Modelagem e previsão da resistência à compressão do concreto

Relatório do trabalho final da A2

Modelagem Estatística

Thiago Franke Melchiors

Professor: Luiz Max Fagundes de Carvalho

RIO DE JANEIRO
2024

Conteúdo

1 Introdução

A jornada do concreto na história da humanidade não se inicia com as revoluções industriais ou mesmo na Roma antiga. Suas raízes remontam ao início da civilização, quando os primeiros humanos começaram a utilizar suas consciências emergentes para melhorar suas condições de vida através da inovação tecnológica. Essa evolução histórica do concreto, desde formações naturais até as sofisticadas misturas modernas, reflete sua importância fundamental na construção e no desenvolvimento humano [Kaefer 1998].

No cenário contemporâneo, o concreto desempenha um papel crucial em quase todas as infraestruturas. A resistência à compressão do concreto é um dos parâmetros mais críticos na engenharia civil, afetando diretamente a segurança e durabilidade das estruturas. Essa característica é determinante no design final de edifícios, pontes e outras estruturas. Em um contexto de desafios crescentes, como eventos climáticos extremos, a necessidade de desenvolver modelos preditivos precisos que estimem a resistência à compressão a partir das propriedades e composições do concreto torna-se ainda mais premente.

Outrossim, além dos aspectos relacionados à segurança e confiabilidade das estruturas, modelos preditivos precisos da resistência à compressão do concreto permitem alcançar desempenhos desejados com custos menores. Em seu estado da arte, esses métodos podem reduzir ou mesmo eliminar a necessidade de testes empíricos extensivos relacionados à resistência, proporcionando economia de tempo e recursos durante a fase de desenvolvimento e construção.

Com essa premissa, este trabalho visa explorar e analisar um conjunto de dados abrangente que documenta diversas fórmulas de concreto e suas correspondentes resistências à compressão. Através de uma análise exploratória detalhada, buscar-se-á entender como as variáveis independentes — os componentes do concreto — interagem e influenciam a propriedades final do material. Alicerçado a essas descobertas, propor-se-ão modelos estatísticos para prever a resistência à compressão a partir de um determinado traço. [1]

2 Conjunto de dados

O conjunto de dados utilizado neste estudo é o "Concrete Compressive Strength" [Yeh 2007], disponibilizado pelo UC Irvine Machine Learning Repository. Este conjunto consiste em 1030 amostras, cada uma compreendendo 8 variáveis independentes que correspondem aos componentes da mistura de concreto, além de uma variável dependente que registra a resistência à compressão do concreto. A Tabela 1 fornece uma visão geral de todas as variáveis incluídas.

[1] O notebook em python contendo a análise exploratória e os métodos estatísticos descritos neste relatório está disponível em: https://github.com/TFrankeM/concrete-compressive-strength.

Tabela 1: Descrição dos componentes da mistura de concreto

Variável	Unidade de Medida	Descrição
Cimento	kg/m^3	Mistura finamente moída de compósitos inorgânicos que quando combinados com água endurecem por hidratação
Escória de alto-forno	kg/m^3	Subproduto de siderurgia usado como adição mineral
Cinza volante	kg/m^3	Resíduo da combustão de carvão utilizado como adição pozzolânica
Água	kg/m^3	Agente de hidratação
Superplastificante	kg/m^3	Aditivo redutor de água
Agregado grosso	kg/m^3	Pedras britadas ou seixos que conferem resistência ao concreto
Agregado fino	kg/m^3	Areia, fundamental para a coesão da mistura
Idade	dias	Período de cura que afeta a hidratação e o desenvolvimento da resistência
Resistência	MPa	Medida da capacidade do concreto de suportar cargas compressivas

Cada componente da mistura de concreto desempenha um papel crucial na definição das propriedades finais do material. O cimento é o principal aglomerante e contribui significativamente para a resistência inicial e final do concreto. Porém, devido ao seu custo e impacto ambiental elevados, outros materiais são adicionados para otimizar a mistura. A escória de alto-forno e a cinza volante são adições minerais que ajudam a reduzir o custo do concreto e também aumentam sua durabilidade e resistência à corrosão. A água é indispensável para a reação de hidratação do cimento, mas sua quantidade precisa ser cuidadosamente controlada para evitar a redução da resistência e durabilidade.

Os superplastificantes são aditivos que permitem a redução do conteúdo de água na mistura, mantendo a trabalhabilidade, o que resulta em um concreto mais forte e mais durável. Os agregados grosso e fino fornecem a massa necessária para o concreto e ajudam a controlar a contração durante o processo de cura. O agregado grosso é usado para dar estrutura e reduzir o custo, pois materiais como pedras britadas são geralmente mais baratos que o cimento. Por sua vez, a areia ou agregado fino preenche os espaços entre os agregados grossos, melhorando a coesão e a resistência à compressão. A idade de cura do concreto influencia diretamente na resistência desenvolvida, sendo um fator crucial em projetos estruturais.

3 Análise exploratória dos dados

Compreender a distribuição e a contribuição individual dos componentes do concreto é fundamental para o desenvolvimento de um modelo robusto que descreva eficazmente as variáveis-chave em estudo. Esta compreensão estabelece uma base sólida para investigar detalhadamente as relações entre os componentes e a resistência à compressão do concreto, permitindo-nos explorar quantitativamente essas interações ao identificar padrões, tendências e anomalias nos dados. A presente análise exploratória pretende satisfazer essas necessidades, orientando a formulação de modelos preditivos mais precisos e eficientes para prever a resistência à compressão do concreto.

A tabela 2 apresenta estatísticas descritivas do conjunto de dados, fornecendo uma visão detalhada da natureza e distribuição dos componentes analisados.

