Universidade Estadual do Ceará
Centro de Ciências e Tecnologia
Programa de Pós-Graduação em Ciência da Computação
Mestrado Acadêmico Em Ciência Da Computação

Relatório Técnico I

Regressão: *Naval Propulsion Plants*

Carolina Araújo Dias

Maio de 2022

Resumo

Neste trabalho introduzimos o conceito de regressão linear através do método dos mínimos quadrados, utilizando técnicas oriundas da Álgebra Linear. A decomposição QR é utilizada para solucionar um sistema de equações e com isso encontrar seus coeficientes lineares. Ademais, são citadas diversas aplicações desse método no mundo real e o utilizamos para resolver um problema relacionado à indústria naval. O conjunto de dados em questão, *Naval Propulsion Plants*, é analisado e seus coeficientes são calculados utilizando a linguagem *Python* em um *Jupyter Notebook*. São mostrados os resultados da raiz quadrada do erro médio quadrático (RMSE) para os dados em estudo.

Palavras-chave: regressão, linear, álgebra.

Conteúdo

1 Introdução

Na matemática, algumas vezes, não é imediata a ligação entre a teoria e a prática. Mas no contexto da álgebra linear conseguimos encontrar muitas aplicações diretas de suas fórmulas e teoremas ao mundo real.

Uma dessas aplicações está presente na tarefa de regressão linear para obter coeficientes de um sistema linear. Isso é relevante em diversas atividades reais, como prever quantas vendas serão realizadas em determinada loja apenas com informações sobre número de pessoas e horário. Ou também, podemos calcular qual a dosagem de remédio deve ser aplicada em um paciente com informações sobre seu peso e outras informações fisiológicas.

Entendimento do cálculo envolvido nesse problema de regressão linear é tão importante quanto conhecer onde podemos aplicá-lo. Para a teoria, temos o método dos mínimos quadrados aliado à decomposição QR de uma matriz de dados A. Na prática, é comum utilizá-lo para cálculos computacionais, pela sua velocidade e praticidade em encontrar os coeficientes de grande matrizes de dados com diversas medições e variáveis.

2 Trabalhos Relacionados

O tema da regressão linear utilizando o método dos mínimos quadrados data do início do século XIX com Legendre em 1805 e Gauss em 1809. Essa primeira aplicação resultou na predição de movimentos planetários. [1]

Atualmente esse método é amplamente utilizado nos mais diversos setores e aplicações, como uma forma mais simples e direta de realizar previsões lineares, antes de utilizar métodos mais avançados e computacionalmente caros como redes neurais.

Em [2], os autores mostram diversos exemplos de aplicação da regressão linear pelo método dos mínimos quadrados, que incluem, mas não se limitam à, estimação de parâmetros para a análise da qualidade de alimentos, utilizando dados como a qualidade da água utilizada; aproximação de dados sobre a poluição do ar pela concentração de NO em uma cidade, com informações sobre a quantidade de carros em cada período durante um dia, entre outros.

As aplicações também estão cada vez mais específicas. Em [3], o método os mínimos quadrados é atualizado e utilizado em espectros infravermelhos para a estimativa da concentração de mistura de multicomponentes em amostras biológicas. Já em [4], vemos aplicações em problemas de química atmosférica, que, por sua natureza incerta, trazem diversos desafios para a aplicação.

Finalmente, em [5], a regressão está intimamente ligada ao problema de prever variáveis dependentes de fatores relacionados à indústria naval. Analisaremos mais a fundo esse problema e conjunto de dados.

3 Fundamentação Teórica

No contexto de realizar uma regressão linear, utilizamos um conjunto de dados relacionado à Fábrica de Propulsões Navais (*Naval Propulsion Plants*). Esses dados foram gerados através de um simulador numérico de turbinas de gás. [6]

Existem 11.934 medições, com 16 atributos somados a mais duas variáveis dependentes, totalizando um vetor com 18 valores para cada umas das 11.934 medições. Tratamos das variáveis dependentes uma por vez, separadamente. Inicialmente, utilizamos a *GT Compressor decay state coefficient* (chamaremos a partir de agora de *GTC*). Para a outra variável dependente, a *GT Turbine decay state coefficient* (*GTT*), o processo é análogo.

