Universidade Federal do Rio Grande – FURG
Probabilidade e Estatística Aplicada à Engenharia
Rio Grande/RS, Brasil
Engenharia de Computação

Análise de Correlação e Regressão Linear Aplicada entre Variaveis

Orientador(a): Viviane Leite Dias de Mattos
Aluno: Andrey Vinicius Santos Souza

janeiro 2025

1 Introdução

Compreender o comportamento financeiro, particularmente no mercado de ações, tornou-se essencial tanto para o desenvolvimento de trabalhos científicos quanto para a tomada de decisões estratégicas em negócios [2]. Os dados das ações apresentam padrões distintos, com variações e alterações ao longo do tempo, o que contribui para a imprevisibilidade característica do mercado de ações. Um dos campos interdisciplinares que tem recebido crescente atenção é o de *stock market prediction*, tema de grande interesse entre pesquisadores [1], que busca explorar métodos estatísticos e computacionais para interpretar e prever as flutuações do mercado.

Nesse contexto, destaca-se a importância da análise estatística como ferramenta essencial para identificar padrões, modelar relações entre variáveis e fornecer insights valiosos aplicáveis a diferentes áreas [4]. Técnicas como a análise de correlação e a regressão linear desempenham papéis cruciais ao permitir o estudo de interdependências entre variáveis e a criação de modelos matemáticos que descrevem comportamentos observados em dados financeiros.

Este trabalho enfatiza o uso de técnicas estatísticas, como a análise de correlação e a regressão linear, que possibilitam o estudo das relações entre variáveis e a construção de modelos matemáticos para descrever padrões em dados [6]. Em continuidade ao estudo prévio intitulado "Análise Exploratória e de Correlação entre Duas Variáveis: Volume e Preço Médio com Técnicas de Estatística Descritiva", esta pesquisa amplia a abordagem inicial ao explorar mais profundamente a relação entre as variáveis "**Average**" (média dos preços) e "**Volume**" (volume de transações) da ação **PETR4**.

Os dados utilizados provêm do conjunto *BovDBv2*, um *dataset* de referência para estudos de previsão do mercado de ações[2], publicamente acessível e pré-processado, contendo informações diárias de todas as ações negociadas na B3 (Brasil Bolsa Balcão)[3] no período de 1995 a 2024.

[1] `https://www.forbes.com/sites/investor-hub/article/stock-market-predictions-2025/`
[2] `https://github.com/Ginfofinance/BovDbV2repository`
[3] `https://www.b3.com.br`

Este trabalho está organizado da seguinte forma: a Seção 2 apresenta a metodologia adotada, enquanto a Seção 3 detalha os resultados obtidos. As conclusões e sugestões para trabalhos futuros são discutidas na Seção 4.

2 Metodologia

A análise de dados financeiros, particularmente no contexto do mercado de ações, exige o uso de ferramentas estatísticas robustas que possibilitem a identificação de padrões e a modelagem de relações entre variáveis. Esta seção apresenta a metodologia proposta neste trabalho, cujo principal objetivo é fornecer uma compreensão abrangente da abordagem utilizada, aplicando ferramentas como análise de correlação e análise de regressão. Estas desempenham papéis fundamentais ao permitir o estudo das interdependências entre variáveis e a construção de modelos matemáticos para descrever os comportamentos observados.

A seção está organizada da seguinte forma: na subseção 2.1, são discutidos os dados utilizados e sua origem; a subseção 2.2 apresenta informações sobre o cálculo e a fórmula para a análise de correlação;a subseção 2.3 detalha o processo de análise de regressão e os testes estatísticos. E, por fim temos a seção 2.4 explicando brevemente sobre o *Software* utilizado e suas Blibiotecas.

2.1 Dados

Para o trabalho proposto foi ultilizado o conjunto de dados BovDBv2 apresentado na introdução. Nele, foram extraídas informações das colunas "Volume" e "Preço Médio" da ação **PETR4** (Petroleo Brasileiro S.A. Petrobras) no período de '2024-01-02' a '2024-06-28'. Esses dados estão presentes na tabela *price*, que contém os dados diários das ações, com colunas representando as variáveis e as linhas representando os dias. As variáveis analisadas, "Volume" e "Preço Médio", representam métricas fundamentais para compreender o comportamento do papel **PETR4** [1]. Esta subseção contextualiza essas variáveis, explicando como são usadas e seus respectivos cálculos para formar os valores presentes no banco de dados.