Tabela 2: Estatísticas descritivas das variáveis

Variável	Mínimo	Média	Mediana	Moda	Máximo	Variância	Intervalo	% nulos
Cimento	102.00	281.17	272.90	362.6	540.0	10921.58	438.00	0.0
Escória de Alto-Forno	0.00	73.90	22.00	0.0	359.4	7444.12	359.40	0.0
Cinza Volante	0.00	54.19	0.00	0.0	200.1	4095.62	200.10	0.0
Água	121.80	181.57	185.00	192.0	247.0	456.00	125.20	0.0
Superplastificante	0.00	6.20	6.40	0.0	32.2	35.69	32.20	0.0
Agregado Grosso	801.00	972.92	968.00	932.0	1145.0	6045.68	344.00	0.0
Agregado Fino	594.00	773.58	779.50	594.0	992.6	6428.19	398.60	0.0
Idade	1.00	45.66	28.00	28.0	365.0	3990.44	364.00	0.0
Resistência	2.33	35.82	34.45	33.4	82.6	279.08	80.27	0.0

A partir das métricas resumidas, é evidente que não existem valores faltantes nos dados analisados, proporcionando confiabilidade nas análises subsequentes. Além disso, nota-se que as variáveis como escória de alto-forno, cinza volante e superplastificantes apresentam uma moda de zero, sugerindo que esses componentes não são cruciais em todas as misturas de concreto. Isso pode indicar que tais ingredientes são adicionados especificamente para atender a requisitos particulares de certas aplicações do concreto.

A considerável variância nas características sugere que a composição do concreto é bastante diversificada, refletindo a adaptabilidade da mistura às necessidades específicas de diferentes contextos de construção. Tal diversidade de traços permite otimizar o concreto para diferentes propriedades, como durabilidade e resistência, dependendo do projeto.

Apesar da variabilidade nas composições, o intervalo de valores para a resistência à compressão do concreto, que é a variável-alvo, mostra-se relativamente estreito. Isso ocorre porque o concreto mais comumente utilizado em ambientes urbanos possui uma resistência de cerca de 30 MPa. Os concretos classificados como de alto desempenho (CAD) exibem resistências que variam de 40 a 100 MPa, dependendo das especificações e exigências do projeto. Concretos de resistências mais elevadas - chamados de concretos de ultra-alto desempenho (CUAD) - são ocasionalmente empregados em grandes infraestruturas ou em moldes para peças industriais, situações em que a durabilidade e a resistência mecânica são cruciais [AECweb 2022]. Portanto, os dados refletem as composições mais frequentemente utilizadas, enquanto valores superiores têm aplicações mais especializadas e relevância reduzida neste estudo.

Dado que as distribuições de dados podem variar significativamente mesmo quando compartilham estatísticas descritivas semelhantes, é crucial visualizar essas distribuições para uma compreensão mais profunda. A figura a seguir apresenta gráficos do tipo violino para cada atributo. Nestes gráficos, as curvas representam a densidade de dados em cada valor específico ao longo do eixo y. No centro de cada gráfico, um boxplot vertical revela a mediana através de uma linha branca. As extremidades superior e inferior do retângulo mais grosso correspondem, respectivamente, ao terceiro e primeiro quartis. As linhas finas, os "bigodes", se estendem a partir deste retângulo até $1,5$ vezes o intervalo interquartil. Valores além desse intervalor correspondem a potenciais *outliers*.

