UNIVERSIDADE FEDERAL DO CEARÁ

UNIVERSIDADE ABERTA DO BRASIL

INSTITUTO UFC VIRTUAL

CURSO DE LICENCIATURA EM MATEMÁTICA

Eliude Angelo Carneiro Neto

Importância da Regressão Linear para Análise de Dados

Caucaia -CE

2024

Eliude Angelo Carneiro Neto

Importância da regressão Linear para Análise de Dados

Trabalho de Conclusão apresentado ao Curso de Licenciatura em Matemática Semipresencial do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Licenciado em Matemática.

Orientador: Prof.: Me. Breno Rafael Pinheiro Sampaio.

Caucaia-CE

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

N385i Neto, Eliude Angelo Carneiro.
Importância da Regressão Linear para Análise de Dados / Eliude Angelo Carneiro Neto. – 2024.
95 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Instituto UFC Virtual, Curso de Matemática, Fortaleza, 2024.
Orientação: Prof. Me. Breno Rafael Pinheiro Sampaio.

1. Regressão Linear . 2. Análise de Dados. 3. Regressão Linear Múltipla . I. Título.
CDD 510

Eliude Angelo Carneiro Neto

Importância da regressão Linear para Análise de Dados

Trabalho de Conclusão apresentado ao Curso de Licenciatura em Matemática Semipresencial do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Licenciado em Matemática.

Aprovada em: 03/07/2024

BANCA EXAMINADORA

_____

Prof.: Me. Breno Rafael Pinheiro Sampaio (Orientador)
Universidade Federal do Ceará (UFC)

_____

Prof.: Me. Clodomir Silva Lima Neto
Instituto Federal do Ceará (IFCE)

_____

Prof.: Dr. Miguel Angelo da Silva
Universidade Estadual do Ceará (UFC)

A Deus.

Aos meus pais, Nataniel Soares Carneiro e Samia Soares Carneiro.

AGRADECIMENTOS

À Universidade Federal do Ceará, em específico ao Instituto UFC Virtual, por me proporcionar a oportunidade da minha formação.

Ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (Capes) pelo programa Universidade Aberta do Brasil.

Ao Prof. Breno Sampaio, pela excelente orientação.

Aos professores participantes da banca examinadora, Me. Clodomir Silva Lima Neto e Dr. Miguel Angelo da Silva pelo tempo, pelas valiosas colaborações e sugestões.

Aos colegas da turma de graduação, em especial ao Tiago Santana e Ana Paula, pelas reflexões, críticas e sugestões recebidas.

Agradeço também a todos aqueles que, de alguma forma, colaboraram para a concretização deste projeto, mesmo que não mencionados nominalmente, cada contribuição foi fundamental para o seu êxito.

“Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer.” Alan Mathison Turing.

RESUMO

O objetivo deste trabalho é destacar a importância da regressão linear na análise de dados, uma técnica estatística fundamental em diversos campos de estudo. A regressão linear permite modelar a relação entre variáveis dependente e independentes, facilitando a previsão de resultados e a identificação de tendências. Por meio de ferramentas como o coeficiente de determinação (R^2), gráficos Q-Q, gráficos de resíduos versus valores preditos, e outros métodos de diagnóstico, é possível avaliar a qualidade do ajuste do modelo e a adequação dos dados bem como por eles ter insights.

O estudo também explora a aplicação prática da regressão linear, demonstrando como essa técnica pode ser utilizada para transformar dados em insights úteis, além disso, o trabalho discute as limitações e suposições da regressão linear, como a linearidade, a homoscedasticidade, e a independência dos erros, já que compreender essas suposições é essencial para a aplicação mais consistente da técnica e para evitar conclusões equivocadas.

Palavras-chave: Regressão Linear; Análise de Dados; Regressão Linear Múltipla.

ABSTRACT

The objective of this paper is to highlight the importance of linear regression in data analysis, a fundamental statistical technique in several fields of study. Linear regression allows modeling the relationship between dependent and independent variables, facilitating the prediction of results and the identification of trends. Using tools such as the coefficient of determination (R^2), Q-Q plots, residual versus predicted value plots, and other diagnostic methods, it is possible to assess the quality of the model's fit and the adequacy of the data, as well as to gain insights from them.

The study also explores the practical application of linear regression, demonstrating how this technique can be used to transform data into useful insights. In addition, the paper discusses the limitations and assumptions of linear regression, such as linearity, homoscedasticity, and the independence of errors, since understanding these assumptions is essential for the most consistent application of the technique and to avoid erroneous conclusions.

Keywords: Linear Regression; Data Analysis; Multiple Linear Regression.