Com isso, queremos encontrar solução para o sistema de equações abaixo:

$$\begin{aligned} \alpha_1 x_{11} + \alpha_2 x_{12} + \ldots + \alpha_{16} x_{116} + \alpha_{17} &= GTC_1 \\ \alpha_1 x_{21} + \alpha_2 x_{22} + \ldots + \alpha_{16} x_{216} + \alpha_{17} &= GTC_2 \\ \vdots &= \vdots \\ \alpha_1 x_{119341} + \alpha_2 x_{119342} + \ldots + \alpha_{16} x_{1193416} + \alpha_{17} &= GTC_{11934} \end{aligned}$$

Esse sistema equivale, matricialmente, à

$$\begin{bmatrix} x_{11} & x_{12} & \ldots & x_{116} \\ x_{21} & x_{22} & \ldots & x_{216} \\ \vdots & \vdots & \ddots & \vdots \\ x_{119341} & x_{119342} & \ldots & x_{1193416} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{16} \end{bmatrix} = \begin{bmatrix} GTC_1 \\ GTC_2 \\ \vdots \\ GTC_{11934} \end{bmatrix}$$

Esse sistema $A\mathbf{x} = \mathbf{b}$ é inconsistente e para encontrar uma solução aproximada projetamos o vetor $\mathbf{b}$ no espaço-nulo esquerdo de A e buscamos a solução da equação normal

$$A^T A\tilde{\mathbf{x}} = A^T\mathbf{b}. \tag{1}$$

Calculamos então a decomposição QR de A, $A = QR$, e substituímos na equação acima para obter

$$R\tilde{\mathbf{x}} = Q^T\mathbf{b}.$$

O sistema 1 possui solução única se a matriz quadrada $A^T A$ for invertível. Para que isso ocorra, deve valer a

Proposição 1. *Seja A uma matriz $m \times n$, com $m \geq n$. Se A possuir n colunas linearmente independentes, então $A^T A$ será invertível.*

Para provar essa proposição, precisamos dos seguintes resultados, retirados de [7].

Teorema 1. *Se A é uma matriz $m \times n$, então são equivalentes as seguintes afirmações:*

a. As colunas A são vetores linearmente independentes.

b. O sistema homogêneo $A\mathbf{x} = \mathbf{0}$ possui somente a solução trivial.

c. O sistema $A\mathbf{x} = \mathbf{b}$ possui no máximo uma solução para cada $\mathbf{b}$.

Demonstração. (a) $\Longrightarrow$ (b) Se as colunas de A forem vetores L.I., então a equação vetorial que representa a combinação linear dessas colunas, $A\mathbf{x} = \mathbf{0}$, só possuirá uma única solução, que é $\mathbf{x} = \mathbf{0}$.

(b) $\Longrightarrow$ (a) Reciprocamente, se o sistema $A\mathbf{x} = \mathbf{0}$ possuir apenas a solução trivial, então a única maneira de combinar linearmente as colunas de A a fim de anulá-las será se os coeficientes da combinação linear forem todos nulos. Isto é, a única maneira de $A\mathbf{x} = \mathbf{0}$ será com $\mathbf{x} = \mathbf{0}$.

(b) $\Longrightarrow$ (c) Suponhamos que $A\mathbf{x} = \mathbf{0}$ possua apenas a solução trivial. O sistema $A\mathbf{x} = \mathbf{b}$ ou é inconsistente ou é consistente. Se for inconsistente, não possuirá nenhuma solução e a tese estará provada. Se for consistente, suponhamos que $\tilde{\mathbf{x}}$ e $\tilde{\mathbf{z}}$ sejam duas soluções de $A\mathbf{x} = \mathbf{b}$. Então, $A(\tilde{\mathbf{x}} - \tilde{\mathbf{z}}) = A\tilde{\mathbf{x}} - A\tilde{\mathbf{z}} = \mathbf{b} - \mathbf{b} = \mathbf{0}$, *i.e.*, $\tilde{\mathbf{x}} - \tilde{\mathbf{z}}$ será solução do sistema homogêneo. Mas o sistema homogêneo só possui a solução trivial, logo $\tilde{\mathbf{x}} = \tilde{\mathbf{z}}$.

