UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
ESPECIALIZAÇÃO EM ESTATÍSTICA

DANIELE PEREIRA KAPPES

APLICAÇÃO *DIGITAL TWIN* PARA AUXÍLIO DE DETECÇÃO DE ANOMALIAS NO CONSUMO DE GÁS

BELO HORIZONTE
2021

DANIELE PEREIRA KAPPES

APLICAÇÃO *DIGITAL TWIN* PARA AUXÍLIO DE DETECÇÃO DE ANOMALIAS NO CONSUMO DE GÁS

Monografia de especialização apresentada ao Programa de Especialização em Estatística do Departamento de Estatística da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Especialista em Estatística.

Área de pesquisa: Regressão Linear

Orientadora: Professora Doutora Ilka Afonso Reis

BELO HORIZONTE
2021

Kappes, Daniele Pereira

K17a Aplicação digital Twin para auxílio de detecção de anomalias no consumo de gás [manuscrito] / Daniele Pereira Kappes. Belo Horizonte — 2021.
48f. : il. ; 29 cm corrigir

Orientadora: Ilka Afonso Reis.

Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências : f. 48

1. Estatística. 2. Análise de regressão. 3. Detecção de anomalias (Computação). 4. Gêmeo digital. 5. Gás – Consumo. 6. Energia elétrica – Consumo. I. Reis, Ilka Afonso. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. IV. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz CRB 6ª nº 819.

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 225ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE DANIELE PEREIRA KAPPES.

Aos dois dias do mês de junho de 2021, às 10:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso da aluna **Daniele Pereira Kappes**, intitulado: "*Aplicação Digital Twin para auxílio de detecção de anomalias no consumo de gás*", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Ilka Afonso Reis – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Após a defesa, os membros da banca examinadora reuniram-se sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: a candidata foi considerada Aprovada condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado foi comunicado publicamente à candidata pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 02 de junho de 2021.

Prof.ª Ilka Afonso Reis (Orientadora)
Departamento de Estatística / ICEX / UFMG

Prof.ª Lourdes Coral Contreras Montenegro
Departamento de Estatística / ICEX / UFMG

Prof. Guilherme Lopes de Oliveira
Departamento de Computação / CEFET / MG

AGRADECIMENTOS

Agradeço, primeiramente à Deus, que me deu energia e benefícios para concluir todo este trabalho.

Agradeço aos meus pais que me incentivaram todos os momentos da minha vida.

À minha orientadora, Professora Doutora Ilka Afonso Reis, pelo suporte no pouco tempo que lhe coube, pelas suas correções e paciência.

Enfim, agradeço a todas as pessoas que de alguma forma fizeram parte desta etapa decisiva em minha vida.

“A lei da mente é implacável. O que você pensa, você cria; O que você sente, você atrai; O que você acredita torna-se realidade.”.

Buda

RESUMO

Empresas distribuidoras de gás ou energia devem sempre se atentarem aos medidores (de gás ou de energia) para garantirem que estão cobrando pelo produto de forma correta, ou seja, que estão recebendo pela quantidade correta de produto disponilizada ao cliente. Eventualmente, pode ocorrer alguns problemas nos medidores e essa leitura pode ser feita de forma incorreta, tanto por conta de ações de má fé por conta do cliente, ou até mesmo por mau funcionamento do medidor. Neste trabalho é apresentado o desenvolvimento de uma ferramenta de auxílio à detecção de anomalia na medição de consumo de gás para um cliente específico de uma empresa distribuidora de gás. No sistema desenvolvido, a detecção de anomalia é feita utilizando regressão linear para criação de um *digital twin*. Embora o ajuste do modelo de regressão linear final tenha apresentado problema de violação das suposições de autocorrelação e não normalidade dos erros, ainda é possível ser melhorado para o intuito de detecção de anomalia.

PALAVRAS-CHAVE: Regressão Linear, Gêmeo Digital, Detecção de anomalia, *Digital Twin*.

ABSTRACT

Gas or energy distribution companies must always pay attention to the gas or energy meters in order to ensure that they are charging for the product correctly, that is, that they are receiving for the correct amount of product made available to the customer. Eventually, there may be some problems with the devices and this measuring may be done incorrectly, either due to bad faith actions on behalf of the customer, or even due to malfunction of the meter. This work presents the development of a tool to assist anomaly detection in the measurement of gas consumption for some specific customers of a gas company. In the developed system, anomaly detection is done using linear regression to create a *digital twin*. Although the final linear regression model presented the problem of autocorrelation and non-normality in the residuals, it is still possible to be improved in order to detect anomalies.

SUMÁRIO

Sumário 9

Lista de ilustrações 11

Lista de tabelas 13

1 INTRODUÇÃO 14
1.1 Motivação e Justificativa 14
1.2 O problema de detecção de anomalia 14
1.2.1 Contexto do problema 15
1.2.2 O que é um *digital twin* 15
1.3 Objetivos do Projeto 15
1.4 Estrutura da Monografia 15

2 MATERIAIS E MÉTODOS 17
2.1 Regressão linear 17
2.1.1 As fontes de variabilidade de Y e o teste F da tabela de Análise 18
2.1.2 Análise de Resíduos 20
2.1.3 Multicolinearidade e *Variance Inflation Factor* (VIF) 20
2.1.4 Transformação de Box-cox 21
2.2 Descrição dos dados disponíveis 21
2.3 Análise estatística dos dados 22
2.3.1 Análise exploratória 22
2.3.2 Correlações 22
2.4 Separação dos dados 22
2.5 Construção do *Digital Twin* 23
2.6 Ambiente de realização das análises 23

3 RESULTADOS E DISCUSSÃO 24
3.1 Análise Estatística dos Dados 24
3.1.1 Análise Exploratória 24
3.1.2 Correlações 27
3.2 Ajuste da Regressão Linear 30
3.2.1 Modelo 1 30
3.2.1.1 Análise de colinearidade entre variáveis explicativas 31
3.2.2 Modelo 2 31
3.2.2.1 Análise de resíduos 32

3.2.3 Transformação de box-cox 33
3.2.4 Correlações após transformação da variável alvo 33
3.2.5 Modelo 3 36
3.2.5.1 Análise de resíduos 36
3.2.6 Modelo 4 38
3.2.6.1 Análise de resíduos 39
3.2.7 Modelo 5 40
3.2.7.1 Análise de multicolinearidade entre variáveis explicativas 41
3.2.7.2 Análise de resíduos 41
3.3 Construção do *Digital Twin* **43**

4 CONCLUSÃO E TRABALHOS FUTUROS **46**

Referências **48**

LISTA DE ILUSTRAÇÕES

Figura 1 – Representação de um ajuste de modelo de regressão linear a um conjunto de dados 17
Figura 2 – Fontes de variabilidade de Y - Fonte: [4] 19
Figura 3 – Transformação de Box-Cox - Fonte: [5] 21
Figura 4 – Separação dos dados 23
Figura 5 – Dados de volume 24
Figura 6 – Dados de pressão 25
Figura 7 – Dados de temperatura 25
Figura 8 – Box plot das variáveis 26
Figura 9 – Distribuição de frequência dos dados coletados das variáveis Volume, Pressão e Temperatura 27
Figura 10 – Gráfico de dispersão para pressão e volume e coeficientes de correlação de Pearson e de Spearman entre pressão e volume 28
Figura 11 – Gráfico de dispersão para temperatura e volume e coeficientes de correlação de Pearson e de Spearman entre temperatura e volume 28
Figura 12 – Gráfico de dispersão para $\frac{temperatura}{pressão}$ e volume e coeficientes de correlação de Pearson e de Spearman entre $\frac{temperatura}{pressão}$ e volume 29
Figura 13 – Gráfico de dispersão para volume passado e volume e coeficientes de correlação de Pearson e de Spearman entre volume passado e volume 29
Figura 14 – Resultados do ajuste da regressão linear usando temperatura e pressão (variável alvo: volume) 30
Figura 15 – Ajuste da regressão linear usando $\frac{temperatura}{pressão}$ (variável alvo: volume) . . . 31
Figura 16 – Análise de resíduos - Modelo 2 32
Figura 17 – Teste de normalidade de Shapiro-Wilk - Modelo 2 32
Figura 18 – Histograma do volume de gás antes e depois da transformação de box-cox (raiz quadrada) 33
Figura 19 – Gráfico de dispersão para pressão e raiz quadrada do volume e coeficientes de correlação de Pearson e de Spearman entre pressão e raiz quadrada do volume 34
Figura 20 – Gráfico de dispersão para temperatura e raiz quadrada do volume e coeficientes de correlação de Pearson e de Spearman entre temperatura e raiz quadrada do volume 34
Figura 21 – Gráfico de dispersão para $\frac{temperatura}{pressão}$ e raiz quadrada do volume e coeficientes de correlação de Pearson e de Spearman entre $\frac{temperatura}{pressão}$ e raiz quadrada do volume 35

Figura 22 – Gráfico de dispersão para raiz quadrada do volume passado e raiz quadrada do volume e coeficientes de correlação de Pearson e de Spearman entre raiz quadrada do volume passado e raiz quadrada do volume 35
Figura 23 – Ajuste da regressão linear usando $\frac{temperatura}{pressão}$ (variável alvo: raiz quadrada do volume) 36
Figura 24 – Análise de resíduos - Modelo 3 37
Figura 25 – Teste de normalidade de Shapiro-Wilk - Modelo 3 37
Figura 26 – Teste de Durbin-Watson de autocorrelação dos resíduos - Modelo 3 37
Figura 27 – Ajuste da regressão linear usando temperatura e pressão (variável alvo: raiz quadrada do volume) 38
Figura 28 – Análise de resíduos - Modelo 4 39
Figura 29 – Teste de normalidade de Shapiro-Wilk - Modelo 4 39
Figura 30 – Teste de Durbin-Watson de autocorrelação dos resíduos - Modelo 4 40
Figura 31 – Ajuste da regressão linear usando $\frac{temperatura}{pressão}$ e raiz quadrada do volume passado (variável alvo: raiz quadrada do volume) 41
Figura 32 – Análise de resíduos - Modelo 5 42
Figura 33 – Teste de normalidade de Shapiro-Wilk - Modelo 5 42
Figura 34 – Teste de Durbin-Watson de autocorrelação dos resíduos - Modelo 5 42
Figura 35 – Valores preditos, reais e anomalias - Conjunto de validação 44
Figura 36 – Gráfico de dispersão para valores preditos e valores reais no conjunto de validação - Pontos marcados com X indicam os valores considerados anomalias 44

LISTA DE TABELAS

Tabela 1 – Tabela de Análise de Variância - Fonte: [4] 19
Tabela 2 – Estatística descritiva das variáveis 24

1 INTRODUÇÃO

1.1 Motivação e Justificativa

Empresas distribuidoras de gás ou energia devem sempre se atentarem aos medidores (de gás ou de energia) para garantirem que estão cobrando pelo produto de forma correta, ou seja, que estão recebendo pela quantidade correta de produto disponibilizada ao cliente. Eventualmente, podem ocorrer alguns problemas nos medidores e essa leitura pode ser feita de forma incorreta, tanto por conta de ações de má fé por conta do cliente, ou até mesmo por mau funcionamento do medidor.

Existem algumas técnicas que usam análise de dados e ciência de dados para detectar fraudes. Dentre elas existem algumas técnicas em que o comportamento da fraude é conhecido, ou seja, existem rótulos bem definidos do que é e do que não é uma fraude. Como exemplo podemos citar alguns algoritmos de classificação (Regressão Logística, Naive Bayes, entre outros). Existem ainda outras técnicas em que o comportamento da fraude não é conhecido, como é o caso da técnica usada neste trabalho. Esta técnica se chama *digital twin* e consiste na modelagem do comportamento dos valores da variável alvo para qual algum comportamento anômalo possa ser identificado.