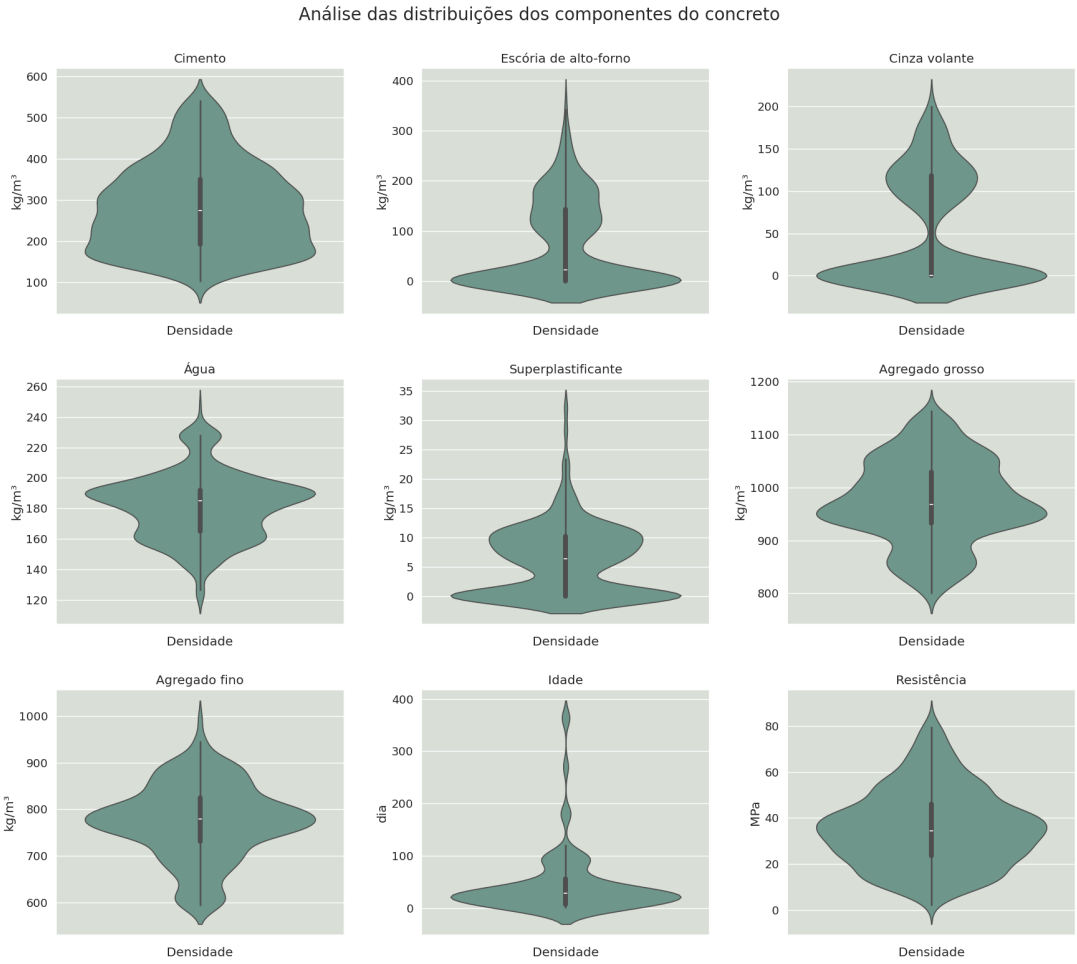

Figura 1: Gráficos do tipo violino para todos os atributo do dataset.

Observa-se uma variação relevante na distribuição do cimento, com densidades elevadas cobrindo uma ampla faixa de valores, o que sugere uma flexibilidade na sua dosagem conforme as necessidades de cada projeto. Em contraste, os componentes como água, agregado grosso e agregado fino apresentam picos de densidade mais definidos, indicando que existem proporções comumente adotadas nessas variáveis. Por outro lado, escória de alto-forno e cinza volante mostram picos mais altos em valores baixos, assinalando que, frequentemente, pequenas quantidades desses materiais são usadas.

Essas revelações apontam para a existência de traços mais padronizados ou preferidos para esses materiais nas misturas de concreto.

Após analisar a distribuição dos componentes do concreto, é crucial entender como as variáveis interagem entre si para influenciar a resistência do concreto. A matriz de correlação é uma ferramenta essencial nessa análise, pois quantifica o grau de relação entre cada par de variáveis. Altas correlações entre variáveis independentes podem indicar redundância, permitindo, em alguns casos, a eliminação de uma das variáveis do modelo para simplificar a análise sem perda expressiva de informação. Correlações próximas de 1 indicam uma relação positiva (quando uma variável aumenta, a outra também aumenta), enquanto valores próximos de -1 indicam uma relação negativa (quando uma variável aumenta, a outra diminui).

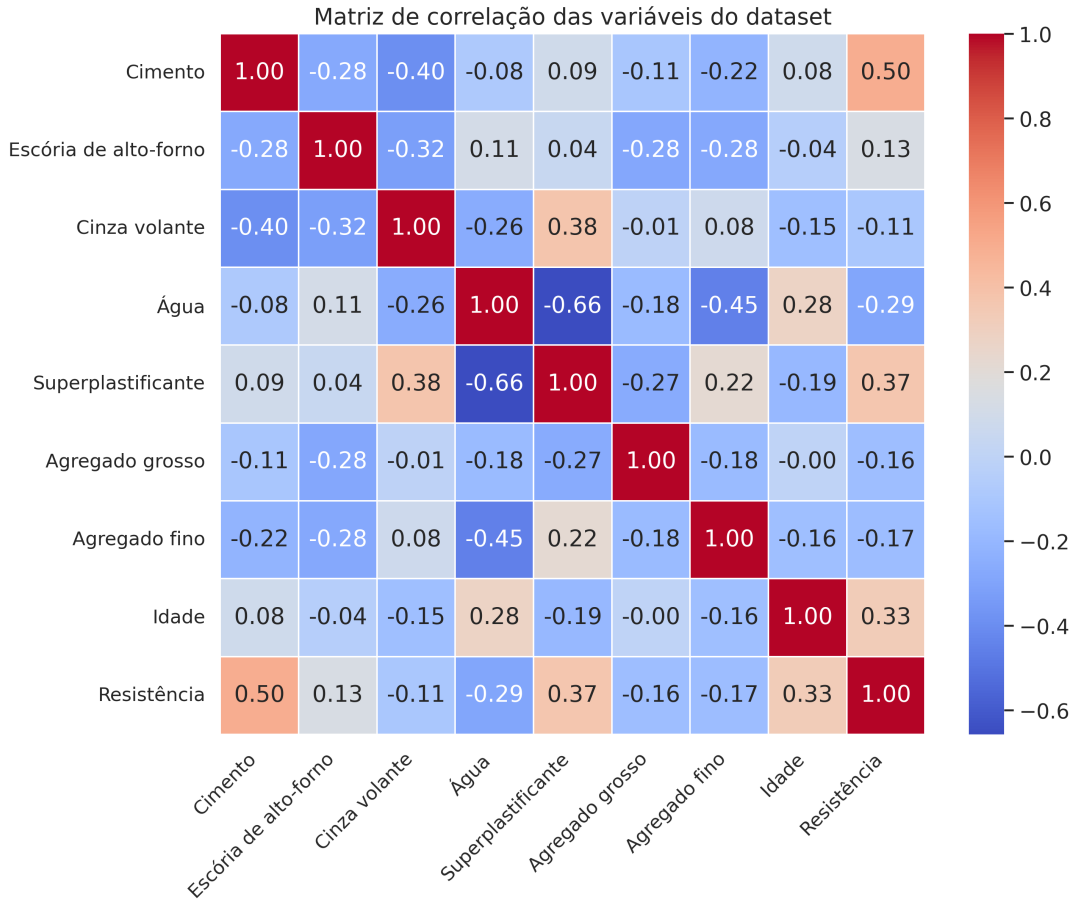

Figura 2: Matriz de correlação das variáveis do dataset, mostrando as interrelações entre os componentes do concreto e sua resistência.