(c) $\Longrightarrow$ (b) Se o sistema $A\mathbf{x} = \mathbf{b}$ possuir no máximo uma solução para cada $\mathbf{b}$, tomemos $\mathbf{b} = \mathbf{0}$, então o sistema $A\mathbf{x} = \mathbf{0}$ possuirá no máximo uma solução, mas sendo ele um sistema homogêneo, haverá sempre a solução trivial, portanto o sistema $A\mathbf{x} = \mathbf{0}$ possuirá apenas a solução trivial. $\square$

Proposição 2. *$A^T A$ possui o mesmo espaço-nulo de A.*

Demonstração. Se $\mathbf{x} \in \mathcal{N}(A)$, então $A\mathbf{x} = \mathbf{0}$. Logo $(A^T A)\mathbf{x} = A^T(A\mathbf{x}) = A^T\mathbf{0} = \mathbf{0}$, isto é, $\mathbf{x} \in \mathcal{N}(A^T A)$, portanto, $\mathcal{A} \subset \mathcal{N}(A^T A)$.

Reciprocamente, se $\mathbf{x} \in \mathcal{N}(A^T A)$, então $A^T A\mathbf{x} = 0$. Mas, então, $||A\mathbf{x}||^2 = (A\mathbf{x})^T(A\mathbf{x}) = \mathbf{x}^T A^T A\mathbf{x} = \mathbf{x}^T\mathbf{0}$, *i.e.*, $A\mathbf{x} = \mathbf{0}$, o que significa que $\mathbf{x} \in \mathcal{N}(A)$. Portanto, $\mathcal{A} \supset \mathcal{N}(A^T A)$, e podemos concluir que $\mathcal{A} = \mathcal{N}(A^T A)$. $\square$

Teorema 2. *Seja A uma matriz quadrada $n \times n$. São equivalentes as seguintes afirmações:*

a. A é invertível.

b. O sistema homogêneo $A\mathbf{x} = \mathbf{0}$ possui somente a solução trivial.

c. posto$(A) = n$.

d. O sistema $A\mathbf{x} = \mathbf{b}$ possui uma única solução para cada vetor $\mathbf{b}$.

Demonstração. (a) $\implies$ (b) Se A for invertível e $A\mathbf{x} = \mathbf{0}$, então $\mathbf{x} = A^{-1}A\mathbf{x} = A^{-1}\mathbf{0} = \mathbf{0}$.

(b) $\implies$ (c) Se $A\mathbf{x} = \mathbf{0}$ possuir somente a solução trivial, então $\mathcal{N}(A) = \{0\}$. Logo, $posto(A) = n - nul(A) = n - 0 = n$.

(c) $\implies$ (d) Se posto(A) = n, então o sistema $A\mathbf{x} = \mathbf{0}$ será consistente. Como posto(A) = n, então, pelo Teorema Fundamental da Álgebra Linear, $nul(A) = 0$, *i.e.*, o sistema $A\mathbf{x} = \mathbf{0}$ possuirá apenas uma solução. Pelo Teorema 1, o sistema consistente $A\mathbf{x} = \mathbf{b}$ possuirá no máximo uma solução, logo, possuirá exatamente uma solução para cada vetor $\mathbf{b}$.

(d) $\implies$ (a) Se $A\mathbf{x} = \mathbf{b}$ possuir uma única solução para cada vetor $\mathbf{b}$, podemos fazer $\mathbf{b} = \mathbf{e_i}$ sucessivamente para cada vetor da base canônica do $\mathbb{R}^n$, obtendo as respectivas soluções (únicas) $\mathbf{c_i}$. Assim,

$$AC = A \begin{bmatrix} | & | & & | \\ \mathbf{c_1} & \mathbf{c_2} & \dots & \mathbf{c_n} \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ A\mathbf{c_1} & A\mathbf{c_2} & \dots & A\mathbf{c_n} \\ | & | & & | \end{bmatrix} =$$

$$= \begin{bmatrix} | & | & & | \\ \mathbf{e_1} & \mathbf{e_2} & \dots & \mathbf{e_n} \\ | & | & & | \end{bmatrix} = I.$$

