Uso de regressão em Machine Learning

Joab da Silva Rocha[1] — Rebeca da Silva Sousa[2]

19 de novembro de 2022

[1] Universidade Federal do Ceará, Fortaleza, Ceará, Brasil - Matemática Industrial
[2] Universidade Federal do Ceará, Fortaleza, Ceará, Brasil - Estatística

Resumo

Dentre todas as habilidades desenvolvidas pelo homem, a possibilidade de prever eventos futuros é, de longe, a mais desejada nos dias atuais. Na computação, alcançamos esse efeito por meio de algoritmos de Machine Learning, que utilizam técnicas estatísticas de Regressão para realizar classificações e previsões com base em dados de entrada que, rotulados ou não, permitirão ao algoritmo produzir uma estimativa sobre os padrões observados nessa entrada.

O objetivo de estudo deste trabalho visa discorrer a respeito do uso de regressão em algoritmos de Machine Learning. Para isso, falaremos de uma forma bem simples e resumida sobre alguns dos modelos de regressão, a saber: Modelos de Regressão Linear Simples e Múltiplo; Modelo de Regressão Polinomial; Modelo de Regressão Logística. Em seguida, introduziremos o assunto de Machine Learning e mostraremos de uma forma prática como a regressão é útil e indispensável na computação preditiva.

Palavras-chave: Regressão. Modelos de Regressão. Modelo. Machine Learning. Aprendizado de Máquinas. Análise de Regressão.

---

1 Introdução

Muito se tem falado a respeito de Machine Learning [7] e do poder que esse método tem para prever eventos futuros, mas poucos buscam realmente entender o que está por trás da capacidade da máquina de aprender, classificar e prever. O aprendizado de Máquinas surgiu para resolver problemas de diversas áreas onde apenas a computação tradicional não era o suficiente e, a cada dia que se passa, temos acompanhado o avanço dessa ferramenta, sendo inserida cada vez mais naquilo que antes era dito impossível para uma máquina. Entretanto, ao mesmo tempo em que muitos procuram utilizar todo o potencial dessa incrível ferramenta de aprendizagem, muitos também ignoram toda a base na qual o Machine Learning é fundamentado. Para que a computação chegasse ao ponto de prever e aprender foram necessários conhecimentos de uma área da Estatística que conhecemos por *Modelos de Regressão*. Com isso em mente, falaremos brevemente sobre alguns Modelos de Regressão e o conceito de Machine Learning e mostraremos no decorrer deste trabalho a maneira pela qual ambos se relacionam e como a Regressão é de fundamental importância para o Aprendizado de Máquinas (ou *Aprendizado Estatístico*). Este artigo está organizado da seguinte forma: A introdução é explicada na seção I. A teoria dos Modelos de Regressão seguirá na seção II. Na III seção falaremos de alguns tipos de Aprendizado de Máquinas. A seção IV apresentará um exemplo prático do uso da Regressão em Machine Learning juntamente do seu Diagnóstico[10]. Na seção V descreveremos nossas conclusões e as limitações que observamos do nosso trabalho.

1.1 Regressão

Quando falamos de Regressão dentro dos conceitos da Estatística, estamos nos referindo a uma poderosa ferramenta que tem por funcionalidades principais estimar e modelar a relação entre variáveis dependentes e independentes. Em outras palavras, o que está sendo feito é um ajuste de uma função em relação aos dados de uma amostra sob alguma função de erro, de modo que essa função possa explicar, o mais próximo do real quanto possível, o comportamento desses dados da amostra. Dessa forma, com o uso da Regressão, é possível determinar a correlação entre as variáveis dependentes e independentes do modelo e ajustar uma função a um conjunto de dados disponíveis, e com isso estimar e/ou prever, com alguma precisão, os valores da variável resposta para determinados valores de pontos em relação às variáveis explicativas.

Quando ajustamos uma função de Regressão, estamos interessados em dois propósitos principais. Chamamos de *Interpolação* quando podemos estimar os dados que faltam dentro do nosso intervalo de dados e *Extrapolação* quando buscamos estimar dados fora do nosso intervalo.