Cálculo do Volume A variável **Volume** é definida como o produto entre o preço de negociação (Preço_i) e a quantidade de ações transacionadas (Quantidade_i) dentro de um intervalo de tempo específico, sendo calculada por:

$$\text{Volume} = \sum_{i=1}^{n} (\text{Preço}_i \times \text{Quantidade}_i).$$

Esta métrica reflete o total financeiro negociado em um período, e no caso dos dados utilizados, o período corresponde a um dia de negociação.

Cálculo do Preço Médio O **Preço Médio** é obtido dividindo o volume financeiro total pela quantidade total de ações negociadas:

$$\text{Preço Médio} = \frac{\text{Volume}}{\text{Quantidade Total}}.$$

Esta variável indica a média ponderada do preço das transações, refletindo o preço médio de negociação das ações em um determinado período.

A escolha da ação **PETR4** se justifica por sua alta liquidez e representatividade no mercado financeiro brasileiro, sendo uma das ações mais negociadas na B3 (Brasil, Bolsa, Balcão). Além disso, sua relevância para o setor de energia e sua volatilidade tornam este papel uma excelente escolha para análise estatística de padrões financeiros. Estudos como o de Fabozzi et al. (2014) [3] destacam a importância de utilizar ações de alta liquidez em análises de correlação e regressão, especialmente em mercados emergentes como o brasileiro.

2.2 Análise de Correlação

A correlação mede o grau de relacionamento entre duas variáveis quantitativas, avaliando como as variações em uma estão associadas às variações na outra. Para visualizar esse relacionamento, utiliza-se o *diagrama de dispersão*, que representa graficamente os pares ordenados de observações das variáveis analisadas. Este gráfico é uma ferramenta fundamental para identificar visualmente a direção e a intensidade do relacionamento entre as variáveis, sendo muito útil para observar os padrão de correlação. Exemplo, Uma correlação positiva apresenta pontos alinhados em uma direção ascendente ou a ausência de correlação resulta em pontos dispersos aleatoriamente. A figura a seguir mostra exemplos de diagramas de dispersão:

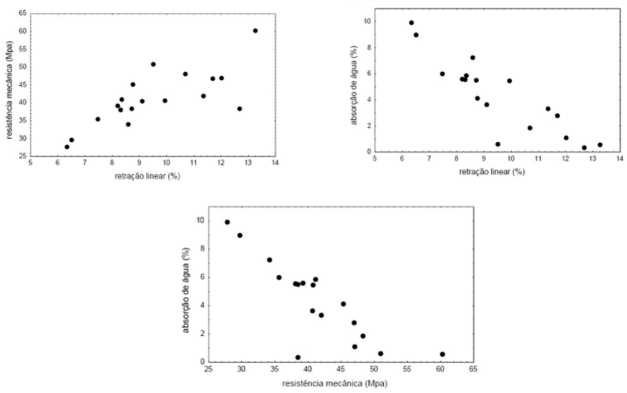

Figure 1: Exemplos de Diagramas de dispersão

Coeficiente de correlação linear de Pearson :

O coeficiente de correlação linear de Pearson (r) é uma métrica amplamente utilizada para quantificar o grau de relação linear entre duas variáveis quantitativas. A fórmula geral do coeficiente é:

$$r_{x,y} = \frac{n\sum(x \cdot y) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2] \cdot [n\sum y^2 - (\sum y)^2]}},$$

onde:

- x e y representam as variáveis observadas;
- n é a quantidade de pares de observações.

Os valores de r variam entre -1 e 1 e podem ser interpretados da seguinte maneira: $r = +1$: indica uma **correlação perfeita positiva**, onde as variáveis crescem proporcionalmente. $0 < r < +1$: indica uma **correlação positiva**, onde o aumento de uma variável está associado ao aumento da outra, mas sem perfeita proporcionalidade. $r = 0$: indica **ausência de correlação linear**, ou seja, não há relação linear significativa entre as variáveis. $-1 < r < 0$: indica uma **correlação negativa**, onde o aumento de uma variável está associado à diminuição da outra. $r = -1$: indica uma **correlação perfeita negativa**, onde o aumento de uma variável está perfeitamente associado à redução da outra.

Além disso a intensidade do coeficiente de correlação também pode ser classificada, conforme ilustrado na escala da figura 2:

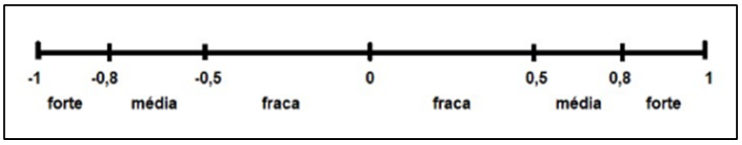

Figure 2: Escala de intensidade da Correlação de Pearson

- $-1,0 \leq r \leq -0,8$ ou $0,8 \leq r \leq 1,0$: correlação **forte**;
- $-0,8 < r \leq -0,5$ ou $0,5 \leq r < 0,8$: correlação **moderada**;
- $-0,5 < r < 0,5$: correlação **fraca**.

