

kek

UFPB - Regressão I

Paulo Ricardo Seganfredo Campana

Gabriel de Jesus Pereira

6 de novembro de 2023

Resumo

O desenvolvimento de um país emergente está diretamente relacionado com o avanço da indústria, particularmente na construção civil, sendo uma das áreas mais importantes para a criação de empregos e o aperfeiçoamento da infraestrutura de uma nação. O concreto, um dos materiais mais importantes e presentes na construção civil e está sempre em constante desenvolvimento, no entanto, não é possível saber como podemos melhorar-lo se não antes conhecermos a seleção adequada de materiais e características da mistura. Dessa forma, entender e aprimorar as propriedades do concreto, é de importância vital para garantir a segurança e o desempenho de estruturas modernas. Com base nisso, este estudo se propõe a investigar os fatores que influenciam na força compressiva do concreto utilizando a regressão linear múltipla e técnicas estatísticas de análise da mesma. Obtendo assim uma relação numérica entre as características da criação do concreto e sua eficácia como material base para grandes estruturas.

Palavras-chave: Regressão linear, concreto, suposições do modelo.

Introdução

Como um dos materiais de construção mais utilizados na engenharia civil devido à sua durabilidade, versatilidade e resistência. O concreto 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.

Em seguida, investigamos as suposições do modelo linear e se nosso modelo satisfaz as mesmas, por meio de testes de hipótese e análises gráficas, descrevemos a eficiência do modelo com base nas métricas de performance comuns, por fim concluímos o estudo resumindo os resultados obtidos e seu possível impacto da escolha da fabricação de novas misturas de concreto.

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.

Construção do modelo

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 por 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 obteve grandes melhoras nas métricas de performance para compensar sua complexidade, então decidimos não incluir análises sobre o segundo modelo 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.

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

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 e buscando um modelo simples quando possível.

$$R^2 = 1 - \frac{SS_{\text{resid}}}{SS_{\text{total}}}$$

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

Verificação das suposições do modelo

Após o ajuste do modelo, realizamos testes para as suposições do modelo: normalidade dos resíduos, linearidade, ausência de multicolinearidade, homocedasticidade e ausência de autocorrelação. Essas suposições são necessárias para termos um modelo linear válido. Usamos um nível de significância de 5%.

- Resíduos não normais afetam a construção de intervalos de confiança e predição para a variável resposta e para os coeficientes pois dependem da distribuição normal.
- Relação não linear entre os regressores e a variável resposta sugere o use de um modelo de regressão não linear.
- Multicolinearidade entre os regressores causa grande variância na estimativa dos coeficientes, possivelmente tornando-os não significantes.
- Heterocedasticidade no modelo tornam não confiáveis as estimativas de variâncias e covariâncias para os coeficientes.
- Autocorrelação dos resíduos indica alguma dependência temporal nos dados e perda de eficiência na estimativa dos coeficientes.

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), escória de aço (x'_2), água (x_3) e o tempo de secagem (x'_t):

$$\hat{y}' = 0.769x'_1 + 0.186x'_2 - 0.023x_3 + 0.396x_t'^2 - 0.050x_t'^3$$

Verificação das suposições do modelo

A seguir, veremos os testes de hipótese e gráficos que sustentam as suposições necessárias para criação de um modelo linear válido.