O objetivo principal é construir um modelo de regressão linear para criação de um *digital twin* que possa auxíliar a detecção de anomalias nas medições de consumo de gás em um dos clientes em uma empresa de distribuição de gás. Trata-se de um cliente do ramo de gás natural veicular (GNV), possuindo consumo médio próximo de 200m³ (não sendo um dos grandes consumidores da distribuidora). Segundo a distribuidora de gás, grandes consumidores são clientes em que é menos provável que ocorram anomalias, não sendo necessário sistema de detecção de anomalias para estes.

1.2 O problema de detecção de anomalia

A detecção de anomalias visa identificar momentos em um conjunto de dados em que tenha ocorrido alguma exceção, ruído ou desvio. O principal desafio em problemas desse tipo é quando a anomalia não está bem definida, ou seja, não existem rótulos especificando o que é considerado um comportamento anômalo na variável alvo, como é o caso do problema apresentado nessa monografia. Para contornar esse problema é possível usar o conceito de *digital twin*, que visa modelar o comportamento dos valores da variável alvo, para a qual algum comportamento anômalo possa ser identificado.

1.2.1 Contexto do problema

O problema consiste em identificar possíveis fraudes (anomalias) nas medições de consumo de volume de gás em uma distribuidora desse produto. Essa distribuidora possui vários clientes e é necessário desenvolver uma técnica de detecção de anomalia para cada um. Neste trabalho será desenvolvido a técnica apenas para um cliente. Para isso, será utilizado o conceito de *digital twin* mencionado anteriormente, usando conceitos de estatística para desenvolvimento de um modelo de regressão linear. Além disso, os resíduos do modelo ajustado aos dados serão usados para a criação de um limiar aceitável de diferença entre valor medido (de volume) e valor inferido pelo modelo.

1.2.2 O que é um *digital twin*

Segundo [1], *digital twin* é uma representação virtual dinâmica de um objeto físico ou sistema em todo o seu ciclo de vida, usando dados em tempo real para permitir a compreensão, aprendizagem e raciocínio. No caso deste trabalho, será aplicada a técnica de Regressão Linear para representar de forma virtualmente o valor de volume consumido por cada cliente. Posteriormente, essa representação virtual será usada para tentar identificar possíveis falhas na medição.

1.3 Objetivos do Projeto

Tendo em vista o exposto acima, este projeto tem por objetivos:

a. Construir um modelo de regressão linear que seja capaz de representar o comportamento dos valores da variável alvo: volume de gás;

b. Determinar o limiar aceitável de erro entre valores reais e valores preditos pela regressão linear considerado como não anomalia;

c. Auxiliar a empresa distribuidora de gás na detecção de anomalia em um cliente específico através do *digital twin* desenvolvido;

d. Empregar os conhecimentos adquiridos durante o Curso de Especialização em Estatística, demonstrando e buscando conhecimento dos assuntos estudados durante o curso de especialização;

e. Demonstração de capacidade técnica e científica, e também de caráter ético e cidadão.

1.4 Estrutura da Monografia

O trabalho está dividido em quatro capítulos. Este capítulo apresentou uma introdução ao projeto a ser descrito nesta monografia. O Capítulo 2 apresenta os materiais e o método

utilizado neste trabalho. O Capítulo 3 apresenta os resultados obtidos. No Capítulo 4, são apresentadas as conclusões do trabalho e as considerações finais.

2 MATERIAIS E MÉTODOS

2.1 Regressão linear

Segundo [2], a regressão linear simples é uma maneira bem direta de avaliar quantativamente a resposta de uma variável Y em função de uma variável X. Essa relação pode ser escrita da forma:

$$Y = \theta_0 + \theta_1 x + e$$

na qual, θ_0 e θ_1 são duas constantes desconhecidas que representam a inclinação e a interceptação de um modelo linear, respectivamente, e e é o termo de erro não-observável que está associado à tudo que o modelo simples não foi capaz de captar. Juntos, θ_0, θ_1 e e representam os coeficientes ou parâmetros do modelo. É possível estimar o preço de casas se tivermos uma relação linear dada por $\hat{Y} = \theta_0 + \theta_1 x$ entre o tamanho da casa e o preço por exemplo.

Assume-se que e é um fator aleatório para o qual se atribui uma distribuição de probabilidade apropriada. Ainda segundo [2], os parâmetros a e b são desconhecidos, portanto é necessário a sua estimação. Para isso, é necessário reunir conjunto de dados (x_1, y_1), (x_2, y_2),..., (x_n, y_n), que consistem em medidas de x e de y na mesma unidade de análise, sendo n o tamanho da amostra.

Na Figura 1 a seguir temos a representação gráfica de um conjunto de oito observações (quadrados em vermelho) e a reta que melhor descreveria a relação entre a variável Y (eixo vertical) e a variável x (eixo horizontal). Na regressão linear são estimados parâmetros (θ_0 e θ_1) de uma reta que melhor se ajusta aos dados. Cada ponto da reta azul é uma predição de Y baseado no valor de X. A reta é estimada minimizando a soma dos quadrados do resíduo (Residual sum of squares - RSS). O resíduo é a diferença entre o valor estimado e o valor real.

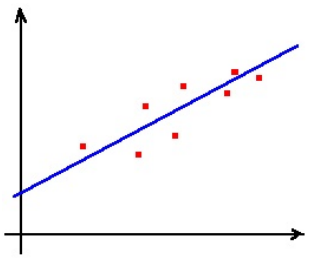


Figura 1 – Representação de um ajuste de modelo de regressão linear a um conjunto de dados

O modelo da regressão linear múltipla se assemelha ao modelo de regressão linear com a diferença que, no modelo de regressão linear múltipla, a variável resposta Y é estimada com base em mais de uma variável explicativa $(x_1, x_2, ..., x_k)$:

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_k X_k + e$$

Onde:

Y é um vetor com n observações da variável resposta, uma para cada elemento da amostra;

$X_1, X_2, ..., X_k$ são, respectivamente, os vetores com n observações das k variáveis explicativas;

$\theta_0, \theta_1, ..., \theta_k$ são os coeficientes do modelo;

e é um vetor com n termos de erro.

Assumindo

$\mathrm{X} = \begin{bmatrix} 1 & X_1 & ... & X_k \end{bmatrix}$

$\theta = \begin{bmatrix} \theta_0 & \theta_1 & ... & \theta_k \end{bmatrix}$

Onde X é uma matriz com (k+1) colunas e θ é um vetor com (k+1) elementos, o modelo de regressão pode ser escrito na forma matricial da seguinte maneira:

$$Y = X\theta^T + e$$

A estimação dos parâmetros do modelo pode ser feita por meio da minimização da soma dos quadrados dos erros (SQE), que é dada por:

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

Para estimar os parâmetros do modelo de regressão linear basta derivar a SQE em relação aos parâmetros da equação de regressão, igualar as expressões a zero para criar as equações normais e então resolver essas equações.

2.1.1 As fontes de variabilidade de Y e o teste F da tabela de Análise

Segundo [3], a análise de variância é um método utilizado para testar a significância da regressão. Esse método trabalha com a ideia de que a variabilidade total da variável resposta Y é o resultado de duas fontes de variação, como mostra a Figura 2 a seguir:

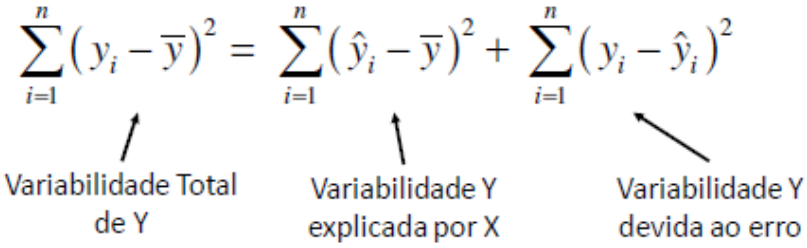


Figura 2 – Fontes de variabilidade de Y - Fonte: [4]

Ainda segundo [3], podemos chamar $SQ_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ de soma do quadrado dos erros e $SQ_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ de soma dos quadrados da regressão. Simbolicamente, a equação mostrada na Figura 2 pode ser escrita da seguinte forma:

$$SQ_T = SQ_R + SQ_E$$

Segundo [3], a análise de variância é uma forma de resumir um modelo de regressão linear através da decomposição da soma dos quadrados para cada fonte de variação no modelo e, utilizando o teste F, testar a hipótese de que os coeficientes de inclinação do modelo são iguais a zero.

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrados Médios	F_0
Regressão	SQ_R	1	QM_R	QM_R/QM_R
Resíduos	SQ_R	n-2	QM_R	
Total	SQ_T	n-1		

Tabela 1 – Tabela de Análise de Variância - Fonte: [4]

Segundo [3], quando fala-se de regressão linear, um teste F geralmente compara os ajustes de diferentes modelos lineares. O teste F da significância global é uma forma específica do teste F. Ele compara um modelo sem preditores com o modelo especificado ajustado. Um modelo de regressão que não contém preditores também é conhecido como um modelo somente com o intercepto.

As hipóteses nula e alternativa do teste F global da tabela ANOVA são:

H0: todos os coeficientes das variáveis explicativas são nulos.

H1: ao menos um dos coeficientes das variáveis explicativas é não-nulo

Ainda segundo [3], se o valor-p para o teste F de teste de significância global for menor que o nível de significância especificado, rejeita-se a hipótese nula.

2.1.2 Análise de Resíduos

Segundo [5], os resíduos de um modelo de regressão linear podem ser usados para investigação de adequação do modelo e das seguintes suposições:

a. O modelo é linear;

b. A variância dos erros é constante, ou seja, os erros são homocedásticos;

c. A forma linear adotada pelas variáveis explicativas no modelo está correta;

d. Os erros do modelo de regressão linear seguem o modelo Normal;

e. Os erros do modelo não são autocorrelacionados.

Portanto, a análise dos resíduos é importante para identificar se o modelo está bem ajustado. Essa análise pode ser feita de forma visual, por meio de gráficos e por meio de testes de hipóteses.

Os gráficos e testes de hipóteses que podem ser utilizados para essa análise são os seguintes:

a. Gráfico de resíduos *versus* preditos pelo modelo: é importante para avaliar a suposição de linearidade de modelo por meio da avaliação da independência entre resíduos e preditos;

b. Gráfico de resíduos *versus* variáveis no modelo: é importante para avaliar a suposição de variância constante dos erros ao longo da reta e para avaliar também se a forma linear adotada para estas variáveis no modelo está correta;

c. Gráfico de Probabilidade Normal dos resíduos: importante para avaliar se os erros apresentam normalidade;

d. Teste de Shapiro-Wilk para normalidade. Nesse teste a hipótese nula é a de que os erros do modelo de regressão seguem o modelo Normal;

e. Teste de Durbin-Watson, que é importante para verificação de autocorrelação dos erros. Nesse teste a hipótese nula é a de que os erros do modelo não são autocorrelacionados.

2.1.3 Multicolinearidade e *Variance Inflation Factor* (VIF)

Segundo [3], em problemas de regressão linear múltipla, é esperado encontrar dependências entre a variável resposta Y e os regressores x_k. No entanto, pode ser encontrado também dependências entre as variáveis regressoras x_k. Em situações onde essas dependências são fortes, é dito que existe multicolinearidade.

Ainda segundo [3], um fator importante para a medida da extensão da presença de multicolinearidade é o *variance inflation factor*, que pode ser definido como:

$$VIF = \frac{1}{1 - {R^2}_k}, k = 1, 2, ..., K$$

Sendo ${R^2}_k$ o coeficiente de determinação múltipla, resultante da regressão de x_k nas outras K - 1 variáveis regressoras.