As classificações são úteis para interpretar o impacto prático da correlação no contexto do estudo. Neste presente trabalho, será utilizado o coeficiente de Pearson para avaliar o grau de correlação linear entre as variáveis *Volume* e *Preço Médio* da ação **PETR4**, apresentando na seção 3 proporcionando uma compreensão estatística do relacionamento entre essas métricas.

2.3 Análise de Regressão

A análise de regressão é uma técnica estatística amplamente utilizada para modelar e investigar a relação entre uma variável dependente (também chamada de

resposta) e uma ou mais variáveis independentes (também chamadas de preditoras ou explicativas), de tal forma que uma variável pode ser predita apartir da outra ou outras, além de avaliar a força e a natureza do relacionamento entre elas.

Neste trabalho, utilizaremos o modelo de regressão linear simples para avaliar a relação entre o *Preço Médio* (Y) da ação **PETR4** como variável dependente e o *Volume* (X) como variável independente. A seguir, apresentamos os aspectos principais da análise realizada.

Enquanto a análise de correlação (apresentada na Seção 2.2) mede apenas o grau de associação linear entre duas variáveis, a análise de regressão vai além, fornecendo uma equação que descreve a relação entre as variáveis. A regressão também permite realizar inferências estatísticas sobre os coeficientes do modelo, avaliar a adequação do modelo ajustado e identificar padrões nos dados por meio de resíduos.

2.3.1 Modelo de Regressão Linear

A reta de regressão e utilizada para prever o valor de uma variável dependente com base em uma independente descrevendo melhor a relação entre duas variáveis. A equação de regressão linear simples assume a seguinte forma:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

onde:

- β_0: intercepto (valor esperado de Y quando $X = 0$);
- β_1: coeficiente angular (taxa de variação de Y para cada unidade de variação em X);
- ε: erro aleatório, que assume distribuição normal com média zero e variância constante.

O valores obtidos para a reta de regressão ajustada será discutida em detalhes na seção 3, considerando os valores estimados de β_0 e β_1, além de uma avaliação gráfica da linha ajustada em relação aos dados observados.

2.3.2 Teste de Significância dos Coeficientes do Modelo

Os coeficientes β_0 e β_1 são avaliados por meio do teste de hipóteses, com foco em determinar a significância do coeficiente β_1, que representa a relação linear entre X e Y.

- **Hipóteses para β_1:**

$$H_0 : \beta_1 = 0 \quad \text{(não há relação linear entre } X \text{ e } Y\text{)},$$

$$H_1 : \beta_1 \neq 0 \quad \text{(existe relação linear entre } X \text{ e } Y\text{)}.$$

O teste de significância é realizado utilizando a estatística t, calculada pela fórmula:

$$t = \frac{\hat{\beta}_1}{\text{EP}(\hat{\beta}_1)},$$

onde:

- $\hat{\beta}_1$ é o valor estimado do coeficiente β_1;
- $\text{EP}(\hat{\beta}_1)$ é o erro padrão da estimativa $\hat{\beta}_1$.

O valor-p associado ao teste é comparado com o nível de significância de 5% ($\alpha = 0,05$). Se o valor-p for menor que α, rejeitamos a hipótese nula H_0, indicando que há evidências estatísticas de uma relação linear significativa entre X e Y. Caso contrário, não rejeitamos H_0, indicando ausência de evidências para tal relação.

2.3.3 Análise de Variância (ANOVA)

A ANOVA é usada para avaliar a qualidade do modelo de regressão, decompondo a variabilidade total (SQT) em dois componentes principais:

Soma dos Quadrados da Regressão (SQR): Variabilidade explicada pelo modelo, associada às variáveis independentes.

Soma dos Quadrados dos Resíduos (SQE): Variabilidade não explicada pelo modelo, relacionada ao erro ou dispersão dos dados.

A hipótese nula (H_0) assume que todos os coeficientes do modelo são nulos ($\beta_1 = 0$ no caso de regressão simples), indicando que o modelo ajustado não é adequado. Já a hipótese alternativa (H_1) sugere que pelo menos um coeficiente é diferente de zero ($\beta_1 \neq 0$), indicando a relevância do modelo.

A decisão estatística é baseada no valor-p associado à razão F_{calc}: $p < 0,05$: Rejeita-se H_0, indicando que o modelo é significativamente melhor do que um modelo nulo. $p \geq 0,05$: Não há evidências suficientes para rejeitar H_0, sugerindo que o modelo não é adequado.