1.2 Machine Learning

De forma simplificada, quando falamos de machine Learning[9, 12], estamos falando da capacidade do computador de "aprender" alguma técnica ou processo fundamentalmente por meio da "interpretação" de um conjunto de dado. Enquanto a inteligência artificial (IA) pode ser definida como a capacidade de dispositivos eletrônicos de funcionar de maneira que lembra o pensamento humano, o Machine Learning[2] pode ser entendido como uma vertente específica da IA, que busca treinar as máquinas para que possam entender e aprender com os padrões em um conjunto de dados e dessa forma predizer eventos adversos. Ou seja, um modelo de Machine Learning *percebe* e aprende os padrões existentes nos dados e cria uma função capaz de gerar as previsões.

2 Modelos de Regressão

O termo *Regressão* foi proposto inicialmente pelo antropólogo, meteorologista, matemático e estatístico inglês Francis Galton[4], por volta de 1885 quando, em seus estudos de Antropometria, analisando as estaturas de pais e filhos, observou que filhos de pais com estaturas altas em relação à média tendem a ser mais baixos que seus pais, e filhos de pais mais baixos do que à média tendem a ser mais altos que seus pais. Em outras palavras, a altura do ser humano tende a regredir à média.

A Regressão nos ajuda a modelar relações entre variáveis e predizer o valor de variáveis de resposta (ou dependentes) com base em um conjunto de variáveis explicativas (independentes ou preditoras). Nesta seção falaremos dos modelos de Regressão Linear, Polinomial e Logística.

2.1 Regressão Linear

A forma geral de um modelo de regressão linear[3, 14] é apresentada a seguir.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \varepsilon_i \tag{1}$$

onde:

- y_i, $(i = 1, ..., n)$, é a i-ésima observação da variável dependente Y.
- $x_j (j = 1, ..., p)$ são as variáveis independentes.
- β_0 e β_j são parâmetros desconhecidos,rotulados parâmetros de regressão.
- β_0 é o valor de y_i quando as observações x_{ji} são nulas $(x_{ji} = 0)$.
- ε_i são os resíduos (ou erro aleatório) do modelo, ou seja, a parte do valor de y_i que difere do valor real a qual o modelo não explica.

Existem dois tipos de Regressão Linear das quais falaremos aqui: SIMPLES e MÚLTIPLA.

2.1.1 Regressão Linear Simples

Um modelo de Regressão[13] é dito Linear Simples (MRLS) quando há somente uma variável independente (X) para realizarmos a predição. O modelo de Regressão Linear Simples segue a seguinte forma funcional:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, ..., n, \tag{2}$$

Onde, para esse caso, β_0 é o valor de y_i quando x_1 é nulo $(x_1 = 0)$ e, β_1 é a taxa de variação do valor de y_i, ou seja, o quanto y_i varia positiva ou negativamente, para cada variação de x_1. Observe que, quando a amostra não incluir ou não não fizer sentido considerar $x_i = 0$, β_0 não possui uma interpretação prática. Isso é contornado centralizando a variável explicativa.

Ao admitir o MRLS, estabelecemos as seguintes pressuposições:

- A função de regressão é linear nos parâmetros.
- x_i é fixo, $\forall i \in [1, n]$, i.e, x_i não é variável aleatória.
- Homoscedasticidade, i.e, a variância do ruído aleatório ε_i é contante e vale σ^2.
  $Var[\varepsilon_i | x_i] = \sigma^2$, $\forall i \in [1, n]$.

- Os erros aleatórios de duas observações diferentes são não-correlacionados.
  $Cov(\varepsilon_i, \varepsilon_j) = 0$, $(\forall i \neq j)$

O MRLS serve principalmente para distinguir a influência de variáveis independentes em relação ao valor de variáveis dependentes.