É possível perceber então que quanto mais forte for a dependência linear de x_k nos regressores restantes, mais forte será a multicolinearidade.

Quanto maior o *variance inflation factor* mais severa será a multicolinearidade. Ainda segundo [3], alguns autores sugerem que se qualquer fator de inflação exceder 10, a multicolinearidade será um problema.

2.1.4 Transformação de Box-cox

Quando a suposição de normalidade dos erros do modelo de regressão linear não pode ser feita, é possível transformar a variável resposta na tentativa de melhorar o ajuste do modelo. Segundo [5], a transformação a ser usada pode ser encontrada por meio da técnica conhecida como transformações de Box-Cox.

Segundo [5], a ideia por trás da transformação de Box-Cox é achar o valor de λ que faça com que a variância da parte de Y que não é explicada pelo modelo de regressão seja a menor possível:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Figura 3 – Transformação de Box-Cox - Fonte: [5]

2.2 Descrição dos dados disponíveis

Os dados medidos para cada cliente consumidor de gás são:

a. Volume em m^3 (variável alvo);

b. Pressão em bar;

c. Temperatura em $°C$.

Para este trabalho, foi selecionado um cliente de forma aleatória dentro do conjunto de clientes com menor probabilidade e suspeita de fraude (segundo análises prévias da distribuidora de gás) e este cliente é do ramo de gás natural veicular (GNV). O conjunto dos

dados possuem 720 amostras com amostragem de uma hora. Esses dados são coletados e historiados pela própria distribuidora de gás, podendo ser acessados através de um banco de dados.

2.3 Análise estatística dos dados

2.3.1 Análise exploratória

A fim de entender um pouco melhor os dados coletados, foi feita inicialmente uma análise exploratória dos dados, observando gráficos de tendência das variáveis (pressão, volume e temperatura), gráficos de boxplot para representar a variação dos dados por meio de quartis, gráficos de densidade de distribuição, assim como gráficos de dispersão para avaliar correlações lineares e não lineares entre as variáveis.

2.3.2 Correlações

Foram usadas duas formas de cálculo de correlação. A primeira é utilizando o coeficiente de correlação linear de Pearson, e a segunda utilizando o coeficiente de correlação de Spearman.

A correlação de Spearman é usada para avaliar se a relação entre duas variáveis pode ser descrita pelo uso de uma função monótona. Enquanto a correlação de Pearson avalia relações lineares, a correlação de Spearman avalia relações monótonas, sejam elas lineares ou não. Segundo [6], uma função monótona em um conjunto S é uma função que pode ser crescente, decrescente, monótona não-decrescente ou monótona não-crescente neste conjunto.

Assim como o coeficiente de correlação de Pearson, o coeficiente de correlação de Spearman varia entre -1 a 1. O valor de -1 significa uma correlação negativa perfeita entre as duas variáveis, o valor 1 significa uma correlação positiva perfeita entre as duas variáveis e valor 0 significa que não há correlação entre as duas variáveis.

2.4 Separação dos dados

Os dados foram separados em dois conjuntos: conjunto de ajuste e conjunto de validação. Essa separação será feita de forma que os primeiros 70% dos dados sejam utilizados para o ajuste da regressão linear e os últimos 30% dos dados sejam usados para a validação do modelo de regressão. Esse conjunto de dados de validação é importante para simular o funcionamento do *digital twin* de forma *online*. Portanto, foi estabelecido uma data como separação entre os dois conjuntos de dados.

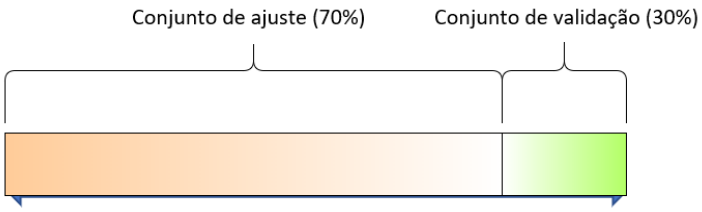


Figura 4 – Separação dos dados

Com essa separação, 504 amostras passaram a compor o conjunto de ajuste e 216 amostras passaram a compor o conjunto de validação.

2.5 Construção do *Digital Twin*

Assumindo que o modelo esteja bem ajustado, para a construção do *digital twin* é necessário estabelecer um limiar de erro entre valores preditos e valores reais que seja aceitável. Momentos em que o valor predito pelo modelo é muito maior que o valor real podem indicar uma fraude (ou anomalia). Portanto, esse limiar foi definido usando os valores de média e de desvio-padrão do erro absoluto encontrado no conjunto de validação:

$$abs(y\_hat - y) > mae\_val + (2 * std\_mae\_val)$$

Sendo:
y_hat = valor predito
y = valor real
mae_val = média do erro absoluto do conjunto de validação
std_mae_val = desvio-padrão do erro absoluto do conjunto de validação

Essa equação é definida pela área da aplicação, sendo usada em problemas de detecção de anomalia. Um ponto importante é o valor 2 que multiplica o valor do desvio-padrão, esse valor pode ser maior ou menor de acordo com a quantidade de falsos positivos ou falsos negativos gerados pelo sistema após a implantação de forma *online*.

A fraude é acusada se a diferença entre valor real e valor predito, em valor absoluto, for maior que o valor médio do erro absoluto do conjunto de validação mais duas vezes o desvio-padrão do erro médio absoluto do conjunto de validação.

2.6 Ambiente de realização das análises

Para análise dos dados foram utilizadas as linguagens de programação *python* (versão 3.7.3) e R (versão 3.6.3) com o auxílio do *software* RStudio e da aplicação *web Jupyter Notebook*.

3 RESULTADOS E DISCUSSÃO

3.1 Análise Estatística dos Dados

3.1.1 Análise Exploratória

Foi feita uma análise descritiva dos dados inicialmente e na Tabela 2 a seguir temos os valores que descrevem as três variáveis disponíveis no problema:

	Volume (m^3)	Pressão (bar)	Temperatura ($°C$)
Média	179.306	7.782	26.262
Desvio-padrão	132.065	0.069	2.419
Mínimo	0.000	7.560	18.640
25%	60.000	7.725	24.695
50%	178.000	7.782	26.795
75%	279.500	7.843	27.922
Máximo	585.000	7.924	32.260

Tabela 2 – Estatística descritiva das variáveis

A seguir, as Figuras 5, 6 e 7 mostram os valores dessas variáveis no tempo:

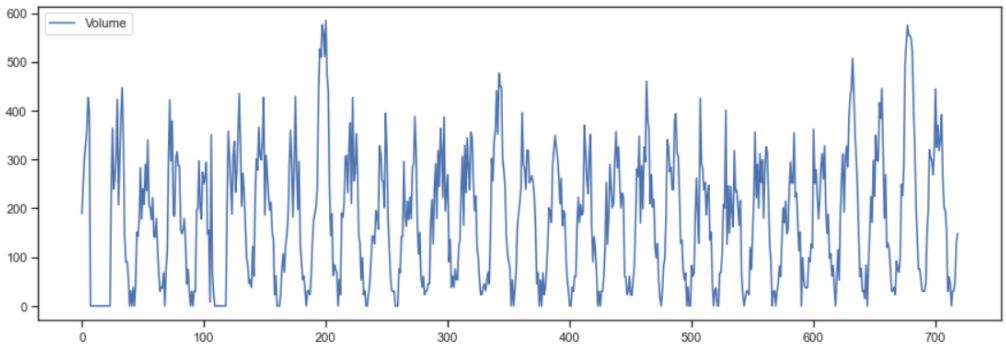


Figura 5 – Dados de volume

Através da Figura 5 é possível perceber que existe uma correlação temporal entre os valores de volume e eles possuem sazonalidade (padrão repetido ao longo do tempo). É possível perceber ainda que os picos de consumo, assim como os valores de consumo zero estão ligados aos horários do dia. O consumo zero de volume geralmente ocorre no período da madrugada.

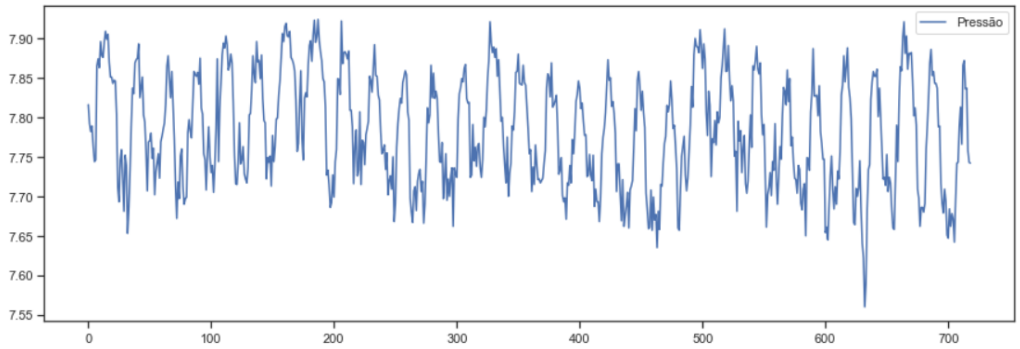


Figura 6 – Dados de pressão

Através da Figura 6 é possível ver o comportamento da variável Pressão ao longo do tempo. Essa variável também possui sazonalidade ligada aos horários do dia.

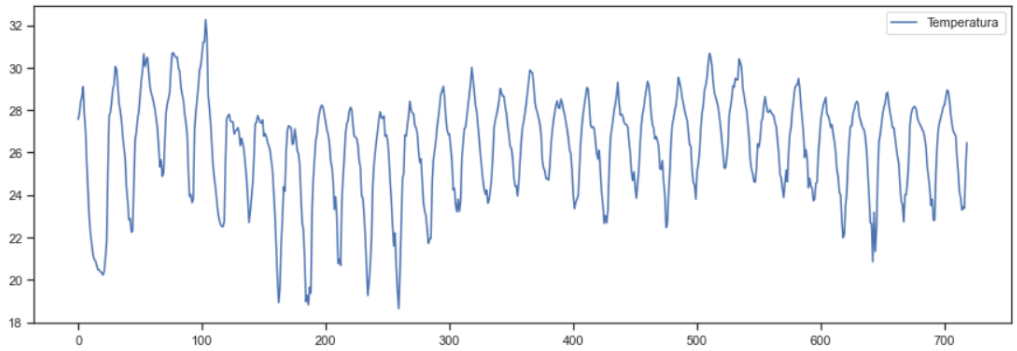


Figura 7 – Dados de temperatura

Através da Figura 7 é possível ver o comportamento da variável Temperatura ao longo do tempo. Essa variável também possui sazonalidade ligada aos horários do dia.

Para as três variáveis foi gerado ainda os box-plots onde é possível verificar os valores máximos, mínimos e mediana para cada variável, como mostra a Figura 8.