As fórmulas para os cálculos são:

$$SQT = \sum (y_i - \bar{y})^2, \quad SQR = \sum (\hat{y}_i - \bar{y})^2, \quad SQE = \sum (y_i - \hat{y}_i)^2,$$

onde y_i são os valores observados, $\hat{y}_i$ os valores ajustados e $\bar{y}$ a média dos valores observados.

A interpretação final depende do valor de F_{calc}: valores altos indicam que a variabilidade explicada pelo modelo é substancialmente maior que a variabilidade residual, validando a relevância das variáveis independentes. A rejeição de H_0 confirma que o modelo ajustado é significativo.

2.3.4 Coeficiente de Determinação (R^2)

O coeficiente de determinação, representado por R^2, mede a proporção da variação total de uma variável dependente (y) que é explicada pela variação da

variável independente (x) em um modelo estatístico, como a regressão linear. Em termos simples, ele avalia o quão bem os dados observados são explicados pelo modelo. A fórmula do R^2 é dada por:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \text{ (Variação explicada)}}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \text{ (Variação total)}}$$

- **Variação explicada (numerador)**: Mede quanto os valores previstos ($\hat{y}_i$) se desviam da média observada ($\bar{y}$).
- **Variação total (denominador)**: Mede o desvio dos valores observados (y_i) em relação à média ($\bar{y}$).

Um valor de R^2 maior ou igual a 0,5 indica que o modelo é adequado para descrever a relação entre as variáveis, pois explica pelo menos 50% da variabilidade total dos dados. Por outro lado, se R^2 for menor que 0,5, significa que o modelo explica menos de 50% da variabilidade, indicando uma baixa capacidade de captura da relação entre as variáveis, possivelmente devido a uma fraca correlação entre elas ou a inadequações do modelo escolhido.

2.3.5 Análise Exploratória dos Resíduos

A análise dos resíduos ($e_i = Y_i - \hat{Y}_i$) é fundamental para avaliar a adequação do modelo ajustado. Resíduos bem comportados indicam que o modelo representa adequadamente os dados e que as suposições do modelo foram atendidas. Essa análise inclui os seguintes aspectos:

- **Independência dos resíduos:** Para verificar a independência dos resíduos, será utilizado o teste de Durbin-Watson (DW), cuja estatística é calculada por:

  $$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

  O valor de DW varia entre 0 e 4. Valores próximos de 2 indicam independência, enquanto valores próximos de 0 ou 4 sugerem autocorrelação positiva ou negativa, respectivamente.

- **Normalidade dos resíduos:** Para avaliar a normalidade dos resíduos, será aplicado o teste de Shapiro-Wilk, que testa a hipótese nula de que os resíduos seguem uma distribuição normal. A estatística W do teste é dada por:

  $$W = \frac{\left(\sum_{i=1}^{n} a_i e_{(i)}\right)^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2}$$

  onde $e_{(i)}$ são os resíduos ordenados, a_i são constantes dependentes da média e covariância de uma distribuição normal, e $\bar{e}$ é a média dos resíduos. Um p-valor maior que 0,05 (nível de significância de 5%) indica que os resíduos seguem uma distribuição normal.

- **Homogeneidade das variâncias (homocedasticidade):** A homogeneidade das variâncias dos resíduos será avaliada pelo teste de Breusch-Pagan, que verifica a hipótese nula de que a variância dos resíduos é constante. A estatística do teste é calculada por:

$$LM = \frac{1}{2\sigma^2} \left(\frac{\left(\sum_{i=1}^{n} e_i^2 \cdot Z_i \right)^2}{\sum_{i=1}^{n} e_i^4} \right)$$

  onde Z_i representa as variáveis explicativas e σ^2 é a variância estimada dos resíduos. Um p-valor menor que 0,05 indica a presença de heterocedasticidade (variância não constante).

2.4 Software Utilizado e Bibliotecas

Todas as análises foram realizadas utilizando o ambiente de programação R (*R Programming Language*)[4]. O R é uma linguagem de programação e um ambiente de software livre amplamente utilizado para estatística, ciência de dados e visualização. Ele oferece uma extensa coleção de pacotes e bibliotecas que permitem realizar análises avançadas e criar gráficos de alta qualidade. Além disso, sua ampla comunidade garante suporte contínuo, documentação detalhada e atualizações frequentes.

Os principais pacotes utilizados neste trabalho foram:

- `e1071`: Este pacote foi utilizado para realizar análises estatísticas avançadas, incluindo o cálculo de métricas como a assimetria e curtose, ele também fornece funções úteis para manipulação e exploração de dados [7].