2.1.2 Regressão Linear Múltipla

MRLM[7, 11] é uma técnica estatística para prever o resultado de uma variável resposta usando uma série de variáveis explicativas. O objetivo do MRLM é modelar a relação linear entre as variáveis independentes x e a variável dependente y que serão analisadas. O modelo básico para MRLM é:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon, \tag{3}$$

Também podemos escrever esse modelo de forma matricial:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{4}$$

onde:

$$\hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, Y = \begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

$$X = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{bmatrix}$$

2.2 Regressão Polinomial

Por sua vez, a regressão polinomial[1] é uma técnica usada para ajustar uma equação estocástica não linear por meio de funções polinomiais de variável independente.

Desse modo, o modelo de regressão polinomial para um dado polinômio de grau k em uma variável é dado por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X^2 + ... + \beta_k X^k + \varepsilon \tag{5}$$

2.3 Regressão Logística

A Regressão Logística[5] é uma técnica estatística usada para produzir, a partir de um conjunto de dados, um modelo que possibilite a predição de valores tomados por uma variável categórica, geralmente binária, em função de uma ou mais variáveis independentes que podem ser tanto contínuas como binárias.

A seguir um modelo de Regressão Logística[1]:

$$p = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}} \tag{6}$$

Então, a partir desse modelo gerado é possível calcular ou prever a probabilidade de um evento ocorrer, dado uma observação aleatória. Podemos separar as aplicações da Regressão Logística em:

- Determinar a Probabilidade: Calcula as chances de um evento ocorrer, sendo a saída do modelo um valor entre 0 e 1.
- Classificação: Embora maioria dos modelos de Regressão Logística use função contínua, é possível modelar uma função para realizar classificações. Podemos usar para separar imagens e determinar tipos relacionados.

3 Machine Learning

Abordaremos agora um assunto um pouco mais introdutório a respeito do Machine Learning. Segundo *Aurélien Géron*[8],

> *"Aprendizado de Máquina é a ciência (e a arte) da programação de computadores para que eles possam aprender com os dados.*
>
> *Veja uma definição um pouco mais abrangente:*
>
> *Aprendizado de Máquina é o campo de estudo que dá aos computadores a habilidade de aprender sem ser explicitamente programado. - Arthur Samuel, 1959"*

Atualmente, há uma imensa quantidade de dados circulando na internet provenientes de diversas fontes, cada dado informa algo a respeito de um usuário ou cliente. Para as empresas que conseguem se aproveitar dessa grande quantidade de informações, há um diferencial grandioso, além da oportunidade de se moldar e personalizar atendimentos para melhor receber seus clientes. Há infinitos usos para os dados disponíveis.

Com o uso de Machine Learning é possível realizar diversas ações com base nesses dados, tais como classificações de clientes, ou mesmo fazer previsões dos valores para se cobrar em determinados produtos com base em características observadas.

As máquinas podem aprender por meio de várias formas diferentes, abaixo apresentaremos algumas dessas formas.

3.1 Aprendizado Supervisionado

O aprendizado supervisionado se baseia na regressão básica e classificação. Os dados de entrada são apresentados ao algoritmo juntamente com os resultados desse conjunto de dados, e assim a máquina compreende padrões e aprende. Por exemplo, um humano pode inserir um conjunto de dados que ensina o modelo a identificar um carro, e mesmo que os carros mudem de cor, marca ou tamanho, ainda existem algumas características comuns associadas, como quatro rodas, formato geral, etc.

As etapas básicas do aprendizagem supervisionado podem ser resumidas em:

- Determinar a natureza dos dados e selecionar o melhor conjunto para usar no treinamento.
- Coleta os dados de diversas fontes diferentes e limpa os dados para o treinamento.
- Escolhe o modelo de aprendizado supervisionado a ser usado com base na natureza do seu objeto, se previsão ou classificação.
- Treina o modelo escolhido ajustando a função por meio de várias iterações de dados, melhorando a precisão e a velocidade de aprendizado.
- Após treinado, o modelo pode receber um novo conjunto de dados e classificar ou prever.