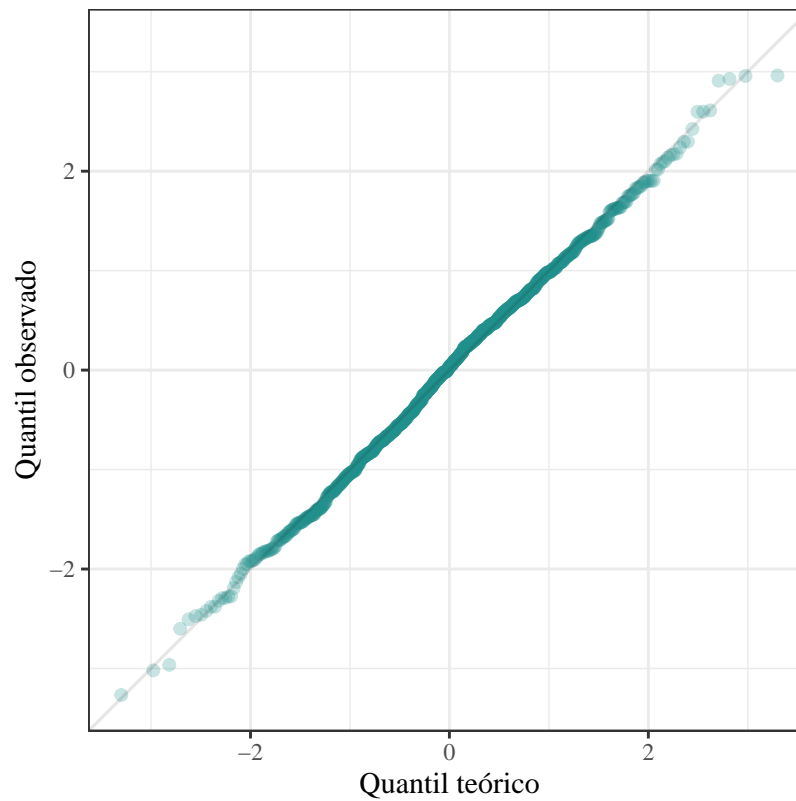
Normalidade

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 graficamente pelo Q-Q plot da Figura 1 os resíduos parecem sim ter distribuição aproximadamente normal pois se assemelham aos quantis teóricos da distribuição normal.