- `lmtest`: Este pacote foi empregado para realizar testes estatísticos em modelos lineares. Em particular, ele oferece suporte para testes como o de Breusch-Pagan para verificar a homogeneidade das variâncias (heterocedasticidade) e o teste de Durbin-Watson para avaliar a autocorrelação dos resíduos [5].

O uso combinado dessas bibliotecas permitiu uma análise estatística robusta e reprodutível, bem como a geração de visualizações e relatórios claros e organizados. O código R foi estruturado para garantir reprodutibilidade, utilizando boas práticas de programação, como organização modular, comentários adequados e a criação de scripts parametrizados.

3 Resultados

Nesta seção mostraremos tosdos os resultados obtidos no codigo R. Inicialmente, verificamos a estrutura do conjunto de dados, que contém 124 observações e 4 variáveis: `ticker`, `date`, `average` e `volume`. Não foram identificados valores ausentes, garantindo a consistência e integridade dos dados.

[4] `https://www.r-project.org/`

```
> # Ler os dados
> dados <- read.csv("price_data_107.csv", header = TRUE, sep = ",", dec = ".")
>
> # Verificar a estrutura e existência de valores ausentes
> str(dados)
'data.frame':   124 obs. of  4 variables:
 $ ticker : chr  "PETR4" "PETR4" "PETR4" "PETR4" ...
 $ date   : chr  "2024-01-02" "2024-01-03" "2024-01-04" "2024-01-05" ...
 $ average: num  37.7 38.6 39 38.8 38 ...
 $ volume : num  9.06e+08 2.02e+09 1.77e+09 1.39e+09 1.34e+09 ...
> head(dados)
  ticker       date average     volume
1  PETR4 2024-01-02   37.66  905513838
2  PETR4 2024-01-03   38.56 2016962803
3  PETR4 2024-01-04   38.99 1768450856
4  PETR4 2024-01-05   38.79 1388296899
5  PETR4 2024-01-08   38.00 1336206062
6  PETR4 2024-01-09   38.25 1043653479
```

Figure 3: Amostra dos dados

3.1 Correlação entre `average` e `volume`

O teste de correlação de Pearson foi realizado para avaliar a relação linear entre `average` e `volume`. O coeficiente de correlação calculado foi $r = -0,053$, indicando uma correlação linear extremamente fraca e negativa. Essa fraqueza na correlação sugere que mudanças nos valores de `average` não estão associadas a mudanças proporcionais em `volume`.

O teste de hipótese associado apresentou um p-valor de 0,556, sendo superior ao nível de significância usual de 5% ($p > 0,05$). Assim, não rejeitamos a hipótese nula de que a correlação verdadeira entre as variáveis seja zero ($H_0 : \rho = 0$).

Além disso, o intervalo de confiança para o coeficiente de correlação, com nível de confiança de 95%, foi estimado como $[-0,228, 0,124]$. Esse intervalo inclui o valor zero, corroborando a ausência de evidências de uma relação linear significativa entre as variáveis.

```
> # Teste de correlação
> cor_test <- cor.test(average, volume)
> print(cor_test)

        Pearson's product-moment correlation

data:  average and volume
t = -0.59043, df = 122, p-value = 0.556
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2275540  0.1241056
sample estimates:
        cor
-0.05337903
```

Figure 4: Correlação de Pearson

3.2 Modelo de regressão linear

Foi ajustado um modelo de regressão linear simples, em que a variável dependente é `volume` e a variável explicativa é `average`. Os resultados do modelo indicam o seguinte:

- **Intercepto ($\hat{\beta_0}$):** O valor estimado do intercepto foi $\hat{\beta_0} = 2,640 \times 10^9$, com erro padrão de $1,760 \times 10^9$. O teste de hipótese associado ($H_0 : \beta_0 = 0$) não foi significativo ($t = 1,50, p = 0,136$), indicando que o intercepto não é estatisticamente diferente de zero ao nível de significância de 5%.
- **Coeficiente de inclinação ($\hat{\beta_1}$):** O coeficiente da variável `average` foi estimado como $\hat{\beta_1} = -2,669 \times 10^7$, com erro padrão de $4,520 \times 10^7$. O teste de hipótese para o coeficiente ($H_0 : \beta_1 = 0$) também não foi significativo ($t = -0,59, p = 0,556$), sugerindo que `average` não tem uma relação linear significativa com `volume`.

O coeficiente de determinação (R^2) foi calculado como 0,0028, indicando que apenas 0,28% da variabilidade total de `volume` é explicada pela variável `average`. O R^2 ajustado foi negativo ($R^2_{ajustado} = -0,0053$), sugerindo que o modelo não melhora a explicação da variabilidade dos dados em relação ao uso apenas da média de `volume`.