Portanto C será a matriz inversa de A, *i.e.*, A será invertível. □

Demonstração. (da Proposição 1) Como $n \leq m$ e A possui n colunas L.I., então, pelo Teorema 1, posto(A) = n. Mas $nul(A) = n - posto(A) = 0$, o que significa que $\mathcal{N}(A) = \{\mathbf{0}\}$. Pela Proposição 2, $(\mathcal{A}^T\mathcal{A}) = \{\mathbf{0}\}$. Pelo Teorema 2, a matriz A^TA é invertível. □

Assim, queremos que nosso conjunto de dados possua n colunas linearmente independentes, onde n = posto(A). Vamos calcular o posto da matriz de dados original após a remoção das duas variáveis independentes. Assim, temos 16 variáveis restantes. Agora adicionamos uma coluna composta apenas do número 1 ao final da matriz. Ficamos, assim, com um vetor de tamanho 17. Ao calcularmos o posto dessa matriz, utilizando a função do *NumPy linalg.matrix_rank()*, obtemos que $posto(A) = 14$. Ou seja, existem 3 colunas que são linearmente dependentes nesse conjunto de dados.

Para encontramos essas colunas, podemos realizar a decomposição LU da matriz A, $A = LU$ e encontrar as colunas correspondentes às colunas sem pivôs na matriz U. Mas, nesse caso, isso não é necessário. Ao olharmos para a matriz A, conseguimos observar que existem duas colunas constantes e uma coluna que é repetição de outra. Confirmamos que esse é realmente o caso, com funções específicas do *NumPy* e do *Pandas*, e removemos essas colunas do conjunto de dados.

Agora possuímos um conjunto de dados em forma de matriz com 14 variáveis e, ao conferir o posto dessa matriz, vemos que ele é 14. Isso nos diz que agora todas as colunas são linearmente independentes, e podemos prosseguir para o cálculo da decomposição QR de A, do modo detalhado acima.

3.1 Questões

1. Mostrar que, usando a decomposição QR de A (isto é, $A = QR$), a equação normal $A^T A\tilde{\mathbf{x}} = A^T\mathbf{b}$ pode ser escrita como $R\tilde{\mathbf{x}} = Q^T\mathbf{b}$.

Solução.

$$\begin{aligned}
A &= QR \\
A^T &= R^T Q^T \\
A^T A &= R^T Q^T A, \text{mas } A = QR, \text{então} \\
A^T A &= R^T Q^T QR \\
A^T A &= R^T R, \text{pois como Q é ortogonal, vale } Q^{-1} = Q^T.
\end{aligned}$$

Daí $A^T A\tilde{\mathbf{x}} = A^T\mathbf{b} \implies R^T R\tilde{\mathbf{x}} = R^T Q^T\mathbf{b}$.
Mas R é invertível, então $R\tilde{\mathbf{x}} = Q^T\mathbf{b}$.

2. Qual é a condição sobre a matriz de dados A, no item anterior, para que a matriz R seja invertível? Demonstrar sua afirmação.

Solução. Para R ser invertível, não podem existir zeros na sua diagonal principal. Isso nos diz que $A_{m\times n}$ têm colunas L.I., ou seja, posto(A) $= n$, logo A é invertível.

Se existirem zeros na diagonal principal de R então algum elemento $||u_i||$ é zero, então A possui colunas L.D.

3. Suponha que as colunas $\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}$ da matriz A sejam **linearmente dependentes**, mas $\mathbf{a_2}, \mathbf{a_3}, \ldots, \mathbf{a_n}$ sejam **linearmente independentes**. O que isso diz sobre o vetor de características $\mathbf{a_1}$? Por que podemos descartá-lo para realizar a regressão linear?

Solução. Isso nos diz que o vetor $\mathbf{a_1}$ é combinação linear dos outros vetores da matriz. Podemos descartá-lo porque ele pode ser encontrado através dessa combinação linear, portanto ele é um vetor redundante para a regressão linear.

4. Verificar numericamente que $Q^T Q = I$, para o respectivo banco de dados.

Solução. Utilizando o comando

```
M = np.matmul(Q.T, Q)
np.allclose(M, np.eye(M.shape[0]))
```

conseguimos descobrir se vale $Q^T Q = I$. O resultado desse comando foi *True*, portanto realmente vale $Q^T Q = I$ para o conjunto de dados utilizado.