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



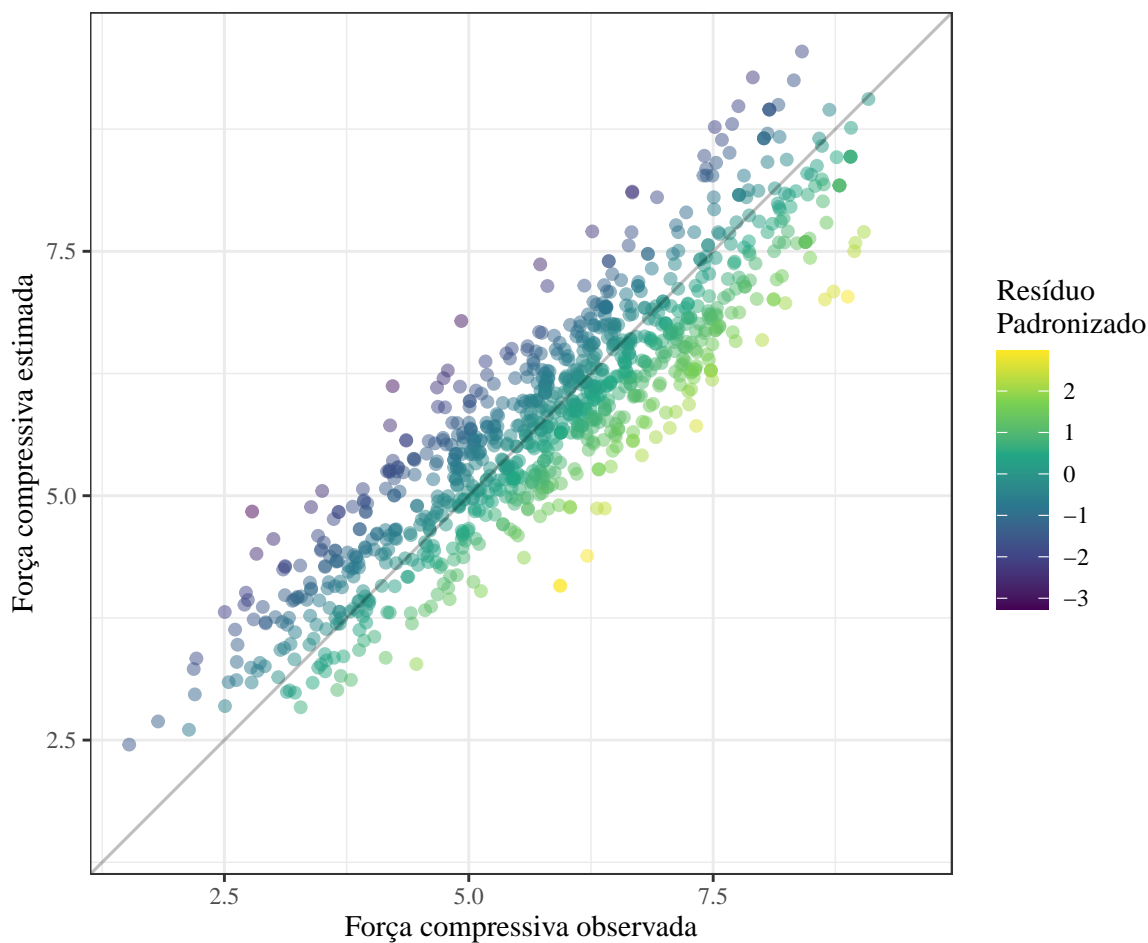
Linearidade

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



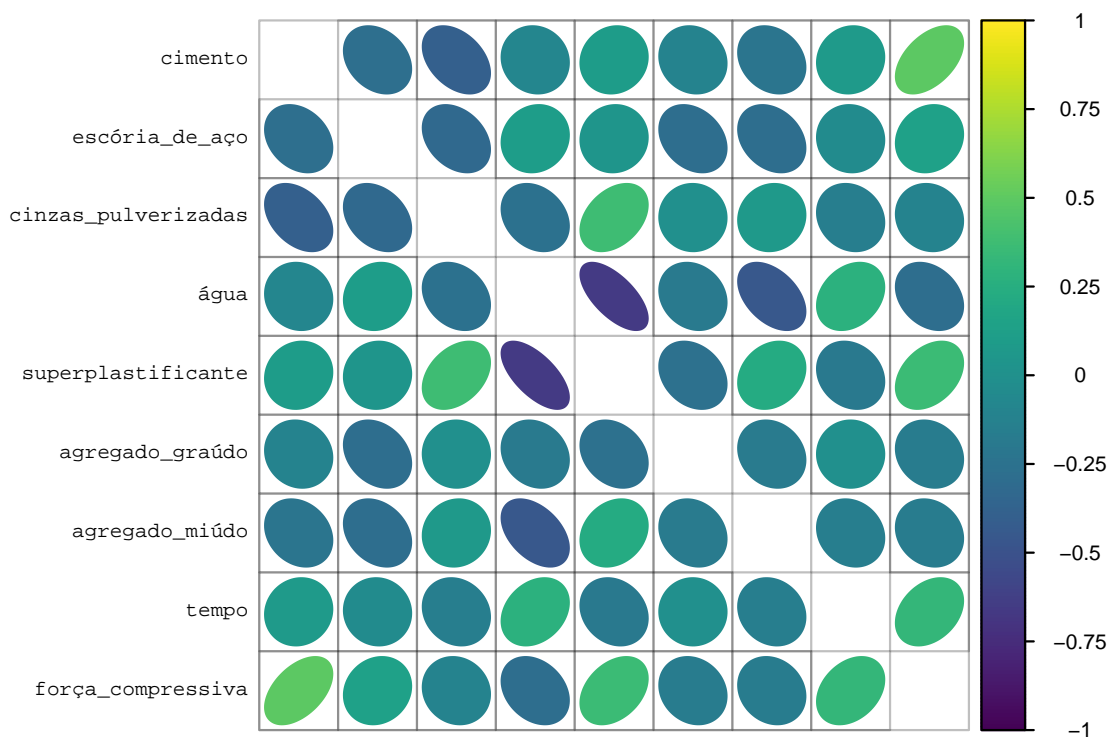
Ausência de multicolinearidade

Tabela 4: Teste t para os coeficientes do modelo

Termo	Estimativa	Erro padrão	Estatística	p-valor
x'_1	0.769	0.0134	57.36	3.36×10^{-322}
x'_2	0.186	0.0079	23.69	2.55×10^{-99}
x'_3	-0.023	0.0007	-32.58	2.16×10^{-160}
$x_t'^2$	0.396	0.0134	29.67	3.50×10^{-140}
$x_t'^3$	-0.050	0.0023	-21.53	4.17×10^{-85}

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 uma correlação de 30% entre as variáveis usadas no modelo segundo a Figura 3. Também vemos na Tabela 4 que todos os coeficientes da regressão são altamente significativos, desse modo, o modelo não possui multicolinearidade entre as variáveis.