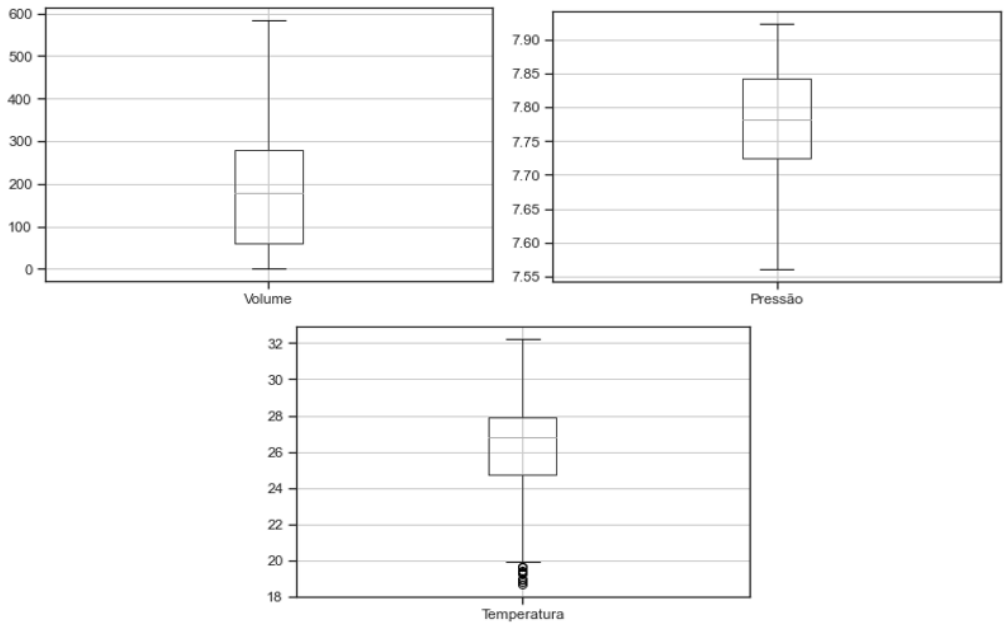


Figura 8 – Box plot das variáveis

Ainda de acordo com a Figura 8, levando em conta o boxplot da variável volume, é possível perceber que a distribuição dessa variável não é simétrica, a mediana gira em torno de $200m^3$, o valor mínimo é $0m^3$ e o máximo de aproxima de $600m^3$. Levando em conta o boxplot da variável pressão, é possível perceber que a distribuição dessa variável também não é simétrica, a mediana gira em torno de 7.8 bar, o valor mínimo é próximo de 7.55 bar e o máximo de aproxima de 7.9 bar. Levando em conta o boxplot da variável temperatura, é possível perceber que a distribuição dessa variável se aproxima mais de uma distribuição simétrica, a mediana gira em torno de $27°C$, o valor mínimo é próximo de $18°C$ e o máximo de aproxima de $32°C$. Além disso a variável temperatura apresenta alguns *outliers* em valores abaixo de $20°C$.

Ainda na análise exploratória foram gerados os gráficos de distribuição de cada variável, como mostra a Figura 9.

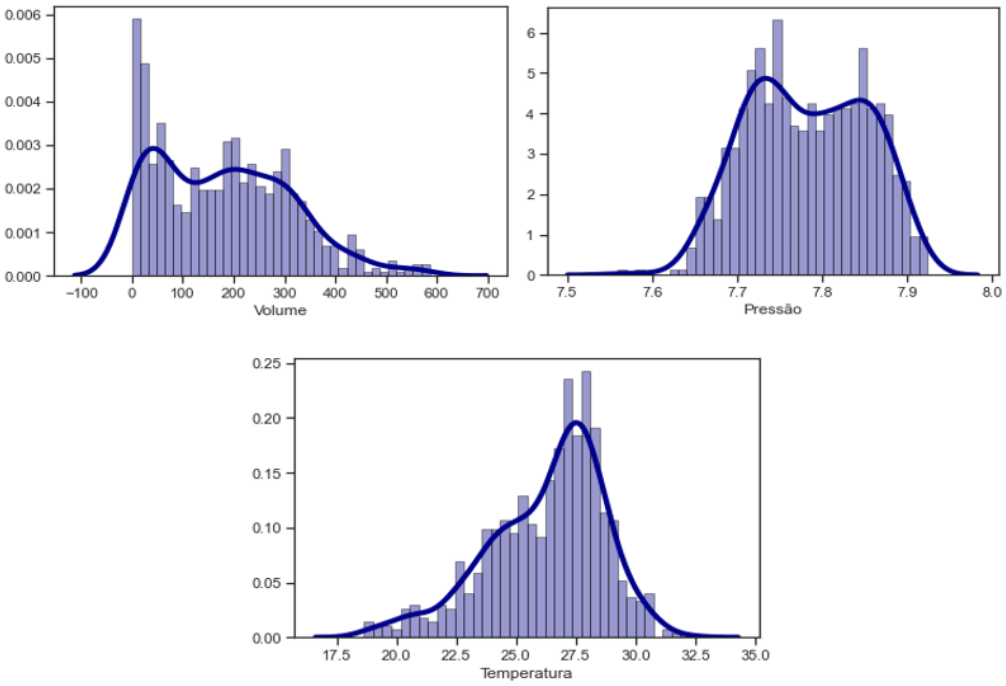


Figura 9 – Distribuição de frequência dos dados coletados das variáveis Volume, Pressão e Temperatura

Ainda de acordo com a Figura 9, levando em conta o gráfico de distribuição de frequência da variável volume, é possível perceber que existe uma alta frequência de zeros fazendo com que a mediana seja inferior à média. Levando em conta o gráfico de distribuição de frequência da variável pressão, é possível perceber que a distribuição é bimodal, apresentando duas modas. Levando em conta o gráfico de distribuição de frequência da variável temperatura, é possível perceber que a distribuição é a que mais se aproxima de uma distribuição simétrica.

3.1.2 Correlações

As Figuras 10 e 11 apresentam os gráficos de dispersão e as respectivas correlações (tanto de Pearson quanto de Spearman) entre todas as combinações das variáveis explicativas e a variável alvo volume.

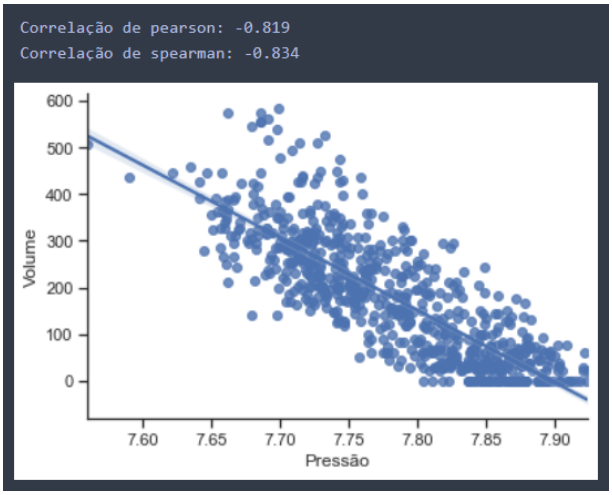


Figura 10 – Gráfico de dispersão para pressão e volume e coeficientes de correlação de Pearson e de Spearman entre pressão e volume

Como mostra a Figura 10, é possível perceber que a correlação entre pressão e a variável alvo volume é forte.

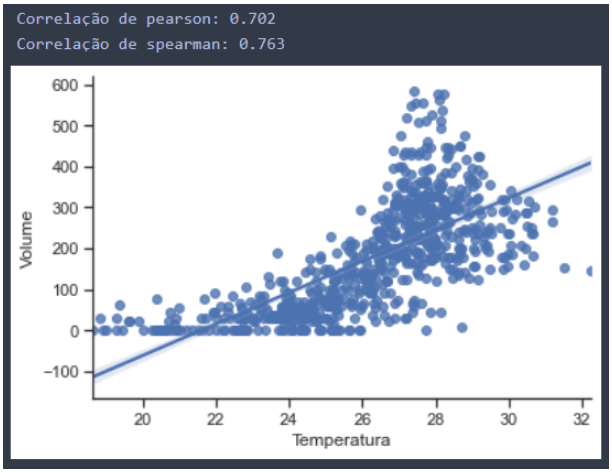


Figura 11 – Gráfico de dispersão para temperatura e volume e coeficientes de correlação de Pearson e de Spearman entre temperatura e volume

Como mostra a Figura 11, é possível perceber que a correlação entre temperatura e a variável alvo volume é forte mas aparenta ser não linear.

Além do cálculo de correlações entre as variáveis explicativas e a variável alvo volume de forma individual, foi ainda calculado a correlação entre a combinação entre temperatura e pressão ($\frac{temperatura}{pressão}$) e a variável alvo volume pelo fato de que temperatura e pressão possuem uma forte multicolinearidade (como será mostrado posteriormente neste trabalho).

A Figura 12 mostra a correlação entre a combinação das variáveis explicativas ($\frac{temperatura}{pressão}$) e a variável alvo volume. Assim como a temperatura, essa combinação de variáveis parece ter uma associação não linear.

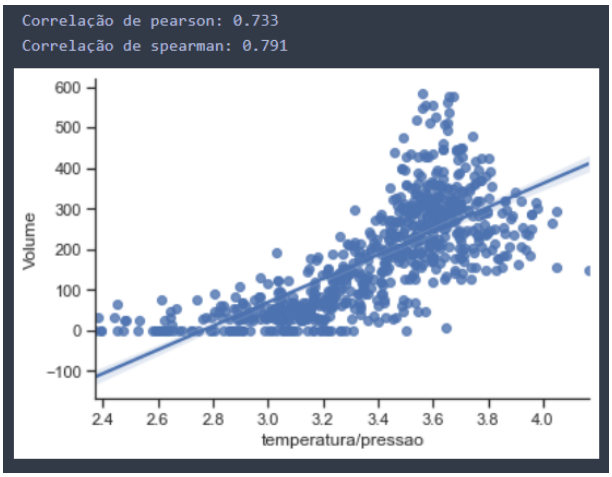


Figura 12 – Gráfico de dispersão para $\frac{temperatura}{pressão}$ e volume e coeficientes de correlação de Pearson e de Spearman entre $\frac{temperatura}{pressão}$ e volume

Os gráficos dos dados no tempo (séries temporais das Figuras 5, 6 e 7) mostram claramente um correlação temporal entre os valores do volume e eles têm sazonalidade (padrão repetido ao longo do tempo), ou seja, um modelo de prediçao para dados desse tipo teria que levar em conta esta dependência no tempo de alguma forma para melhorar a predição. Para isso, foi criada a variável volume_passado, que nada mais é que o valor de volume no instante de tempo t-1. A Figura 13 mostra o gráfico de dispersão entre o volume em t e volume em t-1.

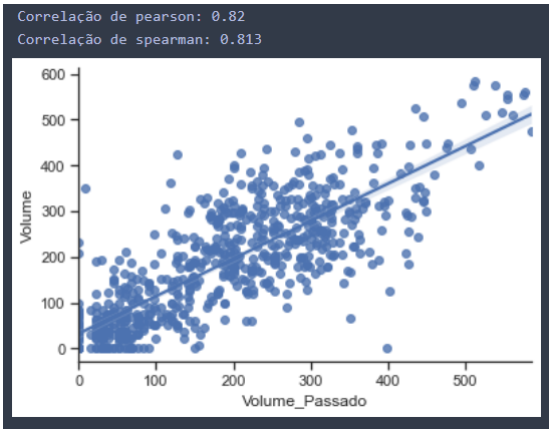


Figura 13 – Gráfico de dispersão para volume passado e volume e coeficientes de correlação de Pearson e de Spearman entre volume passado e volume

3.2 Ajuste da Regressão Linear

3.2.1 Modelo 1

O primeiro ajuste foi feito nos dados de treinamento usando temperatura e pressão para predição do volume, como mostra a Figura 14 a seguir. É possível perceber que tanto temperatura quanto pressão são variáveis estatisticamente significantes (nível de confiança no nível 95%), pois o p-valor de ambas são menores que 0.05, rejeitando a hipótese nula de que são iguais a zero. Além disso os sinais dos coeficientes encontrados das variáveis explicativas fazem sentido, pois temperatura é diretamente proporcional ao volume (sinal positivo no coeficiente) e pressão é inversamente proporcional ao volume (sinal negativo no coeficiente). Usando pressão e temperatura temos um R^2 ajustado de 0.696. A FIgura 14 indica através dos *warnings* que possa haver uma forte multicolinearidade entre as duas variáveis explicativas.