O teste F global do modelo ($H_0 : \beta_1 = 0$) também não foi significativo ($F(1, 122) = 0,349, p = 0,556$), reforçando a falta de capacidade do modelo de explicar a variabilidade dos dados. O erro padrão dos resíduos foi estimado em $\sigma = 1,053 \times 10^9$, um valor elevado, que reflete a grande dispersão dos dados ao redor da linha de regressão ajustada.

```
> # Ajustar o modelo de regressão linear
> modelo <- lm(volume ~ average)
> print(summary(modelo))

Call:
lm(formula = volume ~ average)

Residuals:
       Min         1Q     Median         3Q        Max
-1.152e+09 -5.422e+08 -2.296e+08  1.420e+08  6.659e+09

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2640434043 1760274557    1.50    0.136
average      -26688482   45201585   -0.59    0.556

Residual standard error: 1.053e+09 on 122 degrees of freedom
Multiple R-squared:  0.002849,  Adjusted R-squared:  -0.005324
F-statistic: 0.3486 on 1 and 122 DF,  p-value: 0.556
```

Figure 5: Reta de regressão pelo R

Na Figura 6, apresentamos o diagrama de dispersão entre `average` e `volume`, acompanhado da linha de regressão ajustada. A análise visual confirma a fraca relação entre as variáveis, consistente com os resultados numéricos:

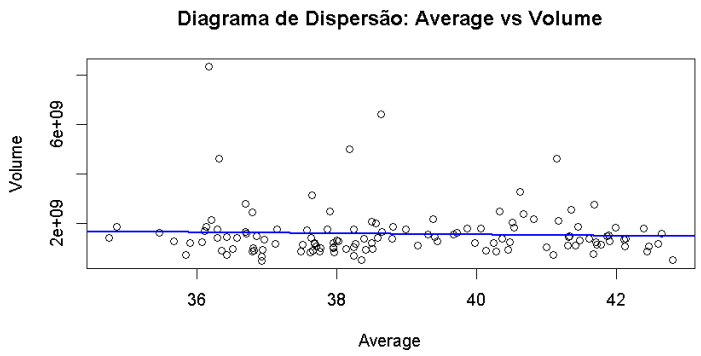

Figure 6: Diagrama de dispersão entre `average` e `volume`, com linha de regressão ajustada.

3.3 Análise de Variância (ANOVA)

A análise de variância (ANOVA) do modelo de regressão linear revelou que a soma dos quadrados explicada pela variável independente (`average`) é $3{,}8653 \times 10^{17}$, enquanto a soma dos quadrados residuais é $1{,}3527 \times 10^{20}$. A razão entre as variâncias (valor F) foi 0,3486, com um p-valor de 0,556. Esses resultados indicam que a variabilidade explicada pelo modelo é insignificante em relação à variabilidade residual ($p > 0{,}05$). Isso reflete a inadequação do modelo para capturar uma relação significativa entre `average` e `volume`, reforçando que a variável independente não contribui de maneira relevante para explicar as variações na variável dependente.

```
> # Análise de Variância (ANOVA) do modelo
> anova_modelo <- anova(modelo)
> print(anova_modelo)
Analysis of Variance Table

Response: volume
           Df     Sum Sq    Mean Sq F value Pr(>F)
average     1 3.8653e+17 3.8653e+17  0.3486  0.556
Residuals 122 1.3527e+20 1.1088e+18
```

Figure 7: Análise da tabela de variância.

3.4 Análise dos resíduos

Os resíduos do modelo foram avaliados para verificar as suposições. A Tabela 1 apresenta as estatísticas descritivas dos resíduos:

Table 1: Estatísticas descritivas dos resíduos do modelo ajustado.

Métrica	Valor
Mínimo	-1.152×10^9
Primeiro Quartil (Q1)	$-5,422 \times 10^8$
Mediana	$-2,296 \times 10^8$
Terceiro Quartil (Q3)	$1,420 \times 10^8$
Máximo	$6,659 \times 10^9$
Média	$4,284 \times 10^{-8}$
Desvio Padrão (DP)	$1,049 \times 10^9$
Assimetria	3,488
Curtose	15,847

O histograma dos resíduos (Figura 8) e o gráfico QQ-plot (Figura 10) indicam uma distribuição assimétrica com curtose elevada. O teste de Shapiro-Wilk rejeitou a hipótese nula de normalidade ($W = 0,656, p < 0,05$).