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

TI	Tecnologia da Informação
IDC	*Internacional Data Corporation*
IBM	*International Business Machines Corporation*
ESPM	Escola Superior de Propaganda e Marketing
SQT	Soma dos Quadrados Totais
SQR	Soma dos Quadrados da Regressão
SQU	Soma dos Quadrados dos Resíduos
Q-Q *Plot*	Gráfico Quantil-Quantil
Q1	1° Quartil
Q2	Mediana
Q3	3° Quartil
AIQ	Amplitude interquartil
3D	Três Dimensões
ANOVA	*Analysis of Variance*
Max_Regressão	Valor Máximo da Regressão
Min_Regressão	Valor Mínimo da Regressão
Val_Regressão	Valor da Regressão
Medio_Regressão	Valor Médio da Regressão

LISTA DE SÍMBOLOS

β_0	Intercepto ou coeficiente linear;
β_1	Coeficiente angular da reta;
$\hat{\beta}$	Estimativa do coeficiente β obtida a partir da amostra
β_i	Coeficiente de x_i
$\hat{\beta}_i$	Estimativa do coeficiente $\beta_i$obtida a partir da amostra.
ε	Termo de erro que representando a variação não explicada pelo modelo.
x_i	Valores observados da variável independente para cada i
$\overline{x}$	Média dos valores observados da variável independente
y_i	Valores observados da variável independente para cada i
$\hat{y}_i$	Valor previsto pela regressão para a i-ésima observação
$\overline{y}$	Média dos valores observados da variável dependente
ε_i	Resíduo da i-ésima observação;
$cov(x, y)$	Covariância de x e y
σ_x	Desvio padrão de x
σ_y	Desvio padrão de y
r	Coeficiente de Correlação
R^2	Coeficiente de Determinação
$R^2_{AJUSTADO}$	Coeficiente de Determinação Ajustado
$SE(x)$	Erro padrão de x
$t_{n-2,\alpha/2}$	Valor crítico da distribuição t de Student em uma regressão linear simples onde $n-2$ são os graus de liberdade e $\alpha/2$ é o nível de significância a média de Y
$VAR(\beta_i)$	Matriz de variância-covariância dos estimadores,
$t_{\frac{\alpha}{2},n-k-1}$	Valor crítico da distribuição t de Student em uma regressão linear simples onde $n-$k -1 são os graus de liberdade sendo n o número de observações, k o número de variáveis explicativas e $\alpha/2$ é o nível de significância

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da tecnologia da informação (TI), ela se tornou essencial para a sociedade moderna, influenciando vários aspectos da nossa vida. O avanço da TI tem grande impacto em diversos setores, trazendo consigo benefícios e desafios que precisam ser cuidadosamente considerados.

Com sua presença quase onipresente desde a forma como nos comunicamos até a maneira como trabalhamos e consumimos bens e serviços a quantidade de dados gerados por essa interação amentou drasticamente, segundo o relatório "Data Age 2025" da Internacional Data Corporation (IDC), em 2015, cada pessoa gerou cerca de 1,3 GB de dados todo dia. E esse número foi projetado para crescer ainda mais. A previsão era que, em 2025, cada pessoa estivesse gerando 5,3 GB de dados por dia e segundo a International Business Machines Corporation (IBM), diariamente, o mundo gera cerca de 2,5 quintilhões de dados. E 90% dos dados disponíveis hoje foram gerados nos últimos 3 anos.

Com isso surgiu um problema, o que fazer com essa quantidade enorme de dados? Desse problema surgiu o Big Data, segundo a Escola Superior de Propaganda e Marketing (ESPM): Em tecnologia, o conceito de Big Data é aplicado a uma área que coleta e analisa informações diversas a partir de um grande volume de dados.

A era do Big Data revolucionou a maneira como as empresas coletam, armazenam e analisam informações. Com tantos dados a capacidade de extrair insights a partir desses dados tornou-se uma "arma" crucial na competição comercial. Nesse contexto, a regressão linear surge como uma ferramenta estatística essencial para a análise de dados.

A regressão linear é uma técnica usada para compreender a relação entre variáveis e fazer previsões solidas. E este é o conteúdo que será abordado nesse trabalho: A Importância da Regressão Linear para Análise de Dados.

1.1 Objetivo

O objetivo deste trabalho é explorar detalhadamente o uso eficaz da regressão linear para extrair insights cruciais na análise de dados, a fim de informar decisões estratégicas. Serão abordados os benefícios dessa técnica, como sua simplicidade e a interpretabilidade dos resultados, bem como suas limitações, incluindo sensibilidade a outliers e a suposição de linearidade. Este estudo visa proporcionar uma visão abrangente de como a regressão linear pode contribuir significativamente para análises de dados robustas e decisões embasadas. O objetivo específico é analisar os fundamentos teóricos da regressão linear

simples e múltipla, explorando os conceitos matemáticos subjacentes e suas aplicações práticas na análise de dados

1.2 Metodologia

A metodologia deste trabalho tem como objetivo descrever os métodos utilizados para a pesquisa e desenvolvimento do conteúdo referente à importância da regressão linear para análise de dados. O estudo foi realizado com base em pesquisa bibliográfica e análise de dados.