A correlação de 0.5 entre cimento e resistência indica uma relação moderadamente positiva, sugerindo que um aumento na quantidade de cimento geralmente resulta em um aumento na resistência do concreto. Da mesma forma, a resistência também mostra correlações positivas com a idade do concreto e a quantidade de superplastificante, implicando que o concreto tende a se tornar mais resistente com o tempo e com o uso de aditivos que melhoram suas propriedades. Por outro lado, a correlação negativa entre a resistência e a quantidade de água reforça a noção de que a água, embora essencial para ativar o cimento, pode diminuir a resistência do concreto quando usada em excesso.

Apesar de que evidenciam-se correlações significativas em algumas variáveis, a complexidade das interações químicas e físicas no concreto muitas vezes complica a interpretação direta de certos valores com resultados empíricos. Por exemplo, é possível quantificar e correlacionar diferentes proporções de substituição do cimento por escória de alto-forno e cinza volante com a resistência do concreto, bem como analisar diferentes teores das relações água/aglomerante e sua influência na resistência. Essas correlações são mais evidentes quando lidamos com proporções e teores, mas não necessariamente se aplicam a variáveis como o agregado grosso e o cimento, onde as relações podem não ser tão diretas ou previsíveis [Melchiors 2023]. Cada variável, conforme evidenciado na matriz de correlação, contribui de maneira única para as propriedades finais do concreto e, considerando a complexidade dessas interações,

nenhuma pode ser prontamente descartada sem uma análise cuidadosa do impacto potencial em todo o sistema.

Para complementar a análise quantitativa fornecida pela matriz de correlação, gráficos de dispersão foram gerados para cada componente do concreto em relação à sua resistência à compressão. Esses gráficos corroboram as correlações numéricas observadas anteriormente, e oferecem uma perspectiva visual sobre a forma e a consistência dessas relações. Cada gráfico é acompanhado por uma linha de regressão, que facilita a compreensão das tendências e a identificação de possíveis discrepâncias ou outliers que poderiam influenciar futuras investigações e ajustes no modelo de previsão.

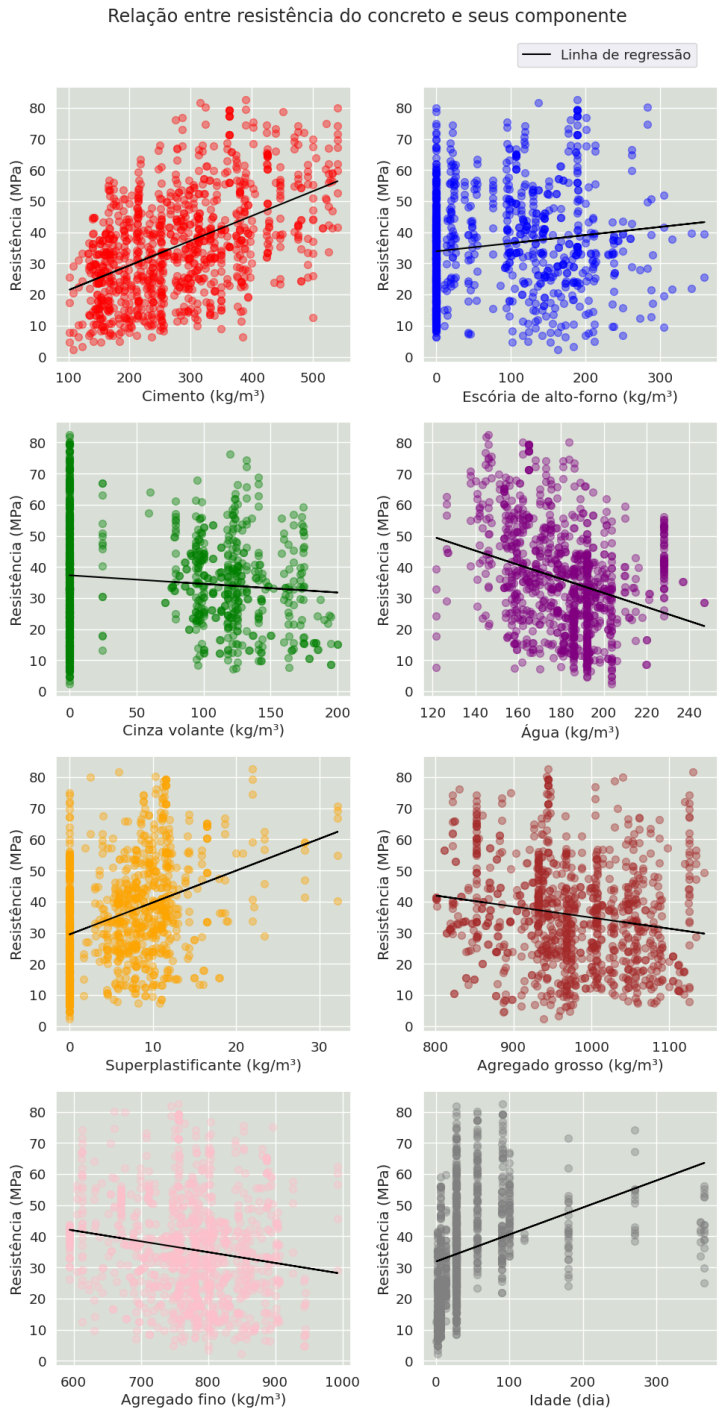

Figura 3: Análise de dispersão mostrando a relação entre cada componente do concreto e sua resistência à compressão. Linhas de regressão são incluídas para destacar as tendências predominantes e auxiliar na visualização do grau e direção da correlação entre as variáveis.