OLS Regression Results

Dep. Variable:	Volume	R-squared:	0.697
Model:	OLS	Adj. R-squared:	0.696
Method:	Least Squares	F-statistic:	576.3
Date:	Sat, 27 Mar 2021	Prob (F-statistic):	1.60e-130
Time:	13:39:41	Log-Likelihood:	-2863.8
No. Observations:	503	AIC:	5734.
Df Residuals:	500	BIC:	5746.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8860.4279	562.250	15.759	0.000	7755.764	9965.092
Temperatura	14.8499	1.784	8.323	0.000	11.344	18.355
Pressão	-1165.0385	67.815	-17.180	0.000	-1298.276	-1031.801

Omnibus:	60.239	Durbin-Watson:	0.488
Prob(Omnibus):	0.000	Jarque-Bera (JB):	84.524
Skew:	0.842	Prob(JB):	4.43e-19
Kurtosis:	4.093	Cond. No.	4.82e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.82e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Figura 14 – Resultados do ajuste da regressão linear usando temperatura e pressão (variável alvo: volume)

3.2.1.1 Análise de colinearidade entre variáveis explicativas

A avaliação da possível presença de multicolinearidade entre as duas variáveis explicativas foi feita por meio do cálculo do valor VIF (*variance inflation factor*). O valor VIF calculado foi de 92.23 indicando que existe multicolinearidade entre as variáveis, pois é um valor bastante alto se comparado ao valor de referência 10.00 usado por alguns autores segundo [3].

3.2.2 Modelo 2

A fim de contornar o problema da multicolinearidade, o segundo ajuste foi feito nos dados de treinamento usando a combinação entre temperatura e pressão ($\frac{temperatura}{pressão}$) para predição do volume, como mostra a Figura 15 a seguir. É possível perceber que a variável $\frac{temperatura}{pressão}$ é estatisticamente significante (nível de confiança de 95%), pois o p-valor é menor que 0.05, rejeitando a hipótese nula de que o coeficiente da variável ($\frac{temperatura}{pressão}$) é igual a zero. Usando $\frac{temperatura}{pressão}$ como variável independente temos um R^2 ajustado de 0.554.

OLS Regression Results

Dep. Variable:	Volume	R-squared:	0.555
Model:	OLS	Adj. R-squared:	0.554
Method:	Least Squares	F-statistic:	624.3
Date:	Sat, 27 Mar 2021	Prob (F-statistic):	4.38e-90
Time:	13:39:41	Log-Likelihood:	-2960.9
No. Observations:	503	AIC:	5926.
Df Residuals:	501	BIC:	5934.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-758.4197	37.557	-20.194	0.000	-832.209	-684.631
temperatura/pressao	278.3215	11.139	24.987	0.000	256.437	300.206

Omnibus:	36.441	Durbin-Watson:	0.640
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49.370
Skew:	0.572	Prob(JB):	1.90e-11
Kurtosis:	4.023	Cond. No.	35.4

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figura 15 – Ajuste da regressão linear usando $\frac{temperatura}{pressão}$ (variável alvo: volume)

3.2.2.1 Análise de resíduos

A Figura 16 mostra os gráficos dos resíduos do Modelo 2. A primeira imagem mostra resíduos *versus* valor predito, a segunda imagem mostra resíduos *versus* $\frac{temperatura}{pressão}$, e a terceira imagem mostra um gráfico Q-Q para verificação de normalidade dos resíduos. Além dos gráficos foi calculado o p-valor do teste de Shapiro-Wilk, como mostra a Figura 17.

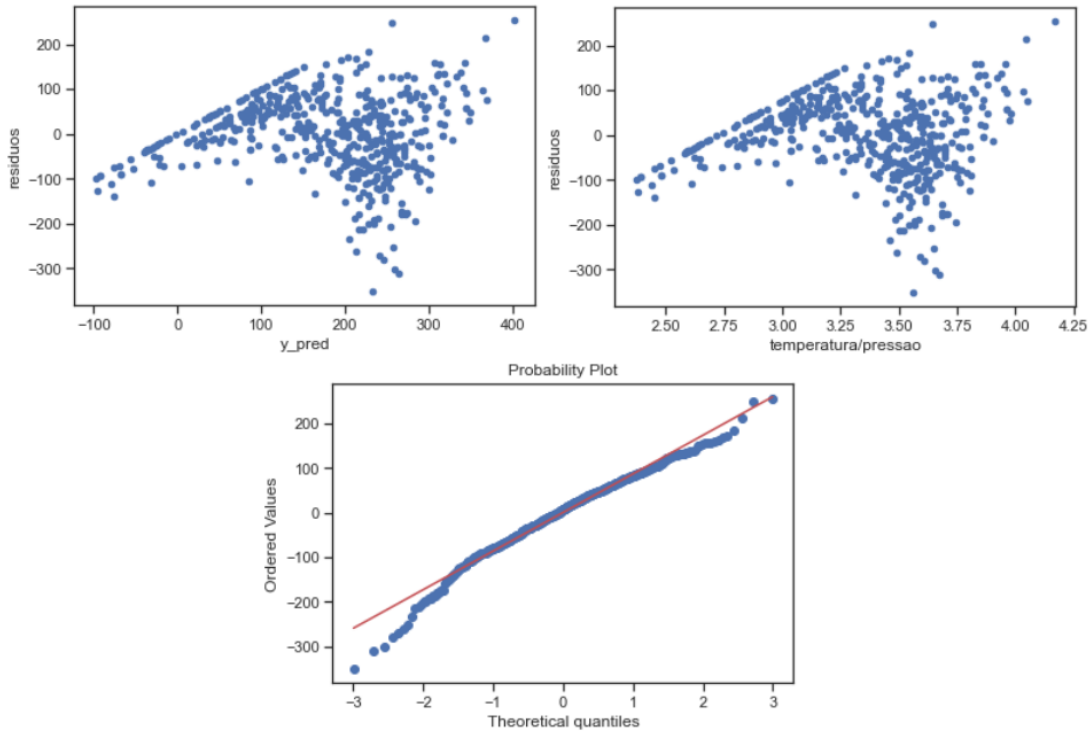


Figura 16 – Análise de resíduos - Modelo 2

```
Shapiro-Wilk normality test
W = 0.98017
data: residuos
p-value = 0.0
```

Figura 17 – Teste de normalidade de Shapiro-Wilk - Modelo 2

Temos então que os erros aparentam ser heterocedásticos de acordo com a Figura 16. O valor-p do teste de normalidade de Shapiro-Wilk é menor que 0.05, rejeitando a hipótese nula de que os erros vêm de uma distribuição normal (nível de confiança de 95%).

3.2.3 Transformação de box-cox

Como a suposição de normalidade dos erros do Modelo 2 não foi considerada válida, a variável resposta foi transformada usando a transformação de box-cox. Usando essa transformação foi encontrado o valor de $\lambda = 0.5$ e então a variável alvo foi transformada para a raiz quadrada.

A Figura 18 a seguir mostra a distribuição da variável alvo antes e depois da transformação.

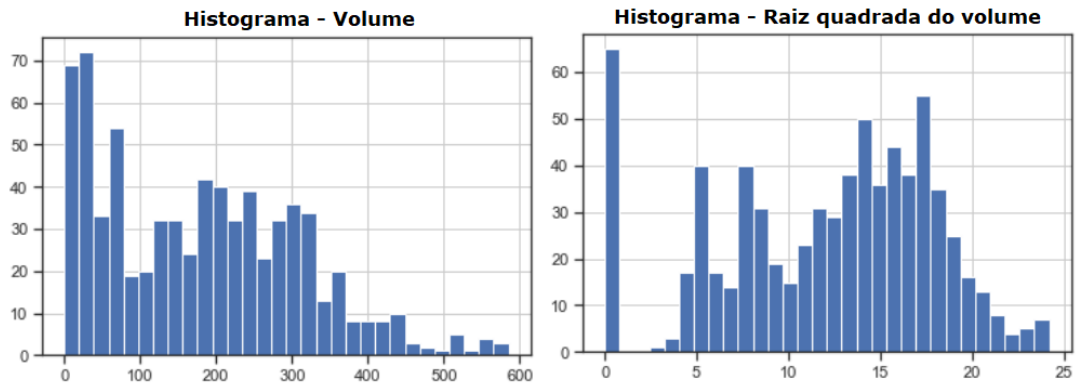


Figura 18 – Histograma do volume de gás antes e depois da transformação de box-cox (raiz quadrada)

3.2.4 Correlações após transformação da variável alvo

A seguir são apresentadas as correlações (tanto de Pearson quanto de Spearman) entre todas as combinações das variáveis explicativas e a variável alvo transformada raiz quadrada do volume.

Como mostra a Figura 19, é possível perceber que a correlação linear de Pearson entre pressão e a variável alvo transformada (raiz quadrada do volume) é forte.

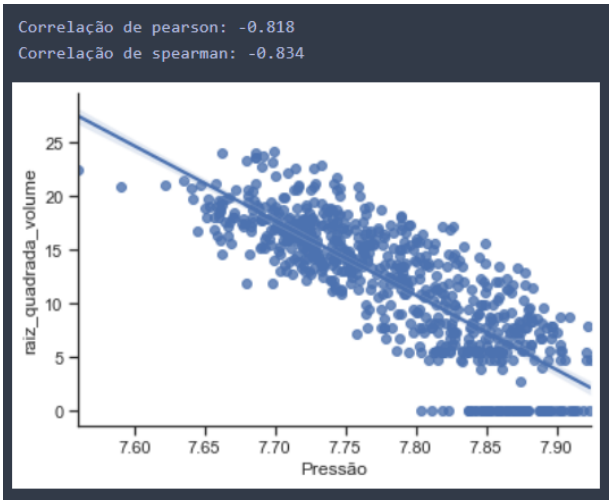


Figura 19 – Gráfico de dispersão para pressão e raiz quadrada do volume e coeficientes de correlação de Pearson e de Spearman entre pressão e raiz quadrada do volume

Como mostra a Figura 20, é possível perceber que a correlação entre temperatura e a variável alvo transformada (raiz quadrada do volume) é moderada e que essa transformação fez com que a correlação linear (Pearson) entre as variáveis aumentasse de 0.702 para 0.761.

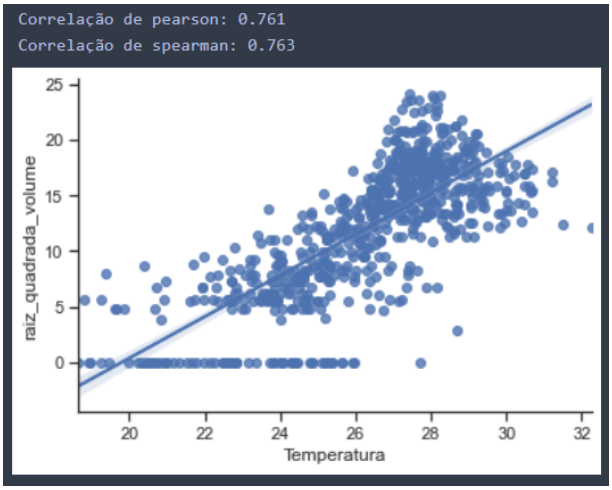


Figura 20 – Gráfico de dispersão para temperatura e raiz quadrada do volume e coeficientes de correlação de Pearson e de Spearman entre temperatura e raiz quadrada do volume

Como mostra a Figura 21, foi calculado ainda a correlação entre a combinação das variáveis explicativas ($\frac{temperatura}{pressão}$) e a variável alvo transformada (raiz quadrada do volume). Assim como a temperatura, essa combinação de variáveis tem uma correlação forte e essa transformação fez com que a correlação linear (Pearson) entre as variáveis aumentasse.