3.2 Aprendizado Não-Supervisionado

Diferente do aprendizado supervisionado, no aprendizado Não-Supervisionado o modelo deve aprender por conta própria com o uso de dados não rotulados. Nesse processo, uma vez que não há rótulos ou categorização pré-definida, o algoritmo agrupa os dados de acordo com suas semelhanças e padrões observados.

Alguns algoritmos desse tipo de Aprendizado:

- Aprendizado Hierárquico(HCA)
- Agrupamento de k-médias
- Modelos de mistura gaussiana
- Agrupamento Difusos

3.3 Aprendizado por Reforço

Muitas das vezes, quando estamos procurando aprender algo novo, nós tentamos, erramos e buscamos corrigir nosso erro para não cometê-lo novamente. Quando uma criança, por exemplo, começa a engatinhar e tenta se levantar para andar pela primeira vez, provavelmente ela vai cair, e será assim algumas vezes mais, até que, com ajuda de seus pais, ela será ensinada a não errar de novo. Esse é um aprendizado que ocorre com base em experiências passadas. Nesse processo, o sistema de aprendizado é chamado de agente, e ele interage com o ambiente externo para realizar suas ações. A cada ação da máquina, o ambiente irá penalizar ou recompensar o agente, desse modo, o agente precisa analisar aquilo que fez e qual retorno ele teve do ambiente anteriormente, desse modo ele buscará uma melhor abordagem para realizar sua atividade minimizando suas penalizações e maximizando suas recompensas.

4 Demonstração

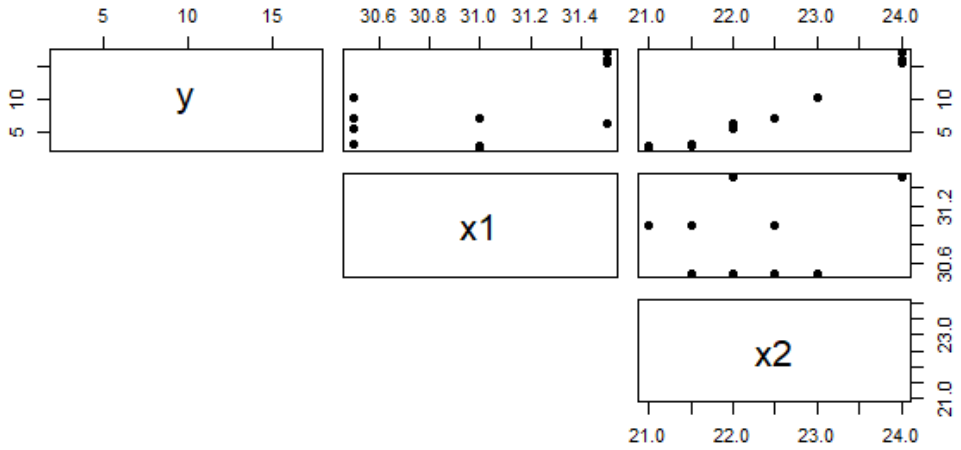
Para deixarmos mais claro o assunto tratado, traremos nessa seção uma aplicação prática para analisarmos o uso da regressão em um processo de Machine Learning. Para nos ajudar nesse processo, usaremos a base de dados de um problema encontrado no livro Introduction to Linear Regression Analysis[6].

O quadro de dados possuí 12 observações sobre carbonatação de refrigerantes, em que y é a carbonatação, x_1 representa a temperatura e x_2 a pressão.

