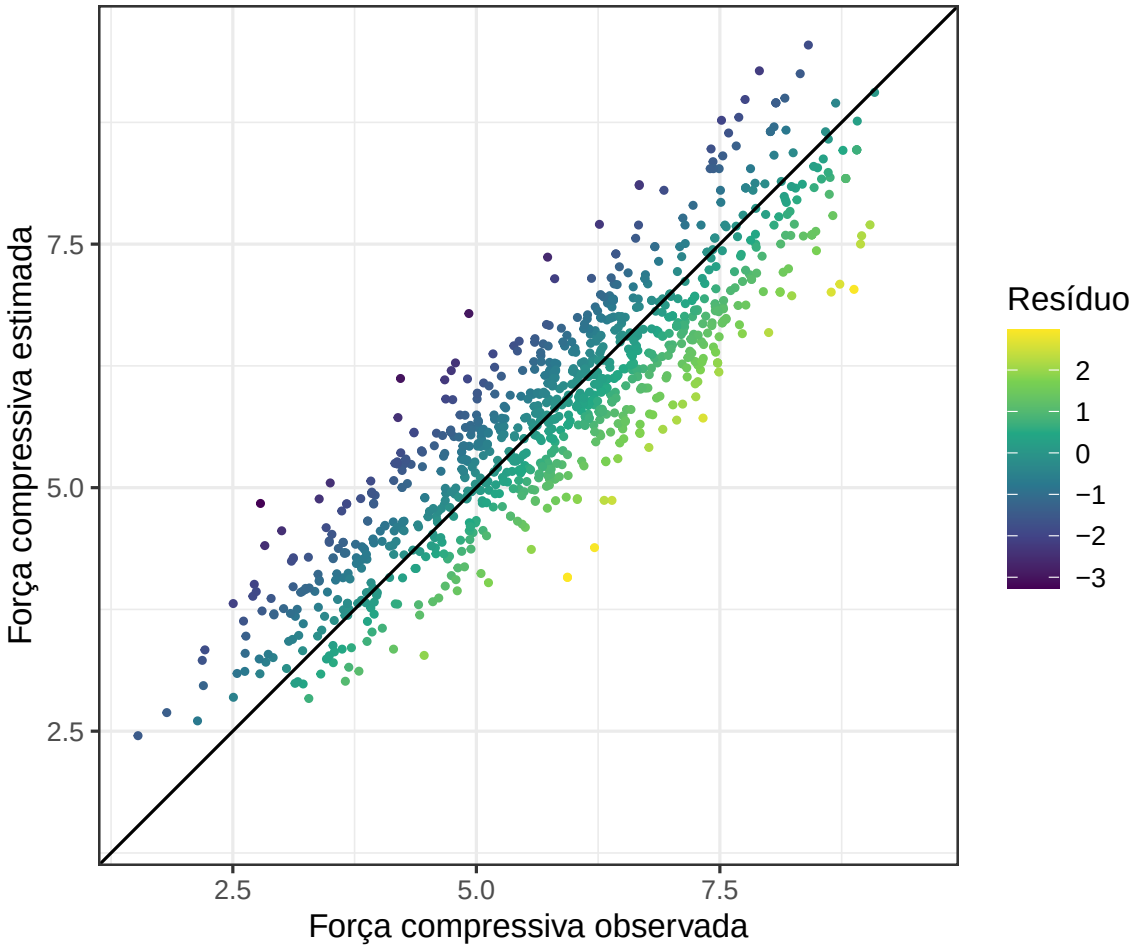
Suposição essencial para o uso de um modelo de regressão linear

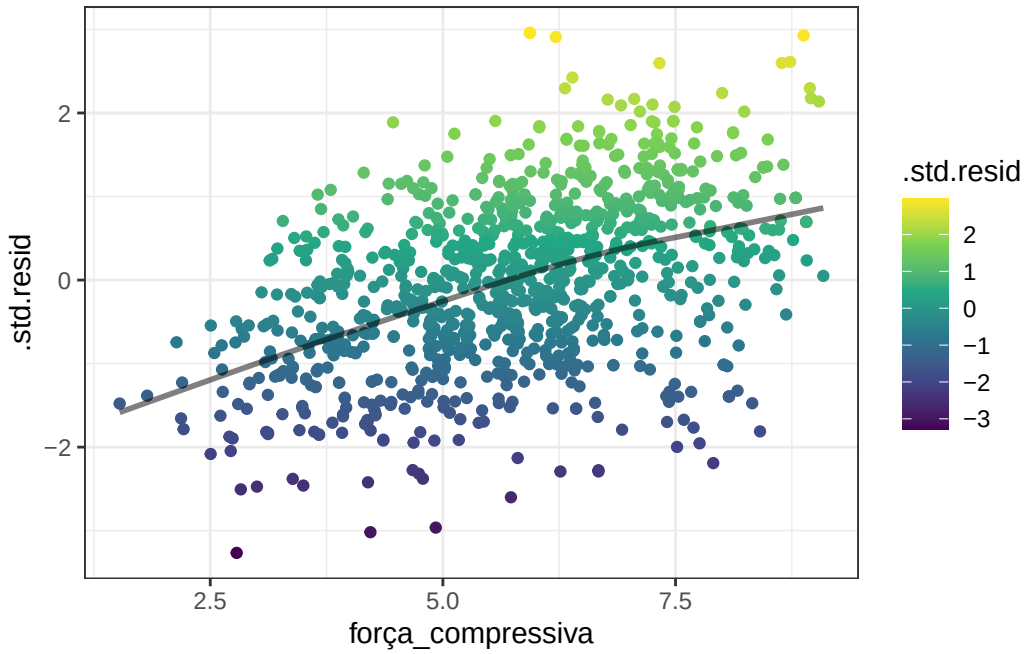
Tabela 3: Resultado dos testes para linearidade

Teste	Estatística	p-valor
RESET	$R = 0.800$	0.371
Rainbow	$R = 1.407$	5.797×10^{-5}

Os testes para linearidade do modelo discordam entre si, pelo gráfico da Figura 2, os valores estimados do modelo para a força compressiva do concreto parecem estar de acordo com os valores observados, exceto talvez para valores baixos da força compressiva, onde o modelo parece superestimar a mesma como vista na cauda esquerda do gráfico.

Figura 2: Gráfico dos valores observados versus valores estimados pelo modelo





Como se tratam de dados experimentais sobre o concreto, esse conjunto inclui várias combinações de valores diferentes entre os regressores, desse modo, a correlação entre as variáveis é baixa, atingindo no máximo 50%.

Referências

Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, e Christophie Dervieux. 2022. «Quarto». 2022. https://quarto.org.

Campana, Paulo R. S., e Gabriel de J. Pereira. 2023. «Códigos dos modelos de regressão e análise». 2023. https://github.com/cowvin0/conkrekt.

Kuhn, Max, e Hadley Wickham. 2020. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing* (versão 4.3.1). Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Yeh, I-Cheng. 2006. «Analysis of Strength of Concrete Using Design of Experiments and Neural Networks». *Journal of Materials in Civil Engineering* 18 (4): 597–604. https://doi.org/10.1061/(ASCE)0899-1561(2006)18:4(597).

Yeo, In-Kwon, e Richard A. Johnson. 2000. «A New Family of Power Transformations to Improve Normality or Symmetry». *Biometrika* 87 (4): 954–59. http://www.jstor.org/stable/2673623.