Os gráficos de dispersão confirmam e expandem as análises previamente realizadas, ilustrando as dinâmicas entre os componentes do concreto e sua resistência à compressão. O cimento, o superplastificante e a idade do concreto apresentam uma correlação positiva com a resistência, sugerindo que aumentos nestes parâmetros tendem a melhorar a resistência do concreto. Por outro lado, outros componentes mostram corre-

lações negativas com a resistência, indicando que seu aumento pode não ser benéfico para a resistência à compressão. Importante notar que a escória de alto-forno, a cinza volante e até certo ponto o superplastificante não são essenciais para atingir altos níveis de resistência à compressão, sendo muitas vezes utilizados para atender a outros requisitos técnicos ou ambientais do traço.

Outrossim, a dispersão verificada nos dados realça a complexidade das interações entre os componentes do concreto. Portanto, para alcançar alta precisão, um modelo preditivo deve ser capaz de capturar essas variações difusas em toda a gama de valores possíveis da resistência. Alternativamente, pode-se concentrar em um subconjunto mais restrito da variável dependente, onde as relações sejam mais consistentes e previsíveis.

4 Análise estatística e modelagem preditiva

Com base nos diagnósticos e análises preliminares realizadas, é possível propor um modelo de previsão mais preciso que capture o padrão geral dos dados, evitando o sobreajuste causado pelo ruído e robusto o suficiente para prever novos registros de maneira eficaz.

4.1 Avaliação do modelo

Para avaliar a qualidade dos preditores e definir o mais adequado para o problema proposto, será usada uma combinação de métodos, que são descritos a seguir.

1. **P-valor e Valor t**: o p-valor é usado para determinar a significância estatística da hipótese de que o coeficiente de uma variável é zero (não tem efeito). Um valor t pequeno e um p-valor abaixo de um limiar (geralmente 0,05) indicam que é improvável que o coeficiente seja zero, sugerindo que a variável é importante para o modelo. O valor t é uma medida da magnitude e direção desse efeito, baseada na estimativa do coeficiente e no erro padrão.

2. **Intervalo de confiança de 95%**: fornece uma faixa dentro da qual os valores verdadeiros dos coeficientes provavelmente se encontram com 95% de confiança. Intervalos que incluem zero sugerem que a variável pode não ser significativa.

3. **Coeficiente de Determinação (R^2)**: mede a proporção da variação na constante dependente que é explicada pelas variáveis preditoras. A fórmula é:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}},$$

sendo que SS_{res} é a soma dos quadrados dos resíduos, dado por $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, onde $\hat{y}_i$ é o valor estimado de y_i; e SS_{tot} é a soma total dos quadrados, dado por $\sum_{i=1}^{n}(y_i - \bar{y})^2$, onde $\bar{y}$ é a média das observações.

O R^2 varia entre 0 e 1 e expressa a quantidade da variância dos dados que é explicada pelo modelo linear. Quanto maior o R^2, mais o modelo se ajusta à amostra. Valores excessivamente altos podem indicar sobreajuste, especialmente em modelos com muitas variáveis.

4. **Root Mean Square Error (RMSE)**: mede o erro médio realizado pelo modelo ao prever o resultado para uma instância. Valores mais baixos indicam um melhor ajuste do modelo. É definido como:

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

   sendo y_i os valores reais e $\hat{y}_i$ os valores preditos pelo modelo.

5. **Critério de Informação de Akaike (AIC)**: estima a quantidade relativa de informação perdida pelo modelo: quanto menos informações um modelo perde, maior a qualidade desse modelo e menor a pontuação AIC. O valor de AIC do modelo considerado é dado por:

$$AIC = 2k - 2\ln(L),$$

   no qual k é o número de parâmetros no modelo e L é a verossimilhança máxima do modelo.

A formulação do AIC é essencialmente uma resposta ao fenômeno de que a adição de variáveis adicionais a um modelo tende a aumentar artificialmente o R^2 e a diminuir o RMSE, indicando um melhor desempenho durante o treinamento, mesmo que essas variáveis não contribuam profundamente para a capacidade preditiva do modelo. Por equilibrar complexidade e eficácia, o AIC incorpora uma penalidade por cada variável adicional incluída.

Adicionalmente, será verificado se os **resíduos do modelo** seguem uma distribuição normal. Os resíduos, definidos matematicamente como $e_i = y_i - \hat{y}_i$, onde y_i são os valores reais e $\hat{y}_i$ são os valores estimados pelo modelo, devem idealmente exibir características específicas para confirmar a adequação do modelo. Estas características incluem:

- **Média próxima de zero**: Isso indica que o modelo não possui viés sistemático nos erros de predição.
- **Independência**: Os resíduos não devem exibir autocorrelação, o que sugere que cada erro de predição é independente dos outros.
- **Homocedasticidade**: Os resíduos devem ter variância constante em relação aos níveis de predição, indicando que a precisão do modelo é uniforme em toda a gama de valores estimados.
- **Distribuição normal do resíduo**: Para justificar o uso de muitos testes estatísticos que dependem dessa premissa, como o teste t.

4.2 Modelo de Regressão Linear Multivariada

Considerando a faixa relativamente estreita dos valores observados da variável dependente, postula-se que a resistência à compressão do concreto, apesar de sua natureza altamente não linear, pode ser aproximada por um hiperplano dentro do intervalo

limitado de dados disponíveis. Para explorar essa hipótese, será empregado um modelo de regressão linear multivariada utilizando o método de Mínimos Quadrados Ordinários (Ordinary Least Squares - OLS), cuja fórmula que estima os coeficientes $\boldsymbol{\beta}$ que minimizam os erros de predição é:

$$\boldsymbol{\beta} = (X^T X)^{-1} X^T \mathbf{y},$$

na qual

- X representa a matriz de características com uma coluna adicional de uns para o intercepto, tornando-a uma matriz $n \times (p+1)$, onde n é o número de registros e p é o número de variáveis preditoras.
- $(X^T X)^{-1}$ existe se $X^T X$ for uma matriz de posto completo, isto é, todas as colunas de X são linearmente independentes.
- $\mathbf{y}$ é o vetor da variável dependente.