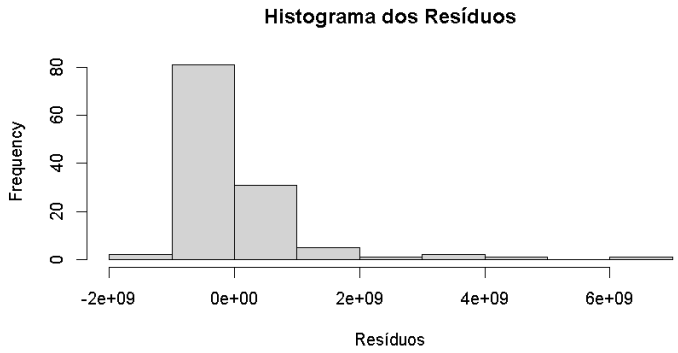

Figure 8: Histograma de Resíduos

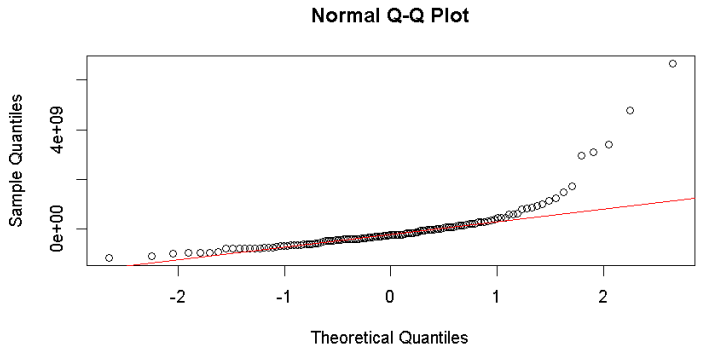

Figure 9: Normal Q-Q Plot

3.5 Testes de Diagnóstico

Foram realizados os seguintes testes para avaliar a adequação do modelo de regressão linear ajustado:

- **Autocorrelação dos resíduos:** O teste de Box-Ljung apresentou $X^2 = 20,161$ com $p < 0,05$, indicando a presença de autocorrelação significativa nos resíduos. Este resultado sugere que os resíduos não são independentes, o que pode comprometer a validade do modelo.

- **Normalidade dos resíduos:** O teste de Shapiro-Wilk apresentou $W = 0,6556$ com $p < 0,05$, indicando que os resíduos não seguem uma distribuição normal. Essa violação dos pressupostos do modelo pode afetar a interpretação dos resultados estatísticos.

- **Heterocedasticidade:** O teste de Breusch-Pagan apresentou $BP =$

$1,651$ com $p = 0,199$, indicando a ausência de heterocedasticidade significativa. Portanto, as variâncias dos erros podem ser consideradas homogêneas.

Os resultados dos testes indicam que, embora não haja evidências de heterocedasticidade, os pressupostos de independência e normalidade dos resíduos não foram atendidos. A autocorrelação dos resíduos sugere que pode ser necessário incluir outras variáveis explicativas ou utilizar modelos alternativos para capturar melhor a estrutura dos dados. Além disso, a violação da normalidade reforça que o modelo atual não se ajusta bem aos dados, limitando a sua capacidade preditiva e interpretativa.

```
> # Testes de diagnóstico
> print(Box.test(residuos, lag = 2, type = "Ljung-Box"))

        Box-Ljung test

data:  residuos
X-squared = 20.161, df = 2, p-value = 4.189e-05

> print(shapiro.test(residuos))

        Shapiro-Wilk normality test

data:  residuos
W = 0.6556, p-value = 1.195e-15

> print(bptest(modelo))

        studentized Breusch-Pagan test

data:  modelo
BP = 1.6511, df = 1, p-value = 0.1988
```

Figure 10: Diagnósticos Teste

4 Conclusão

Neste trabalho, utilizamos um modelo de regressão linear simples para avaliar a relação entre as variáveis `average` e `volume`, considerando um conjunto de dados composto por 124 observações sem valores ausentes. A metodologia aplicada incluiu a análise de correlação, ajuste do modelo de regressão, análise de variância (ANOVA) e testes de diagnóstico para verificar a adequação do modelo e o atendimento dos pressupostos.

Os resultados obtidos mostraram uma relação linear extremamente fraca e não significativa entre as variáveis analisadas, conforme evidenciado pelo coeficiente de correlação de Pearson ($r = -0,053$) e pelo coeficiente de determinação ($R^2 = 0,0028$). O teste F global e a análise de variância reforçaram a ausência

de significância do modelo, indicando que a variável independente `average` explica menos de 1% da variabilidade na variável dependente `volume`. Além disso, a análise dos resíduos revelou violações importantes nos pressupostos do modelo, como a presença de autocorrelação significativa e a falta de normalidade, embora não tenha sido identificada heterocedasticidade.