4.0.1 Análise de regressão

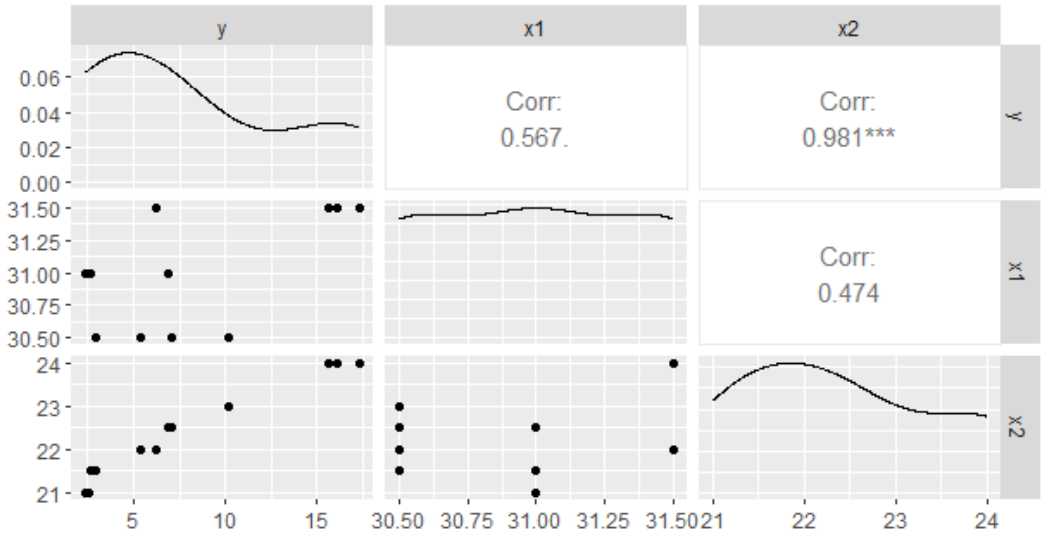
Na Figura 1 é apresentada uma matriz com gráficos de dispersão das relações entre as variáveis. è possível perceber que a relação entre o índice de carbonatação (y) e a pressão é grande, veremos na figura 2 o valor da correlação entre elas.

Figura 1: Matriz de Gráfico de dispersão



A figura a seguir nos apresenta uma comparação entre a correlação, densidade e os respectivos gráficos de dispersão entre as variáveis de interesse. Como mencionado anteriormente, a correlação entre a carbonatação e a pressão é grande e positiva.

Figura 2: Matriz de Gráfico de dispersão, correlação e densidade



Ajustando o modelo de regressão múltiplo (MRLM) obtemos as seguintes informações: as estimativas dos parâmetros, o erro padrão associado a cada estimativa, uma estatística t e um p-valor associado, resultado do teste t utilizado para saber se as estimativas são diferentes de zero e podemos observar que as variáveis incluídas no modelo foram significativas, e que sua contribuição chegou a explicar 97% da variabilidade presente nos dados ($R^2 = 0.9763$).

Figura 3: Valores do Modelo Ajustado

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4047 -0.4936  0.0860  0.5473  1.3004