Do ponto de vista estatístico frequentista, o OLS procura o hiperplano que melhor se ajusta aos dados através da minimização da soma dos quadrados dos resíduos, de modo que a diferença entre os valores verdadeiros e os valores modelados seja a menor possível.

A seleção das características que serão usadas no modelo como variáveis preditoras é conduzida através da seguinte heurística de eliminação para trás:

1. Inicialmente, todas as variáveis são incluídas para estabelecer uma linha de base.
2. O modelo é ajustado, e a variável com o maior `p-valor` acima de 0.05 é removida.
3. A etapa 2 é repetida até que todas as variáveis remanescentes tenham `p-valores` abaixo ou igual ao limiar de significância.

Esta heurística é utilizada para garantir que apenas variáveis com uma contribuição efetiva sejam mantidas, em linha com o princípio da parcimônia, inspirado pela Navalha de Ockham: "entre modelos equivalentes, o mais simples deve ser escolhido". Em outras palavras, busca-se um modelo que se ajuste bem aos dados e mantenha a simplicidade para evitar a complexidade desnecessária. Para isso, o modelo com o menor AIC é priorizado.

5 Apresentação dos resultados

Os dados foram separados em conjunto de treinamento e teste, com 824 e 206 registros (proporção de quatro para um), respectivamente. Em seguida, o conjunto de treinamento foi submetido ao procedimento descrito anteriormente. A tabela 3 resume as características dos três modelos de regressão "ensinados".

Os resultados obtidos indicam que todos os três modelos de regressão testados apresentam desempenhos bastante similares, mas com diferenças sutis que destacam alguns pontos interessantes sobre a natureza dos dados e a eficácia dos modelos. O modelo 1, que inclui todas as oito variáveis preditoras, alcançou o mais alto Coeficiente de Determinação (R^2), 0.610524, sugerindo que é capaz de explicar aproximadamente

Tabela 3: Características dos modelos de regressão aprendidos

Modelo	R^2	AIC	RMSE	RMSE no teste	Num preditores	Variáveis preditoras
1	0.6105	6234.4232	10.5187	9.7964	8	Cimento, Escória de alto-forno, Cinza volante, Água, Superplastificante, Agregado grosso, Agregado fino, Idade
2	0.6091	6235.4046	10.5378	9.8151	7	Cimento, Escória de alto-forno, Cinza volante, Água, Superplastificante, Agregado fino, Idade
3	0.6085	6234.5984	10.5454	9.7796	6	Cimento, Escória de alto-forno, Cinza volante, Água, Superplastificante, Idade

Nota: Todos os modelos incluem um intercepto como parâmetro. O número de preditores listado não inclui o intercepto.

61.05% da variância da resistência à compressão do concreto. Este modelo também apresentou o menor AIC, 6234.423209 e o menor RMSE, 10.518787 no treinamento, com um desempenho sólido também no conjunto de teste (RMSE de de 9.7964), indicando um bom ajuste aos dados de treinamento e uma boa generalização para dados novos.

À medida que variáveis menos relevantes são removidas nos modelos subsequentes, há uma ligeira redução no R^2 e um pequeno aumento no RMSE, tanto no treinamento quanto no teste. O terceiro modelo, que mantém apenas seis variáveis preditoras, ainda apresenta um desempenho comparável ao dos outros modelos, com uma perda mínima de poder explicativo.

As variáveis removidas nos modelos 2 e 3, embora menos críticas, ainda contribuem marginalmente para a precisão do modelo. Este resultado suporta a ideia de que um modelo mais complexo, neste contexto, não necessariamente implica em sobreajuste, mas em uma captação mais completa da complexidade dos dados de resistência à compressão do concreto. Isso é corroborado pela análise exploratória, onde foi constatado que todas as variáveis apresentavam alguma independência na matriz de correlação.

Além disso, a imagem 4 ilustra os resíduos dos três modelos. Na primeira coluna, o histograma sugere que os resíduos dos 3 modelos possuem uma distribuição aproximadamente normal, centrada em torno de zero, indicando a ausência de viés sistemático nos modelos. Os gráficos Q-Q, na segunda coluna, complementam essa verificação ao mostrar os quantis dos resíduos do modelo contra os quantis teóricos de uma distribuição normal. Os quantis teóricos são gerados artificialmente pelo código, equivalentes a uma distribuição normal ideal para a amostra de dados. A conformidade dos pontos com a linha diagonal nesses gráficos indica que a distribuição dos resíduos se aproxima da normalidade: quanto mais alinhados os pontos estiverem com a linha, mais próxima da normalidade está a distribuição dos resíduos.

Por fim, os gráficos de dispersão na terceira coluna relacionam os valores previstos pelo modelo com os resíduos. A distribuição homogênea dos pontos em um intervalo vertical ao longo de todo o eixo x sugere homocedasticidade, ou seja, a variância dos resíduos é constante em relação aos valores previstos - a validade de diversos testes estatísticos necessita dessa característica -, um indicativo de que o modelo se ajusta equitativamente bem em toda a gama de valores previstos. Por outro lado, tem-se um caso de heterocedasticidade quando a variância dos resíduos varia com os valores previstos - os resíduos não estão distribuídos uniformemente.