Diante da inadequação do modelo, sugerimos que futuros trabalhos explorem abordagens alternativas. A inclusão de variáveis explicativas adicionais pode melhorar a capacidade preditiva do modelo, enquanto transformações nos dados podem corrigir violações nos pressupostos. Modelos não-lineares ou técnicas mais robustas, como modelos de regressão com componentes autorregressivos ou aprendizado de máquina, podem ser alternativas promissoras para capturar melhor as relações subjacentes nos dados. Além disso, uma análise mais aprofundada da estrutura temporal dos dados pode ser relevante, dado o padrão de autocorrelação identificado nos resíduos.

Assim, este estudo evidencia a importância de avaliar criticamente os pressupostos e resultados de modelos estatísticos, bem como de considerar métodos complementares para melhorar a compreensão das relações entre variáveis em contextos complexos.

References

[1] Bodie, Z., Kane, A., Marcus, A.J.: Investments. McGraw-Hill Education, 11th edn. (2018)

[2] Cardoso, F.C., Malska, J.A.V., Ramiro, P.J., Lucca, G., Borges, E.N., de Mattos, V.L.D., Berri, R.A.: Bovdb: a data set of stock prices of all companies in b3 from 1995 to 2020. Journal of Information and Data Management **13**(1) (2022)

[3] Fabozzi, F.J., Focardi, S.M., Kolm, P.N.: Financial Modeling of the Equity Market: From CAPM to Cointegration. Wiley, 1st edn. (2014)

[4] de Freitas Parreiras, L., da Silva, R., Sinhorino, S., Ouki, P., Beraha, A., da Silveira, A., de Moraes, D., Granja, S., Alves, A., Kamakura, A.: Arbitragem estatística e inteligência artificial. Ph.d. thesis, Universidade de São Paulo (2007)

[5] Hothorn, T., Zeileis, A.: lmtest: Testing Linear Regression Models (2023), `https://cran.r-project.org/package=lmtest`, r package version 0.9-40

[6] Mattos, V.L.D.; Konrath, A.C.A.A.V.: Introdução à estatística: aplicações em ciências exatas. Editora LTC (1) (2017)

[7] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien (2024), `https://CRAN.R-project.org/package=e1071`, r package version 1.7-16

Apêndice A: Código R Utilizado na Pesquisa

```r
# Bibliotecas necessarias
library(e1071)
library(lmtest)

# Configurar o diretorio de trabalho
setwd("C:/bovdb/docs/rstudio/data")

# Ler os dados
dados <- read.csv("price_data_107.csv", header = TRUE, sep = ",",
    ↪ dec = ".")

# Verificar a estrutura e existencia de valores ausentes
str(dados)
head(dados)
if (any(is.na(dados))) {
  stop("O dataset contem valores ausentes!")
}

# Definir variaveis
average <- dados$average
volume <- dados$volume

# Teste de correlacao
cor_test <- cor.test(average, volume)
print(cor_test)

# Ajustar o modelo de regressao linear
modelo <- lm(volume ~ average)
print(summary(modelo))

# Diagrama de dispersao com linha de regressao
plot(average, volume,
     main = "Diagrama de Dispersao: Average vs Volume",
     xlab = "Average",
     ylab = "Volume",
     pch = 1)
abline(modelo, col = "blue", lwd = 2)

# Teste da significancia dos coeficientes do modelo
summary_coef <- summary(modelo)$coefficients
print(summary_coef)

# Analise de Variancia (ANOVA) do modelo
anova_modelo <- anova(modelo)
print(anova_modelo)

# Coeficiente de Determinacao (R  )
r_squared <- summary(modelo)$r.squared
adj_r_squared <- summary(modelo)$adj.r.squared
cat("R  :", r_squared, "\n")
cat("R  Ajustado:", adj_r_squared, "\n")

# Analise dos residuos
residuos <- residuals(modelo)
```

```r
plot(residuos, main = "Residuos do Modelo", ylab = "Residuos", xlab
    ↪  = "Indice")
hist(residuos, main = "Histograma dos Residuos", xlab = "Residuos")
qqnorm(residuos)
qqline(residuos, col = "red")

# Estatisticas dos residuos
resumo_residuos <- list(
  Min = min(residuos),
  Q1 = quantile(residuos, 0.25),
  Mediana = median(residuos),
  Q3 = quantile(residuos, 0.75),
  Max = max(residuos),
  Media = mean(residuos),
  DP = sd(residuos),
  Assimetria = skewness(residuos),
  Curtose = kurtosis(residuos)
)
print(resumo_residuos)

# Testes de diagnostico
print(Box.test(residuos, lag = 2, type = "Ljung-Box"))
print(shapiro.test(residuos))
print(bptest(modelo))
```

Listing 1: Código R para análise de Correlção e Regressão Linear