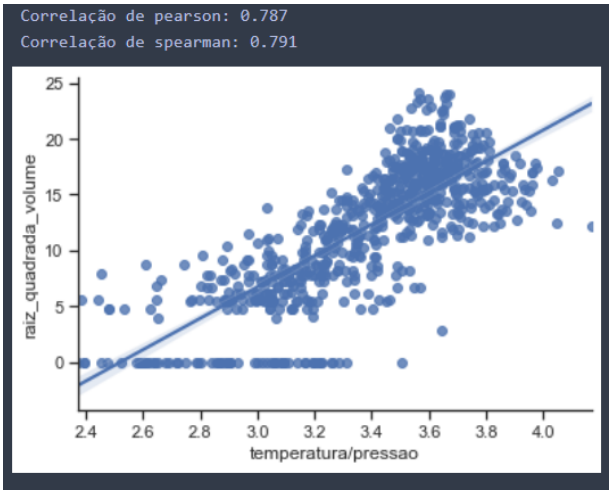


Figura 21 – Gráfico de dispersão para $\frac{temperatura}{pressão}$ e raiz quadrada do volume e coeficientes de correlação de Pearson e de Spearman entre $\frac{temperatura}{pressão}$ e raiz quadrada do volume

Como já foi dito anteriormente, os gráficos dos dados no tempo (séries temporais das Figuras 5, 6 e 7) mostram claramente um correlação temporal entre os valores do volume e eles têm sazonalidade (padrão repetido ao longo do tempo). Portanto, após a transformação da variável alvo, além da variável volume_passado, foi criada também a variável raiz_quadrada volume_passado que é o valor de volume tranformado (raiz quadrada do volume) no instante de tempo t-1. A Figura 22 mostra o gráfico de dispersão entre a raiz quadrada do volume em t e raiz quadrada do volume em t-1.

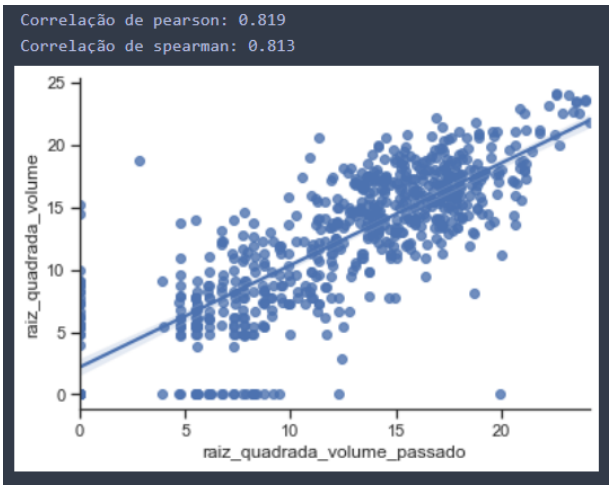


Figura 22 – Gráfico de dispersão para raiz quadrada do volume passado e raiz quadrada do volume e coeficientes de correlação de Pearson e de Spearman entre raiz quadrada do volume passado e raiz quadrada do volume

3.2.5 Modelo 3

O terceiro ajuste foi feito então nos dados de treinamento usando $\frac{temperatura}{pressão}$ para predição da raiz quadrada do volume, como mostra a Figura 23. É possível perceber que a variável explicativa $\frac{temperatura}{pressão}$ é estatisticamente significante (nível de confiança de 95%), pois o p-valor é menor que 0.05, rejeitando a hipótese nula de que o coeficiente dessa variável seja igual a zero. Usando $\frac{temperatura}{pressão}$ temos um R^2 ajustado de 0.633.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     raiz_quadrada_volume   R-squared:                       0.634
Model:                              OLS   Adj. R-squared:                  0.633
Method:                   Least Squares   F-statistic:                     868.3
Date:                  Sat, 06 Mar 2021   Prob (F-statistic):          1.84e-111
Time:                          22:31:44   Log-Likelihood:                -1366.9
No. Observations:                   503   AIC:                             2738.
Df Residuals:                       501   BIC:                             2746.
Df Model:                             1
Covariance Type:              nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                    -34.5192      1.579    -21.862      0.000     -37.621     -31.417
temperatura/pressao       13.7995      0.468     29.467      0.000      12.879      14.720
==============================================================================
Omnibus:                       18.583   Durbin-Watson:                   0.672
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               19.673
Skew:                          -0.455   Prob(JB):                     5.35e-05
Kurtosis:                       3.332   Cond. No.                         35.4
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figura 23 – Ajuste da regressão linear usando $\frac{temperatura}{pressão}$ (variável alvo: raiz quadrada do volume)

3.2.5.1 Análise de resíduos

Na Figura 24 são apresentados os gráficos da análise de resíduos do modelo ajustado. A primeira imagem mostra resíduos *versus* valor predito, a segunda imagem mostra resíduos *versus* $\frac{temperatura}{pressão}$ e a terceira imagem mostra um gráfico Q-Q para verificação de normalidade dos resíduos. Além dos gráficos foi calculado o p-valor do teste de Shapiro-Wilk e o p-valor do teste de Durbin-Watson, como mostram as Figuras 25 e 26 respectivamente.

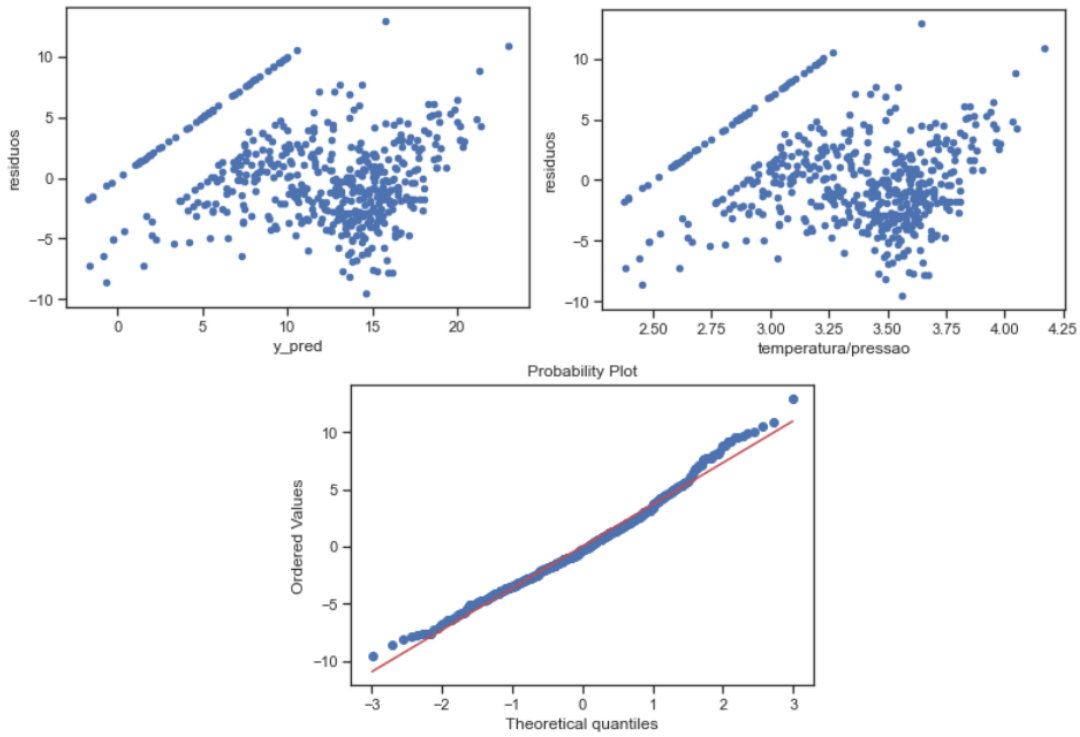


Figura 24 – Análise de resíduos - Modelo 3

```
Shapiro-Wilk normality test
W = 0.98632
data: residuos
p-value = 0.00012
```

Figura 25 – Teste de normalidade de Shapiro-Wilk - Modelo 3

```
lag Autocorrelation D-W Statistic p-value
  1       0.6632301     0.6723264       0
Alternative hypothesis: rho != 0
```

Figura 26 – Teste de Durbin-Watson de autocorrelação dos resíduos - Modelo 3

Temos então que os erros aparentam ser heterocedásticos de acordo com a Figura 25. O valor-p do teste de normalidade de Shapiro-Wilk é menor que 0.05, indicando que a hipótese nula de que os erros vem de uma distribuição normal é rejeitada (nível de confiança de 95%). Temos ainda que o valor-p do teste de Durbin-Watson é menor que 0.05, rejeitando a hipótese nula de que a autocorrelação dos erros seja nula (nível de confiança de 95%), indicando que há autocorrelação nos erros.

3.2.6 Modelo 4

Como o Modelo 3 não apresentou erros com distribuição normal, o quarto ajuste foi feito então nos dados de treinamento usando temperatura e pressão para predição da raiz quadrada do volume, como mostra a Figura 27 a seguir. É possível perceber que tanto temperatura quanto pressão são variáveis estatisticamente significantes (nível de confiança de 95%), pois o p-valor de ambas são menores que 0.05, rejeitando a hipótese nula de que os coeficientes das duas variáveis explicativas são iguais a zero. Além disso os sinais dos coeficientes encontrados das variáveis explicativas fazem sentido, pois temperatura é diretamente proporcional ao volume (sinal positivo no coeficiente) e pressão é inversamente proporcional ao volume (sinal negativo no coeficiente). Usando pressão e temperatura temos um R^2 ajustado de 0.734. Como já foi apresentado, as variáveis explicativas temperatura e pressão apresentam o problema da multicolinearidade.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     raiz_quadrada_volume   R-squared:                       0.735
Model:                              OLS   Adj. R-squared:                  0.734
Method:                   Least Squares   F-statistic:                     791.8
Date:                  Sat, 06 Mar 2021   Prob (F-statistic):          1.65e-165
Time:                          19:12:48   Log-Likelihood:                -1462.7
No. Observations:                   575   AIC:                             2931.
Df Residuals:                       572   BIC:                             2945.
Df Model:                             2
Covariance Type:              nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        380.8765     21.822     17.454      0.000     338.015     423.738
Temperatura    0.8853      0.069     12.741      0.000       0.749       1.022
Pressão      -50.3872      2.638    -19.102      0.000     -55.568     -45.206
==============================================================================
Omnibus:                        0.008   Durbin-Watson:                   0.617
Prob(Omnibus):                  0.996   Jarque-Bera (JB):                0.031
Skew:                          -0.009   Prob(JB):                        0.984
Kurtosis:                       2.969   Cond. No.                     4.70e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.7e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figura 27 – Ajuste da regressão linear usando temperatura e pressão (variável alvo: raiz quadrada do volume)

3.2.6.1 Análise de resíduos

Na Figura 28 temos gráficos em que é possível analisar o comportamento dos resíduos do modelo ajustado. A primeira imagem mostra resíduos versus valor predito, a segunda imagem mostra resíduos versus temperatura, a terceira imagem mostra resíduos versus pressão, e a quarta imagem mostra um gráfico Q-Q para verificação de normalidade dos resíduos. Além dos gráficos foi calculado o p-valor do teste de Shapiro-Wilk e o p-valor do teste de Durbin-Watson, como mostram as Figuras 29 e 30 respectivamente.