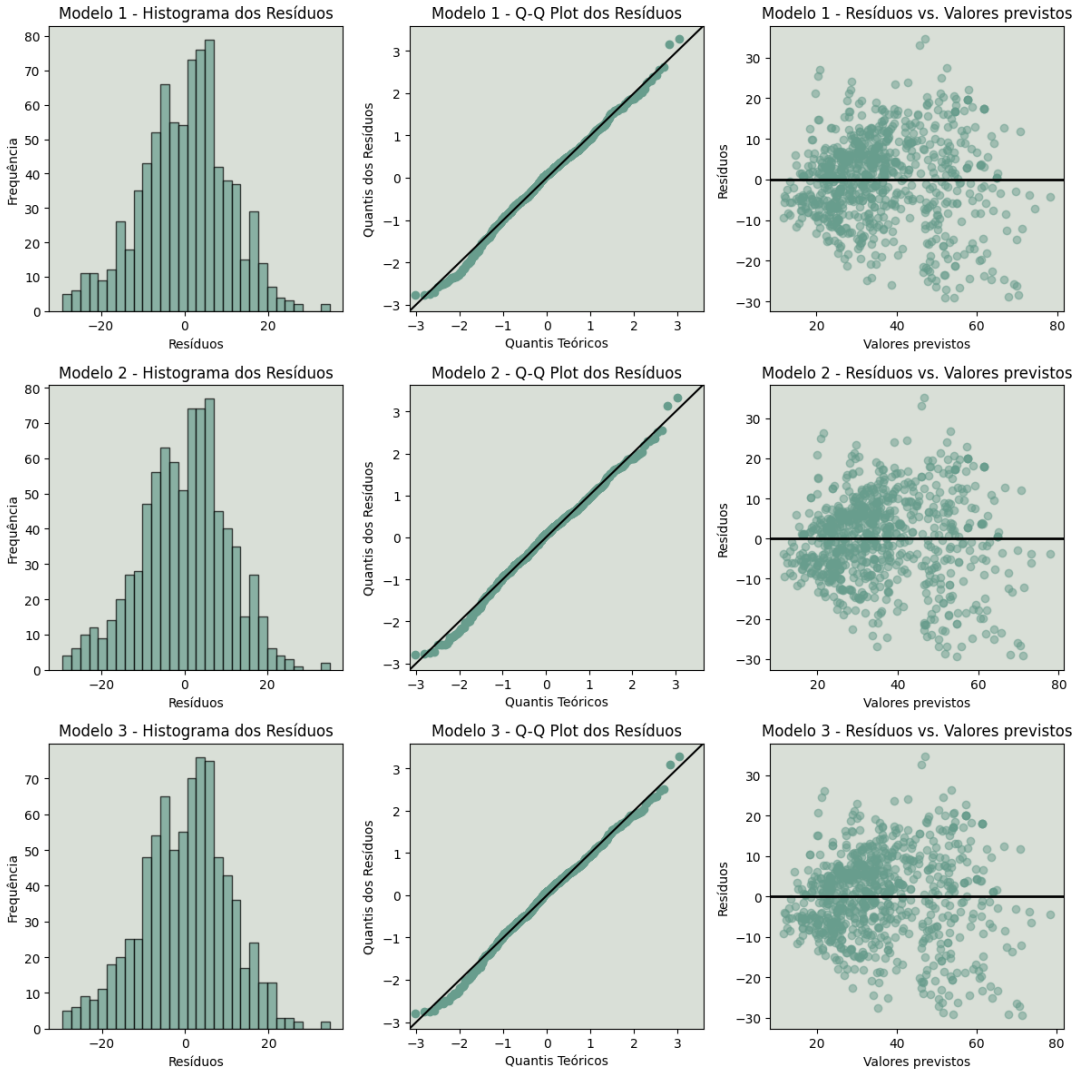

Figura 4: Análise dos resíduos para os três modelos de regressão estudados, mostrando histogramas, gráficos Q-Q e gráficos de dispersão de resíduos vs. valores ajustados. Cada linha corresponde a um modelo, evidenciando a distribuição e o ajuste dos resíduos.

Embora observe-se uma aparente tendência de aumento nos resíduos à medida que os valores previstos crescem, seria necessário uma investigação mais aprofundada para verificar se esse padrão é indicativo de intensa heterocedasticidade - apuração que não será feita no presente trabalho.

Os gráficos de dispersão, que manifestam um aumento na variância dos resíduos à medida que os valores previstos aumentam, sugere que o modelo está sujeito a erros maiores quando faz previsões de maior magnitude. Isto implica que, em média, o modelo é mais preciso ao prever valores menores; consequentemente, os erros tendem a ser menores para essas previsões. Esse fenômeno pode indicar uma inadequação do modelo em capturar todas as variabilidades dos dados, especialmente nas faixas superiores da variável dependente.

A tabela 4 aprofunda a avaliação dos parâmetros do modelo 1, identificado como o de melhor desempenho.

Tabela 4: Resultados da regressão linear para o modelo 1

Variável	Coef	Std Err	t	P > |t|	[0.025, 0.975]
const	-28.8236	29.331	-0.983	0.326	[-86.396, 28.749]
Cimento	0.1190	0.009	12.682	0.000	[0.101, 0.137]
Escória de alto-forno	0.1085	0.011	9.589	0.000	[0.086, 0.131]
Cinza volante	0.0822	0.014	5.809	0.000	[0.054, 0.110]
Água	-0.1353	0.044	-3.066	0.002	[-0.222, -0.049]
Superplastificante	0.3106	0.104	2.978	0.003	[0.106, 0.515]
Agregado grosso	0.0179	0.010	1.719	0.086	[-0.003, 0.038]
Agregado fino	0.0241	0.012	2.034	0.042	[0.001, 0.047]
Idade	0.1164	0.006	18.120	0.000	[0.104, 0.129]

As variáveis cimento, escória de alto-forno, cinza volante e idade são altamente significativas, com um p-valor (praticamente) nulo e um intervalo de confiança que não inclui zero, reforçando sua forte associação positiva com a resistência à compressão, assim como a importância do processo de cura.