Figura 3: Gráfico da matriz de correlação entre as variáveis estudadas



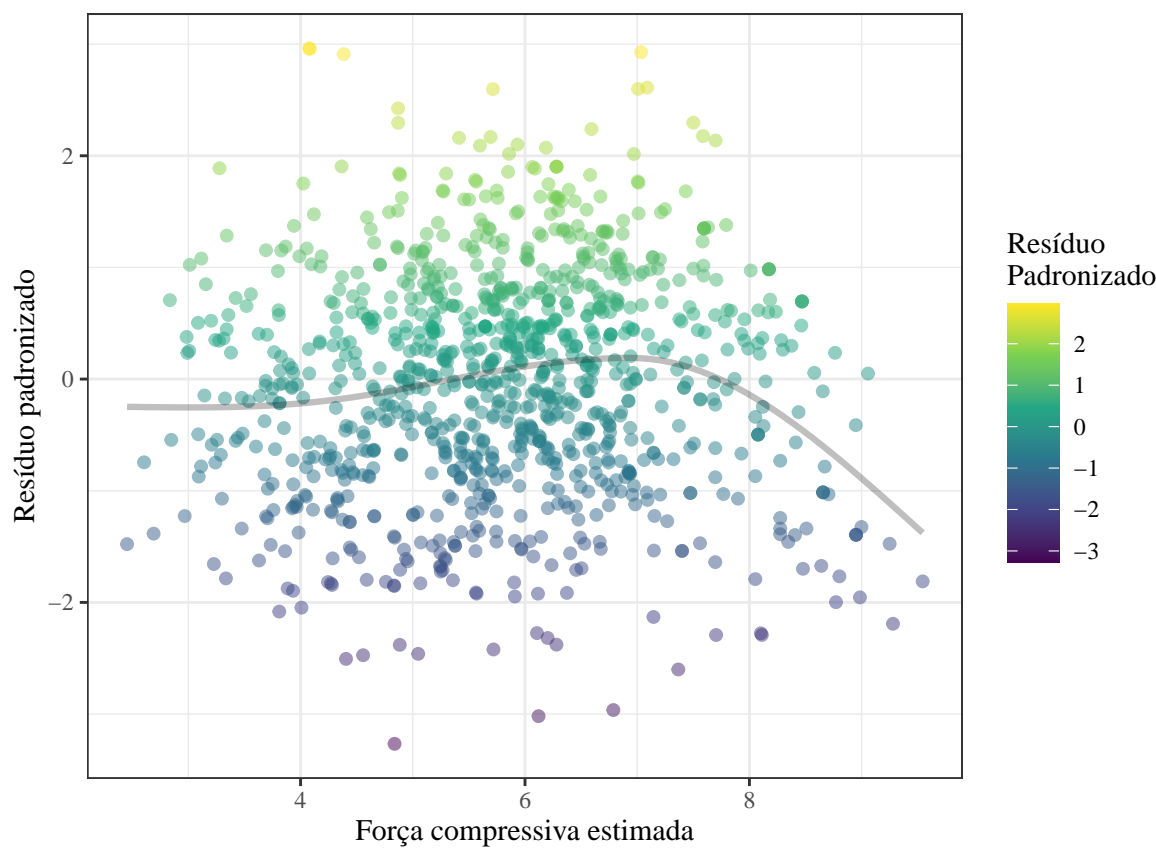
Homocedasticidade

Tabela 5: Resultado dos testes para homocedasticidade

Teste	Estatística	p-valor
Breusch-Pagan	$BP = 4.776$	0.311
Koenker	$BP = 5.155$	0.272
Goldfeld-Quandt	$GQ = 1.059$	0.258
Harrison-McCabe	$HMC = 0.505$	0.608

Todos os testes da Tabela 5 não rejeitam a hipótese do modelo ser homocedástico, com p-valores acima de 5%, isso indica que a variância do modelo é constante para todos os valores estimados da variável resposta, o gráfico da Figura 4 apoia esta suposição, porém pode existir uma pequena relação entre a média dos resíduos padronizados para diferentes valores estimados do modelo.

Figura 4: Gráfico dos valores estimados versus resíduos padronizados



Ausência de autocorrelação

Tabela 6: Resultado dos testes para autocorrelação

Teste	Estatística	p-valor
Breusch-Godfrey	$LM = 169.3$	1.05×10^{-38}
Durbin-Watson	$R = 1.187$	1.01×10^{-39}

Os resíduos do modelo são de fato correlacionados entre observações sequenciais, pois se tratam de dados experimentais, as diferentes combinações entre substâncias na mistura do concreto e tempo de secagem são feitas de modo estruturado. No conjunto de dados, é comum a ocorrência de 5 observações sequenciais com a mesma mistura porém com tempos de secagem diferentes: 3, 14, 28, 56 e 100 dias, ou dezenas de observações seguidas em que não houve presença de escória de aço na mistura.

Sendo assim, a autocorrelação dos resíduos se dá pela estrutura dos dados experimentais, não por uma relação temporal entre cada experimento.

Performance do modelo

Tabela 7: Intervalo de confiança dos coeficientes de regressão

Coeficientes	LI	LS
x'_1	0.743	0.795
x'_2	0.171	0.201
x_3	-0.024	-0.022
$x_t'^2$	0.370	0.422
$x_t'^3$	-0.055	-0.046

Aqui podemos ver o intervalo de confiança para os coeficientes de regressão, que nos permitem entender o impacto médio causado na força compressiva do concreto por cada uma variável regressora. Por exemplo, para cada kg de cimento em uma mistura de m_3 há um aumento médio na força compressiva do concreto entre 0.742 e 0.795.

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\(597\)](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>.