4 Metodologia

Para a realização da regressão linear pelo método dos mínimos quadrados para o presente banco de dados foi utilizada a linguagem *Python*, versão 3.8.10, em um *Jupyter Notebook*. Também foram utilizadas bibliotecas que auxiliam na manipulação de dados e matrizes, como *NumPy* e *Pandas*, e bibliotecas de visualização de dados, como a *Matplotlib*. Por fim, foram utilizadas duas função da biblioteca de aprendizado de máquina *Scikit-Learn*, a `train_test_split` para a separação dos dados em treino e teste, e a `mean_squared_error` para o cálculo da raiz quadrada do erro médio quadrático (RMSE).

Com isso, iremos aplicar o método dos mínimos quadrados para o conjunto de dados em questão e analisar seus resultados tanto numéricos, através do RMSE, como visuais, através dos gráficos produzidos.

Faremos todo o processo duas vezes, uma para cada variável dependente, que aqui chamamos de GTC e GTT.

5 Experimentos

Após reduzirmos a matriz original em uma matriz apenas com colunas linearmente independentes, separamos os dados em dados de treino e dados de teste. Ficamos, assim, com 4 matrizes: *X_train, y_train, X_test, y_test.*

Finalmente, calculamos a decomposição QR da matriz de treinamento *X_train*, já com a coluna de 1s $[1\ 1\ \ldots\ 1]^T$ adicionada. Para isso, utilizamos:

```
Q, R = np.linalg.qr(add_ones_column(X_train))
```

Agora encontramos os coeficientes lineares $[\alpha_1, \alpha_2, \ldots, \alpha_{14}]$ com o comando

```
coefs_lineares = np.linalg.solve(R, np.dot(Q.T, y_train))
```

Com os coeficientes lineares podemos calcular os valores de GTC, a variável dependente, para cada um dos vetores medidos, tanto para o conjunto de treino, como para o conjunto de teste. Calculamos para o conjunto de treino apenas para comparar os resultados com o resultado obtido para o conjunto de teste. Para obter um vetor *y_train_preds* com as predições para *y_train* fazemos

```
y_train_preds = []
for i in range(len(X_train)):
 y_train_preds.append(np.dot(np.squeeze(coefs_lineares),
                             add_ones_column(X_train)[i]))
```

E analogamente para o vetor de teste.

Possuímos, então, vetores com valores reais e vetores com valores calculados a partir dos coeficientes lineares obtidos. Com isso podemos calcular o erro entre essa predição e o valor de fato. Utilizamos, aqui, a métrica da **raiz quadrada do erro médio quadrático** (RMSE), dada pela equação

$$RMSE = \sqrt{\frac{(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_m - \hat{y}_m)^2}{m}}, \tag{2}$$

sendo m a dimensão dos vetores y e *y_preds*, tanto para treino como para teste. Nesse caso, $m_{treino} = 8353$ e $m_{teste} = 3581$.

Assim, utilizando a função `mean_squared_error` da biblioteca *Scikit-Learn* podemos calcular a RMSE. Ao passarmos o argumento *False* para o parâmetro *squared* dessa função, ela nos retorna a RMSE, ao invés da MSE, como diz seu nome.

Obtemos os seguintes valores de RMSE, para a variável independente GTC:

Comando:

```
mean_squared_error(y_train, y_train_preds, squared=False)
```

Resultado:

```
0.005861203460006047
```

Comando:

```
mean_squared_error(y_test, y_test_preds, squared=False)
```

Resultado:

```
0.005766994938583484
```

Já para a variável dependente GTT, temos:

Comando:

```
mean_squared_error(y_train, y_train_preds, squared=False)
```

Resultado:

```
0.002250458361614906
```

Comando:

```
mean_squared_error(y_test, y_test_preds, squared=False)
```

Resultado:

```
0.0022108340999288877
```

Note que aqui os valores são relativamente pequenos pois a faixa de valores que as variáveis dependentes possuem é bem baixa. GTC varia entre 0,95 e 1, e GTT entre 0,975 e 1.