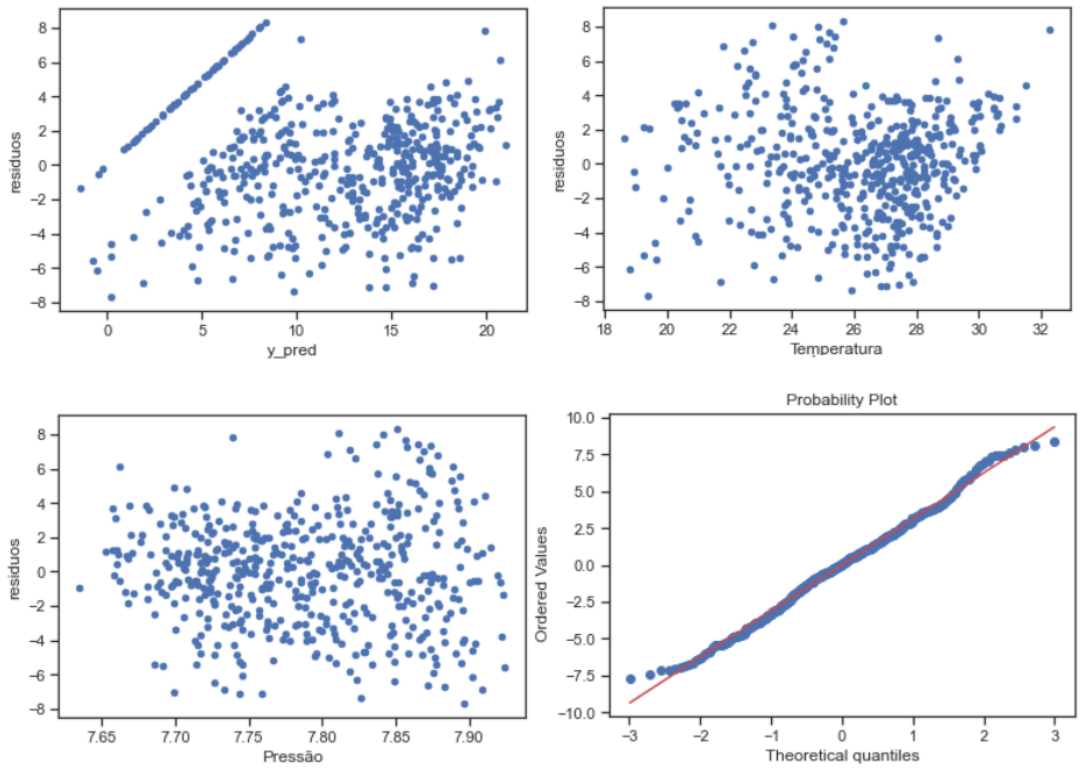


Figura 28 – Análise de resíduos - Modelo 4

```
Shapiro-Wilk normality test
W = 0.99465
data: residuos
p-value = 0.07665
```

Figura 29 – Teste de normalidade de Shapiro-Wilk - Modelo 4

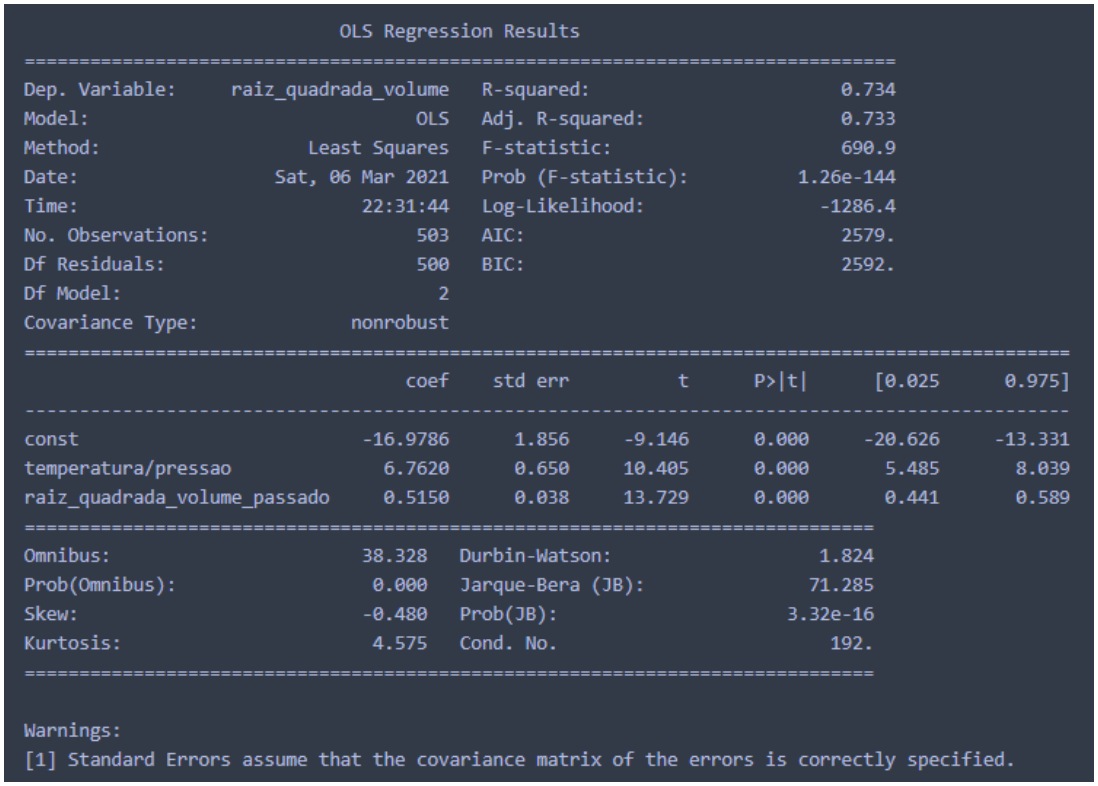
```
lag Autocorrelation D-W Statistic p-value
  1        0.698         0.617          0
 Alternative hypothesis: rho != 0
```

Figura 30 – Teste de Durbin-Watson de autocorrelação dos resíduos - Modelo 4

Temos então que os erros aparentam ser homocedásticos de acordo com a Figura 29. O valor-p do teste de normalidade de Shapiro-Wilk é maior que 0.05, indicando que a hipótese nula de que os erros vêm de uma distribuição normal não é rejeitada (nível de confiança de 95%). Mas o valor-p do teste de Durbin-Watson é menor que 0.05, rejeitando a hipótese nula de que a autocorrelação dos erros seja nula (nível de confiança de 95%), indicando que há autocorrelação nos erros.

3.2.7 Modelo 5

Como o Modelo 4 apresentou problemas de autocorrelação dos erros, foi ajustado ainda o quinto modelo. Esse quinto ajuste foi feito nos dados de treinamento usando duas variáveis explicativas: i) a combinação entre temperatura e pressão ($\frac{temperatura}{pressão}$) e ii) a variável raiz quadrada do volume no instante de tempo t-1 (raiz_quadrada_volume_ passado), como mostra a Figura 31. É possível perceber que tanto a variável $\frac{temperatura}{pressão}$ quanto a variável raiz_quadrada_volume_passado são variáveis estatisticamente significantes (nível de confiança de 95%), pois o p-valor de ambas são menores que 0.05, rejeitando a hipótese nula de que os coeficientes dessas variáveis são iguais a zero. Usando $\frac{temperatura}{pressão}$ e raiz_quadrada_volume_ passado temos um R^2 ajustado de 0.733



OLS Regression Results

Dep. Variable:	raiz_quadrada_volume	R-squared:	0.734
Model:	OLS	Adj. R-squared:	0.733
Method:	Least Squares	F-statistic:	690.9
Date:	Sat, 06 Mar 2021	Prob (F-statistic):	1.26e-144
Time:	22:31:44	Log-Likelihood:	-1286.4
No. Observations:	503	AIC:	2579.
Df Residuals:	500	BIC:	2592.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-16.9786	1.856	-9.146	0.000	-20.626	-13.331
temperatura/pressao	6.7620	0.650	10.405	0.000	5.485	8.039
raiz_quadrada_volume_passado	0.5150	0.038	13.729	0.000	0.441	0.589

Omnibus:	38.328	Durbin-Watson:	1.824
Prob(Omnibus):	0.000	Jarque-Bera (JB):	71.285
Skew:	-0.480	Prob(JB):	3.32e-16
Kurtosis:	4.575	Cond. No.	192.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figura 31 – Ajuste da regressão linear usando $\frac{temperatura}{pressão}$ e raiz quadrada do volume passado (variável alvo: raiz quadrada do volume)

3.2.7.1 Análise de multicolinearidade entre variáveis explicativas

Para verificar se existe uma forte multicolinearidade entre as duas variáveis explicativas foi calculado o valor VIF (*variance inflation factor*). O valor VIF calculado foi de 6.65. É possível perceber que não existe multicolinearidade entre as variáveis pois o valor VIF não é considerado alto se comparado ao valor de referência 10.00 usado por alguns autores segundo [3].

3.2.7.2 Análise de resíduos

Na Figura 32 a seguir temos gráficos da análise de resíduos do modelo ajustado. A primeira imagem mostra resíduos *versus* valor predito, a segunda imagem mostra resíduos *versus* $\frac{temperatura}{pressão}$, a terceira imagem mostra resíduos versus raiz_quadrada_volume _passado, e a quarta imagem mostra um gráfico Q-Q para verificação de normalidade dos resíduos. Além dos gráficos foi calculado o p-valor do teste de Shapiro-Wilk e o p-valor do teste de Durbin-Watson, como mostram as Figuras 33 e 34 respectivamente.

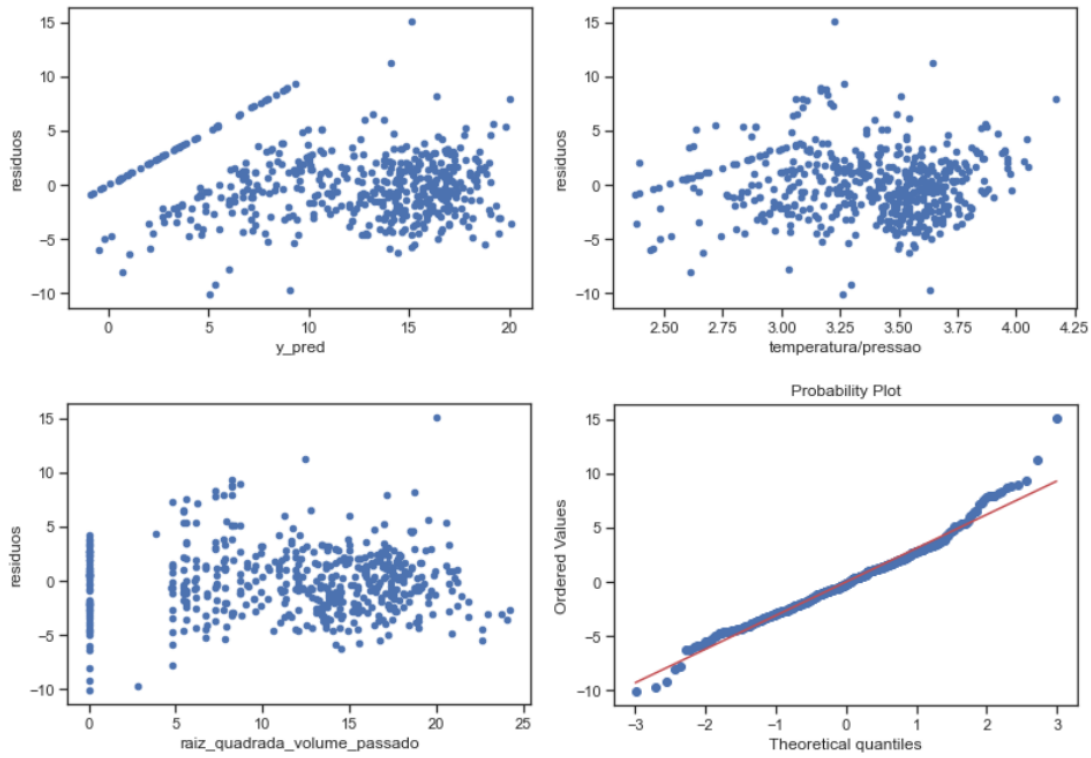


Figura 32 – Análise de resíduos - Modelo 5

```
Shapiro-Wilk normality test
W = 0.9797
data: residuos
p-value = 0.0
```

Figura 33 – Teste de normalidade de Shapiro-Wilk - Modelo 5

lag	Autocorrelation	D-W Statistic	p-value
1	0.087	1.858	0.044

Alternative hypothesis: rho != 0

Figura 34 – Teste de Durbin-Watson de autocorrelação dos resíduos - Modelo 5

Temos então que os erros aparentam ser homocedásticos de acordo com a Figura 32. O valor-p do teste de normalidade de Shapiro-Wilk é menor que 0.05, rejeitando a hipótese nula de que os resíduos vem de uma distribuição normal (nível de confiança 95%). Além disso, o valor-p do teste de Durbin-Watson é menor que 0.05, rejeitando a hipótese de que a autocorrelação dos erros seja nula (nível de confiança 95%), indicando que há indícios de autocorrelação nos erros.