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -147.4892    21.3572  -6.906 7.02e-05 ***
x1             1.7188     0.7629   2.253   0.0508 .
x2             4.5570     0.2892  15.756 7.35e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9501 on 9 degrees of freedom
Multiple R-squared:  0.9763,    Adjusted R-squared:  0.971
F-statistic:   185 on 2 and 9 DF,  p-value: 4.896e-08
```

Utilizando a função anova obtemos o quadro com os respectivos valores de graus de liberdade, soma dos quadrados, quadrados médios, estatística F e Valor-p para a estatística F.

Figura 4: Quadro ANOVA

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 109.964 109.964  121.82 1.565e-06 ***
x2         1 224.101 224.101  248.26 7.354e-08 ***
Residuals  9   8.124   0.903
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

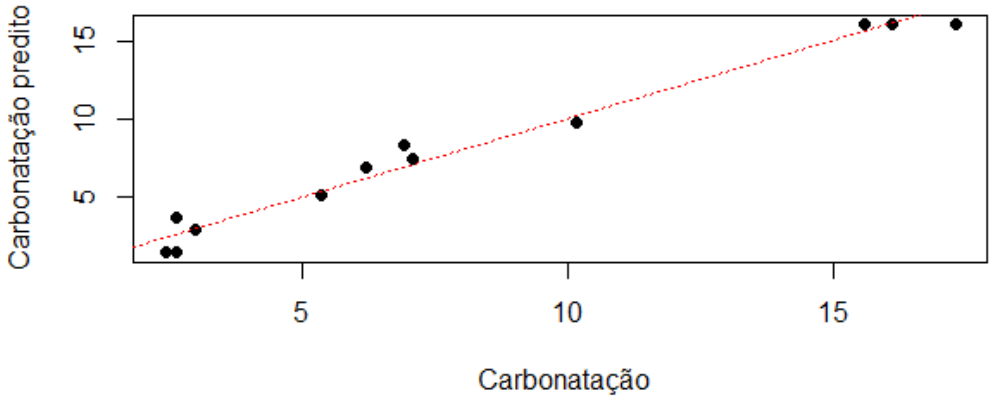
A tabela a seguir apresenta uma comparação entre os valores preditos do modelo e os reais valores observados, juntamente com o erro residual, o que já nos mostra um ajuste adequado, mas veremos na Figura 5 esses valores comparados a reta de regressão.

Tabela 1: Valores observados e preditos

Observado	Predito	Resíduo
2.60	1.489	1.111
2.40	1.489	0.911
17.32	16.020	1.300
15.60	16.020	-0.420
16.12	16.020	0.100
5.36	5.187	0.173
6.19	6.906	-0.716
10.17	9.744	0.426
2.62	3.768	-1.148
2.98	2.908	0.072
6.92	8.325	-1.405
7.06	7.465	-0.405

No gráfico 5 é possível observar que os dados estão bem ajustados a reta de regressão.

Figura 5: Gráfico de dispersão com a reta de regressão



4.0.2 Análise de Resíduos

As figuras abaixo apesentam informações se há pontos *outliers*, de alavanca, influentes e quais observações são essas.

Figura 6: Resíduos versus Ajustes

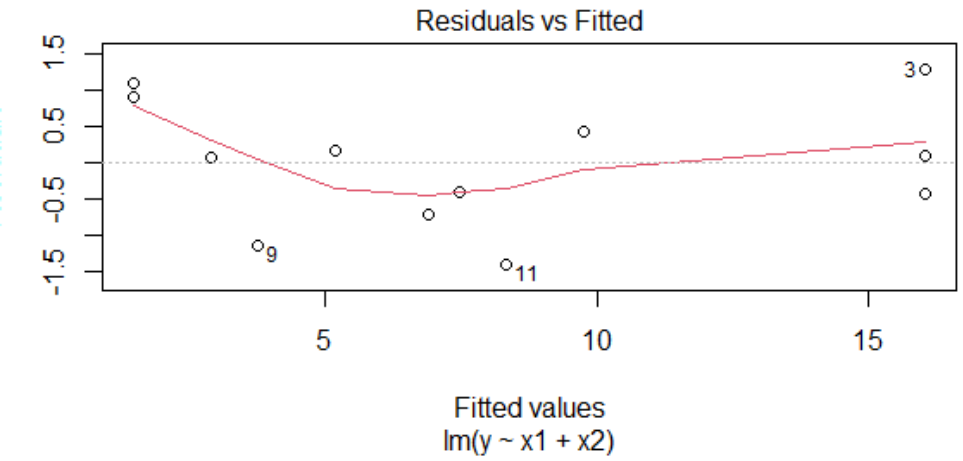


Figura 7: Quantil-Quantil normalidade

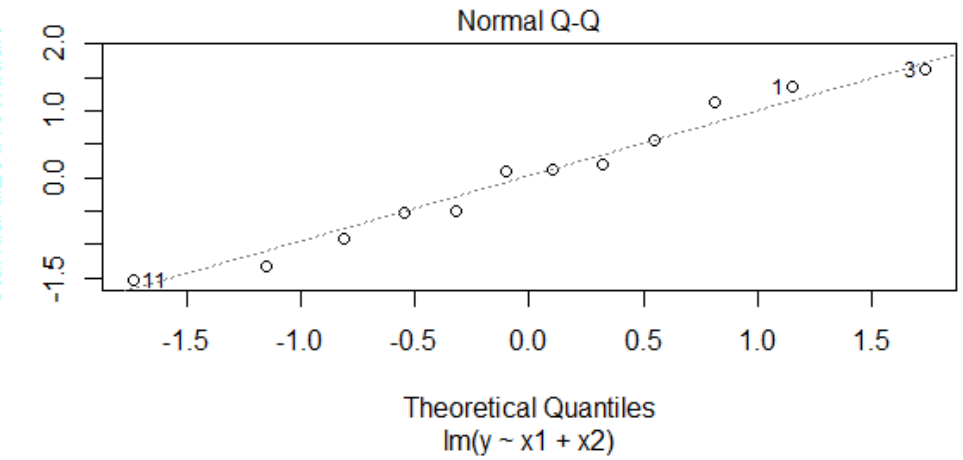


Figura 8: Gráfico de localização de propagação

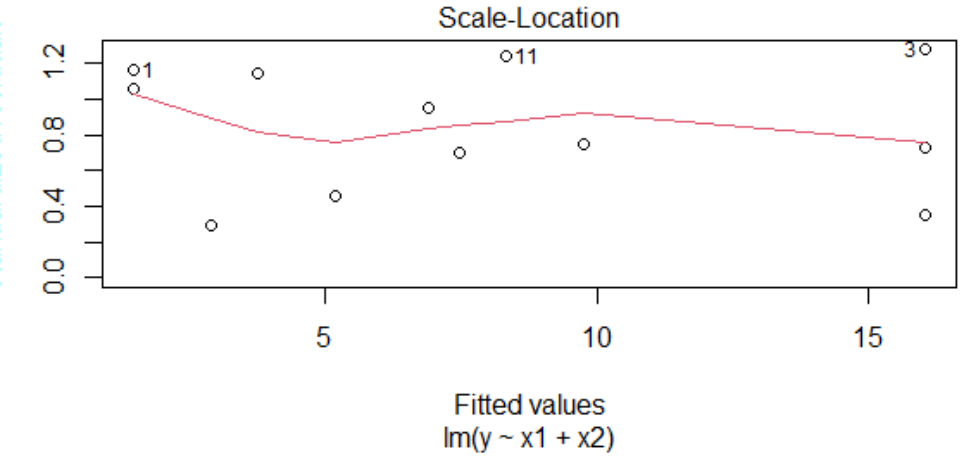