6 Resultados

Após calcularmos os coeficientes lineares correspondentes as variáveis dependentes GTC e GTT e realizarmos as previsão para os dados de teste, obtemos os seguintes gráficos. Aqui mostramos tanto para os dados de treino como para os dados de teste, para comparação.

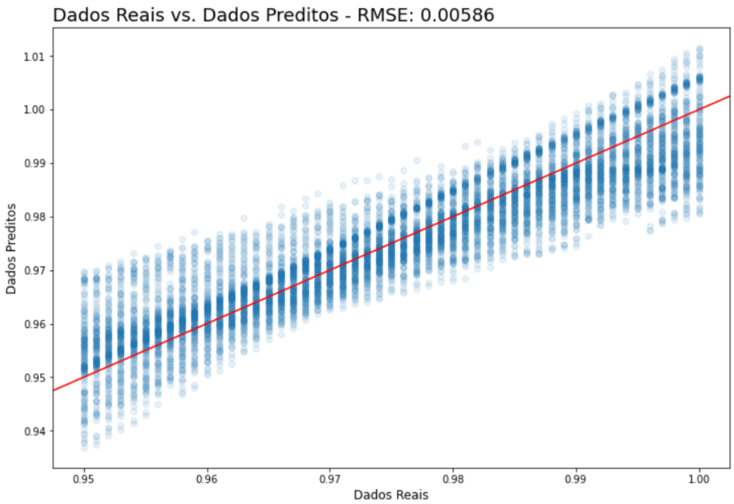

Figura 1: Resultado da predição de treino para a variável dependente GTC.

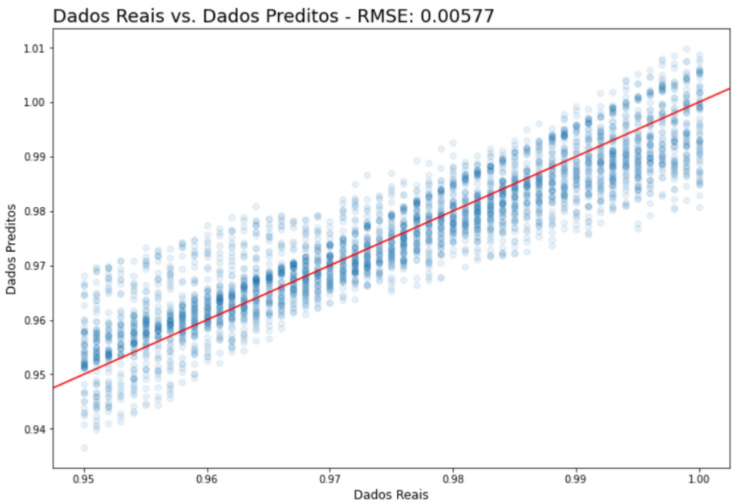

Figura 2: Resultado da predição de teste para a variável dependente GTC.

Para a segunda variável dependente GTT:

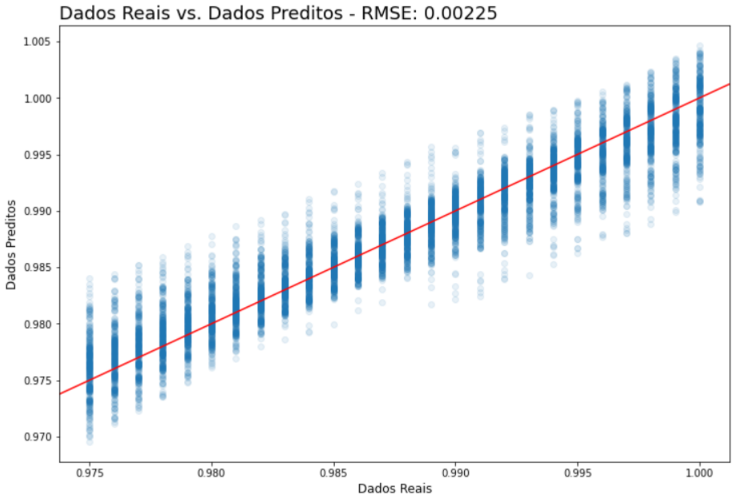

Figura 3: Resultado da predição de treino para a variável dependente GTT.