Após as análises desses cinco modelos ajustados foi escolhido como melhor modelo o Modelo 5. No Modelo 5, apesar do teste de Shapiro-Wilk ter rejeitado a hipótese de normalidade dos erros, a autocorrelação dos erros diminuiu muito em relação ao Modelo 4 (não apresentou evidências contra a suposição de normalidade dos erros). Ainda que o teste de Durbin-Watson tenha acusado autocorrelação significante, o valor da autocorrelação é muito baixa. Tendo em vista que os dados têm uma clara relação temporal, a melhor escolha seria o Modelo 5.

3.3 Construção do *Digital Twin*

A equação do modelo ajustado é mostrada a seguir:

$$\sqrt{Volume}_{(t)} = -16.97857706 + 6.76198614 \frac{temperatura}{pressão}(t) + 0.51503907 \sqrt{Volume}_{(t-1)}$$

Essa equação mostra que para cada aumento de unidade ($\frac{^{\circ}C}{bar}$) da variável $\frac{temperatura}{pressão}$ no tempo t, a variável raiz quadrada do volume no tempo t aumenta, em média, 6.76198614 unidades de medida (raiz quadrada do volume em m^3: $m\sqrt{m}$). Ela mostra ainda que para cada aumento de unidade ($m\sqrt{m}$) da variável raiz quadrada do volume no tempo t-1, a variável raiz quadrada do volume no tempo t aumenta, em média, 0.51503907 unidades de medida (raiz quadrada do volume em m^3: $m\sqrt{m}$).

O valor de média do erro absoluto no conjunto de validação foi de 2.2712 e o valor do desvio-padrão do erro absoluto do conjunto de validação foi de 1.9582. Portanto, a *flag* de anomalia é levantada quando o valor de erro entre valor_predito e valor_real (na escala raiz quadrada do volume) for maior que 2.2712 + (2 * 1.9582) = 6.1876. Esse valor é um valor grande na escala dos dados, mas pode ser ajustado conforme o sistema for testado de forma *online*, baseado na quantidade de falsos positivos e falsos negativos.

Usando os dados do conjunto de validação (que não foi usado para o ajuste da regressão linear), foram preditos os valores do volume. A Figura 35 a seguir mostra os valores preditos, os valores reais e os momentos considerados como anomalia:

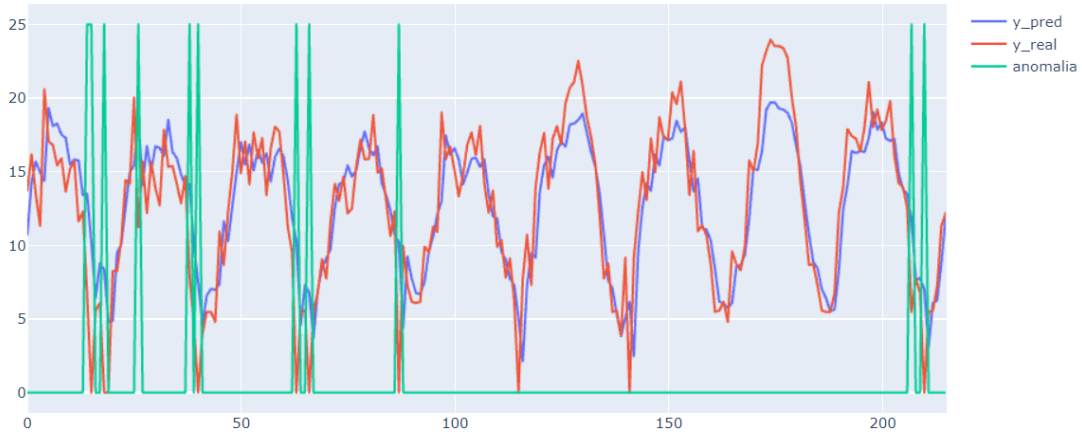


Figura 35 – Valores preditos, reais e anomalias - Conjunto de validação

A Figura 35 mostra que em momentos que o consumo de volume (real) é maior que zero o modelo consegue acompanhar bem o comportamento de consumo. É possível perceber que os momentos tidos como anomalia acontecem em momentos de consumo de volume próximos de zero.

Na Figura 36 a seguir temos um gráfico de dispersão de valores preditos (y) contra os valores reais (x) nesse conjunto de validação. Temos em destaque os valores que foram considerados como anomalia marcados com um x e na cor amarela:

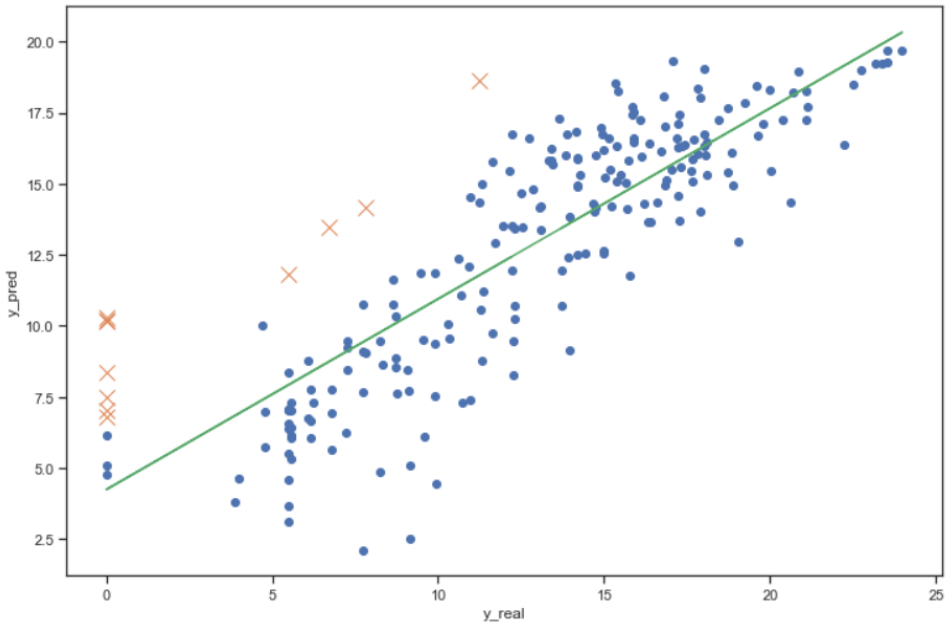


Figura 36 – Gráfico de dispersão para valores preditos e valores reais no conjunto de validação - Pontos marcados com X indicam os valores considerados anomalias

É possível perceber através da Figura 36 que a maioria dos pontos estão próximos da linha de regressão, indicando um bom ajuste, com exceção de momentos que o consumo de volume é baixo e/ou próximo de zero.

4 CONCLUSÃO E TRABALHOS FUTUROS

O desenvolvimento do presente estudo possibilitou a utilização da Estatística, via Regressão Linear, como ferramenta para auxiliar na detecção de possíveis anomalias em medições de volume de gás em distribuidoras de gás.

De um modo geral, o modelo final ajustado conseguiu acompanhar bem o comportamento do consumo de volume real no conjunto de validação, principalmente para valores de consumo de volume maiores que zero, como mostram as Figuras 35 e 36. Apesar do teste de Shapiro-Wilk ter rejeitado a hipótese de normalidade dos erros, a autocorrelação dos erros diminuiu muito em relação aos outros modelos. Mesmo o teste de Durbin-Watson tendo acusado autocorrelação dos erros estatisticamente significante, seu valor é muito baixo. Ademais, as variáveis explicativas se mostraram estatisticamente significantes e o modelo apresentou um coeficiente de determinação de 0.733. O valor de média do erro absoluto no conjunto de validação foi de 2.2712 e o valor do desvio-padrão do erro absoluto do conjunto de validação foi de 1.9582. Sendo assim, o limiar usado para a detecção de anomalia foi de 6.1876.

Um ponto sensível no modelo são os valores de consumo iguais a zero. É possível perceber que os momentos indicados como anomalia acontecem quando há consumo com volume igual a zero, podendo indicar que o consumo zero está associado com outros fatores que não estão sendo considerados no modelo. Na análise descritiva dos dados, é possível perceber que valores de consumo zero ocorrem para diversos valores de pressão e temperatura.

Dada a importância do assunto, torna-se necessário uma maneira de melhorar o ajuste do modelo, coletando mais dados, por exemplo, e até mesmo conversando com profissionais que trabalham na distribuidora de gás a fim de tentar identificar quais outras variáveis disponíveis poderiam influenciar no consumo zero de volume de gás. Além disso, para melhorar o modelo ainda é possível a tentativa de ajuste de outros modelos de regressão como GLM (*Generalized Linear Model*) e Regressão de Cumeeira (solução matemática em que se adiciona uma constante à diagonal principal da matriz de correlação a fim de contornar o problema de forte multicolinearidade nas variáveis explicativas) para o caso de modelos de regressão em que as variáveis temperatura e pressão aparecem separadas. Uma outra tentativa de melhoria no modelo seria restringir o horário de predição para excluir os momentos de consumo 0 de volume, uma vez que esses momentos acontecem, geralmente, em um horário específico do dia.

Nesse sentido, este trabalho teve caráter exploratório mostrando que, com o uso das ferramentas de análise de regressão, pode ser possível detectar anomalias na medição de consumo de gás, mas ainda é necessário uma revisão sobre a questão dos zeros e como

acomodá-los no modelo.

REFERÊNCIAS

[1] Bolton, Ruth N.; McColl-Kennedy, Janet R.; Cheung, Lilliemay; Gallan, Andrew; Orsingher, Chiara; Witell, Lars; Zaki, Mohamed (2018). **"Customer experience challenges: Bringing together digital, physical and social realms"**. Journal of Service Management. 29 (5): 776–808. doi:10.1108/JOSM-04-2018-0113.

[2] JAMES, G.; WITTEN, D.; HASTIE T; TIBSHIRANI, R. **An Introduction to Statistical Learning** USA, 2015. ISSN 1431-875X.

[3] MONTGOMERY, D.C.; RUNGER, G.C. **Estatística Aplicada e Probabilidade para Engenheiros.** LTC Editora, (Quarta Edição).

[4] AFONSO REIS, I. (2020). **Regressão Linear: quantificando a relação entre variáveis de um processo** [PowerPoint presentation]

[5] AFONSO REIS, I. (2020). **Regressão Linear Simples. Análise de resíduos: verificando suposições do modelo** [PowerPoint presentation]

[6] Pré-Cálculo (2020) **Funções monótonas** Disponível em: http://www.professores.im-uff.mat.br/hjbortol/disciplinas/2016.2/gma00116/arquivos/gma00116-slides-03.pdf. Acesso em: 05 Janeiro 2021.