Enquanto isso, o superplastificante e o agregado fino exibem um impacto positivo marginalmente relevante, porquanto seus p-valores estão perto do limite de significância (0.05), especialmente esse último atributo.

Por outro lado, o intercepto, água e o agregado grosso têm um efeito negativo sobre a resistência assinalado pelos seus valores t negativos (exceto para o agregado grosso) e pelos p-valores maiores que 0.05 (salvo o da água).

Essas observações podem ter um paralelo interessante com o uso prático desses materiais: os agregados grosso e fino são frequentemente utilizados para reduzir o custo do concreto, não necessariamente para aumentar a resistência; no que tange à água, quantidades excessivas desse componente diminuem a resistência do concreto.

6 Explorações complementares

Simultaneamente às análises utilizando a regressão linear comum, foram explorados modelos alternativos, a fim de aferir qual o ideal para o problema. Porém, por razões que serão discutidas a seguir, eles não participaram da análise principal.

O primeiro foi o Modelo Linear Generalizado (generalized linear model, GLM), utilizando uma família de distribuição gaussiana com uma função de ligação identidade, tornando-o equivalente à rergessão comum. Por essa razão, o GLM replicou os resultados dos modelos apresentados na seção anterior.

A outra experimentação realizada foi com a Regressão Polinomial multivariada, que consiste em elevar as variáveis preditoras a potências superiores e interações entre elas, para capturar relações não lineares entre as variáveis e a resposta. As implementações, limitadas a polinômios de grau dois, alcançaram valores de R^2 superiores e menores AIC e RMSE do que o modelo 1 do OLS - tabela 5. Contudo, as métricas melhores não parecem justificar a grande quantidade de parâmetros do modelos.

O Modelo Polinomial 1, alcunhado PR 1, possui 45 parâmetros, que incluem o intercepto, as 8 variáveis independentes, e todas as combinações únicas das variáveus duas

Tabela 5: Comparação de desempenho entre regressão polinomial (PR) e ordinal (OLS)

Modelo	R^2	AIC	RMSE	RMSE de teste	Número de parâmetros
OLS 1	0.6105	6234.4232	10.5187	9.7964	9
PR 1	0.8131	5701.5183	7.2871	7.4554	45
PR 2	0.8131	5699.5200	7.2871	7.4541	44
PR 3	0.8131	5697.5351	7.2872	7.4564	43
PR 4	0.8131	5695.5824	7.2874	7.4561	42
PR 5	0.8129	5694.1028	7.2897	7.4826	41
PR 6	0.8128	5692.7241	7.2925	7.5372	40
PR 7	0.8127	5691.3423	7.2952	7.5589	39
PR 8	0.8124	5690.5628	7.3006	7.5391	38
PR 9	0.8118	5691.3248	7.3128	7.5189	37

a duas. Os modelos subsequentes foram definidos a partir da regra de eliminação do parâmetro que excede o limiar de significância estabelecido, nesse caso, 0.05.

7 Conclusão

O presente relatório percorreu todas as etapas de exploração, estudo e proposição de inferências para um conjunto de dados que apresenta uma complexa relação entre as variáveis. Esta foi uma excelente oportunidade para colocar em prática estratégias de análise estatística por meio de gráficos e métricas, bem como aparatos técnicos de modelagem estatística, incluindo modelos de previsão e sua avaliação de desempenho.

A regressão linear simples apresentou resultados satisfatórios considerando o objetivo didático do trabalho e que o conjunto de dados reflete de forma limitada a realidade. Contudo, os resultados apresentados não configuram como um avanço expressivo para a academia ou para o mercado da construção civil. Insistir na regressão, mesmo utilizando uma versão polinomial com muitos parâmetros, provavelmente é menos benéfico do que optar diretamente por um modelo de aprendizado de máquina não linear.

Ademais, há uma chance considerável de que tenha sido identificado um caso de heterocedasticidade no OLS, o que reforça a necessidade de revisar o modelo. Isso pode envolver o uso de métodos mais sofisticados que capturem com maior precisão a natureza intrincada dos dados, como Florestas Aleatórias ou Redes Neurais, ou a aplicação de técnicas de transformação dos dados ou de ponderação na estimação, para alcançar uma maior homogeneidade na variância dos resíduos. Não obstante, como esta é uma tarefa de Modelagem Estatística, optou-se por uma abordagem com visão probabilística, em detrimento de modelos de aprendizado de máquina mais profundos.

Finalmente, o conjunto de dados utilizado não é representativo da ampla gama de variações de resistência do concreto que ocorrem na prática, não incluindo, por exemplo, registros de concretos de ultra-alto desempenho (aqueles com resistência superior a aproximadamente 100 MPa). Em um cenário com dados mais variados, uma regressão linear provavelmente não seria satisfatória.

Referências

AECweb 2022 AECWEB. *O que é Concreto de Alto Desempenho*. 2022. AECweb: O portal da Arquitetura, Engenharia e Construção. Disponível em: <https://www.aecweb.com.br/academy/aec-responde/o-que-e-concreto-de-alto-desempenho/23814>.

Kaefer 1998 KAEFER, L. F. *A Evolução do Concreto Armado*. 1998. PEF 5707 – Concepção, Projeto e Realização das Estruturas: Aspectos Históricos. Disponível em: <https://wwwp.feb.unesp.br/lutt/Concreto%20Protendido/HistoriadoConcreto.pdf>.

Melchiors 2023 MELCHIORS, E. F. *Método de Dosagem de Concretos de Alto e Ultra-Alto Desempenho Autonivelantes*. Dissertação (Trabalho de Conclusão de Curso) — Universidade Federal de Santa Maria, Santa Maria, 2023.

Yeh 2007 YEH, I.-C. *Concrete Compressive Strength*. 2007. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5PK67>.