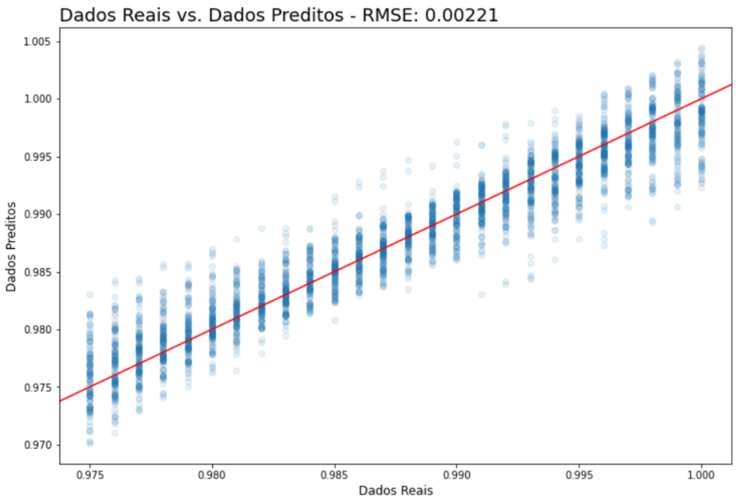

Figura 4: Resultado da predição de teste para a variável dependente GTT.

7 Conclusão

A solução de uma regressão linear com o método dos mínimos quadrados é uma tarefa utilizada em diversas aplicações, com importância para os mais variados nichos do conhecimento, aliando simplicidade e robustez em sua solução.

Já para os experimentos realizados para o conjunto de dados sobre a indústria naval, concluímos que a regressão linear com o método dos mínimos quadrados resolvido com a decomposição QR nos dá uma estimativa e previsão relativamente boas o suficiente das duas variáveis dependentes utilizadas.

8 Trabalhos Futuros

Futuros trabalhos podem se aprofundar ainda mais no método de regressão linear para o problema em questão da indústria naval, buscando qual combinação de variáveis diminui a métrica RMSE em teste. Por exemplo, se usarmos apenas 5 das 16 variáveis, o erro irá diminuir ou aumentar? Quais variáveis mais influenciam e se correlacionam com o valor final de GTC e GTT?

9 Referências Bibliográficas

[1] Stephen M Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press, 1986, pp. 55–61. ISBN: 9780674403406. URL: `https://archive.org/details/historyofstatist00stig`.

[2] Godela Scherer Per Christian Hansen Víctor Pereyra. *Least Squares Data Fitting with Applications*. JHU Press, 2013. ISBN: 9781421407869.

[3] Robert G. Easterling David M. Haaland. "Application of New Least-Squares Methods for the Quantitative Infrared Analysis of Multicomponent Samples". Em: *Applied Spectroscopy* 36 (6 1982), pp. 665–673.

[4] C. A. Cantrell. "Technical Note: Review of methods for linear least-squares fitting of data and application to atmospheric chemistry problems". Em: *Atmospheric Chemestry and Physics* 8 (2008), pp. 5477–5487.

[5] Andrea Coraddu et al. "Machine Learning Approaches for Improving Condition Based Maintenance of Naval Propulsion Plants". Em: *Journal of Engineering for the Maritime Environment* (2014).

[6] *UCI Machine Learning Repository - Condition Based Maintenance of Naval Propulsion Plants Data Set*. URL: `http://archive.ics.uci.edu/ml/datasets/condition+based+maintenance+of+naval+propulsion+plants` (acedido em 05/01/2022).

[7] Thelmo de Araujo. *Álgebra Linear: Teoria e Aplicações*. Colução Textos Universitários. SBM, 2014. ISBN: 9788583370253.

[8] Livro Colaborativo. *REAMAT - Álgebra Linear*. 2020. URL: `https://www.ufrgs.br/reamat/AlgebraLinear/index.html` (acedido em 05/01/2022).

[9] Padraic Bartlett. *Lecture 4: Applications of Orthogonality: QR Decompositions*. 2014. URL: `http://web.math.ucsb.edu/~padraic/ucsb_2013_14/math108b_w2014/math108b_w2014_lecture4.pdf` (acedido em 05/01/2022).