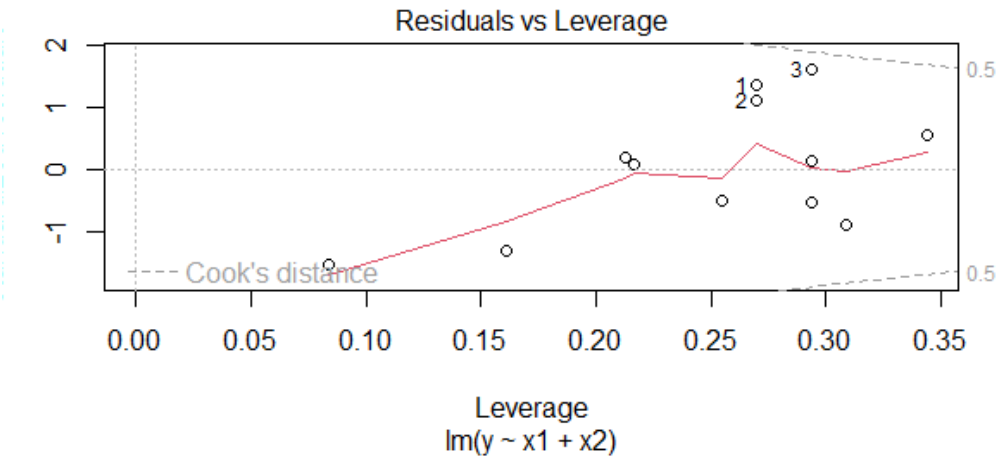


Figura 9: Gráfico de alavancagem



Após análise do modelo, podemos concluir que o mesmo está bem ajustado, o que nos proporciona uma explicação muita boa do índice de carbonatação com relação entre a temperatura e a pressão, apesar da temperatura não explicar sozinha o índice, a combinação com a pressão nos mostrou que os valores se aproximaram do valor que prevemos.

5 Conclusão

A Regressão Linear é um método estatístico bastante usado para classificação, previsão e predição. Não apenas muito procurado e requerido no meio profissional, mas também imensamente útil em observações e análises de casos para estudos, quer seja de teor acadêmico, quer seja profissional. Através da Regressão foi possível construir modelos computacionais ditos capazes de aprender e se desenvolver, realizando tarefas custosas e complexas para humanos mas rápida e simples para máquinas bem treinadas.

Dentro do amplo assunto estatístico de Regressão ainda existem vários outros tipos de modelos de Regressão que são muito usados em diferentes aplicações, cada um atendendo melhor a uma necessidade específica. Em nosso trabalho citamos apenas os modelos de regressão Linear, Polinomial e Logístico, que são modelos mais simples e comuns, visto que temos por finalidade fazer uma breve apresentação dos modelos de regressão e mostrar como eles são usados dentro do Machine Learning.

Dentre os modelos que deixamos de fora, podemos citar alguns como[1]:

- Regressão Quantílica
- Regressão Ridge
- Regressão Lasso
- Regressão ElasticNet
- Regressão de Componentes Principais
- Regressão Parcial de Mínimos Quadrados
- Regressão Vetorial de Suporte
- Regressão Ordinal
- Regressão de Poisson
- Regressão Binomail Negativa
- Regressão Quasi-Poisson
- Regressão de Cox

Vale ressaltar, também, que nosso objetivo nesse trabalho não visava aprofundar os assuntos de Regressão e Machine Learning, mas sim falar do uso da Regressão no Aprendizado de Máquinas, por tal razão, o presente trabalho não buscou explorar de forma incisiva por fórmulas, equações e menos ainda suas demonstrações, mas principalmente utilizar de alguns desses resultados por outros já demonstrados.

Referências

[1] Deepanshu Bhalla. 15 types of regression in data science. url: https://www.listendata.com/2018/03/regression-analysis.html.

[2] Ana Lívia Castro. Machine learning e inteligência artificial na previsão de acidentes de trabalho. URL: https://www.sesi-ce.org.br/blog/machine-learning-e-inteligencia-artificial-na-previsao-de-acidentes-de-trabalho/.

[3] Flávia Chein. *Introdução aos Modelos de Regressão Linear*. Enap, 2019.

[4] Laura Damaceno. Regressão linear? url: https://medium.com/@lauradamaceno/ regressão-linear-6a7f247c3e29, junho 2020.

[5] Leandro de Azevedo Gonzalez. Regressão logística e suas aplicações. 2018.

[6] Montgomery et al. *Introduction To Linear Regression Analysis*. 3rd ed.

[7] Maulud et at. A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, vol. 01(Nº 04):pp. 140–147, 2020.

[8] Aurélien Géron. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn TensorFlow*. 2017.

[9] Matt Harrison. *Authorized portuguese translation of the english edition of Machine Learning Pocket Reference*. Novatec, 2020.

[10] Francisco Marcelo M. Rocha Juvêncio S. Nobre, Julio M. Singer. *Análise de dados longitudinais: Versão parcial preliminar*. 2019.

[11] Alexandre Gori Maia. *Econometria: conceitos e aplicações*. cap. 06, 2017.

[12] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[13] Psicometria Online. O que é regressão linear simples? URL: https://psicometriaonline.com.br/o-que-e-regressao-linear-simples-2/.

[14] Blettner M Schneider A, Hommel G. Linear regression analysis. *part 14 of a series on evaluation of scientific publications*, vol. 107(Nº 44):pp.776–782, 2010.