A partir desse objetivo geral será empregada uma pesquisa exploratória, em que além da consulta a fontes bibliográficas, será feita uma análise e verificação se de fato os elementos usados servem para fazer inferências.

No que se refere à pesquisa qualitativa foi feito uma análise documental onde foram selecionados livros de estatística e análise de dados reconhecidos que discutem o uso da regressão linear afim de identificar padrões e temas recorrentes sobre a eficácia e sua aplicação, como os de Fávero e Belfiore (2014), Triola (2017) assim como Chein (2019); além de matérias online, como os doa blogs Psicometria online e Medium; artigo e anotações do Dr. Iain Pardoe.

Já na pesquisa quantitativa foram analisados dados reais obtido do kaggle, assim como dados aleatórios gerados pelo autor, a fim de realizar análises de regressão linear simples e múltipla, demonstrando como essas técnicas ajudam a identificar relações entre variáveis.

Segundo Yin (2001), o estudo de caso é útil quando o se deseja explorar um fenômeno dentro de seu contexto real, permitindo uma investigação detalhada e contextualizada. E o estudo de caso foi o instrumental. O estudo de caso seguiu as seguintes etapas:

1. Definição do Caso:

- Casos Escolhidos: Seleção de exemplos de livros e do kaggle que demonstram o uso da regressão linear.
- Objetivo: Demonstrar como a regressão linear pode ser aplicada para obter insights e apoiar a tomada de decisões.

2. Coleta de Dados:

- Fontes: Dados extraídos de exemplos em livros de metodologia ou estatística, dados do autor e kaggle.

3. Análise de Dados:

- Aplicação da Regressão Linear: Realização de análises de regressão linear nos casos selecionados.
- Interpretação dos Resultados: Interpretação dos resultados para extrair insights e demonstrar a utilidade da técnica.

4. Interpretação e Discussão:

- Discussão dos Benefícios: Explicação de como a regressão linear ajudou a entender melhor os dados e apoiar a tomada de decisões.
- Limitações: Discussão das limitações encontradas nos exemplos e como essas limitações podem ser tratadas em outros contextos.

2 FUNDAMENTO TEORICO

O proposito deste capítulo é apresentar a teoria que se fundamenta a regressão linear simples e múltipla, assim como também o fundamento de algumas ferramentas para auxiliar a extração de informações uteis da regressão.

2.1 O que é regressão linear simples

Segundo Triola (2017), a regressão linear simples é um modelo de análise que usamos quando modelamos a relação linear entre uma variável de dependente e uma variável independente.

2.2 Modelo de regressão linear simples

Como aponta Chein (2019), a equação geral para uma regressão linear simples é: $y = \beta_0 + \beta_1 \cdot x + \varepsilon$

Onde:

- y é a variável dependente;
- x é a variável independente;
- β_0é o intercepto ou coeficiente linear;
- β_1 é o coeficiente angular da reta;
- ε é o termo de erro que representando a variação não explicada pelo modelo.

2.3 Cálculo dos coeficientes de uma regressão linear simples

Segundo Chein (2019), os coeficientes de uma regressão linear são os valores que descrevem a relação linear entre duas variáveis, esses coeficientes são calculados usando técnicas como a dos mínimos quadrados ordinários, onde a linha de regressão é ajustada de modo a minimizar a soma dos quadrados das diferenças entre os valores observados e os valores previstos pela linha de regressão, em outras palavras na soma dos quadrados das diferenças entre os valores observados e os valores previstos pela linha deve ser a menos possível, logo teremos asseguintes fórmulas dos coeficientes β_0 e β_1 :

- $\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
- $\beta_0 = \overline{x} - \beta_1 \cdot \overline{y}$

Com:

- x_i são os valores observados da variável independente para cada i
- $\overline{x}$ é a média dos valores observados da variável independente
- y_i são os valores observados da variável independente para cada i
- $\overline{y}$ é a média dos valores observados da variável dependente

Exemplo:

x = {1, 2, 3, 4 e 5}

y = {2.5, 3.5, 4, 5 e 5.5}

Tabela 1 – Valores para Cálculo da Regressão

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(y_i - \overline{y})(x_i - \overline{x})$	$(x_i - \overline{x})^2$
1	2.5	-2	-1.6	3.2	4
2	3.5	-1	-0.6	0.6	1
3	4	0	-0.1	0	0
4	5	1	0.9	0.9	1
5	5.5	2	1.4	2.8	4

Fonte: Dados hipotéticos do autor (2024)

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.75$$

$$\beta_0 = \overline{x} - \beta_1 \cdot \overline{y} = 1.85$$

2.4 Cálculo de Resíduos

Segundo Rebeca, Gómez (2004), a análise residual é uma técnica útil para a avaliação do ajuste do modelo, identificação de padrões ou tendências não capturadas pelo modelo, verificação dos pressupostos da regressão linear e identificação de outliers e observações influentes.

Conforme apontado por Chein (2019), os resíduos de uma regressão linear são as diferenças entre os valores observados da variável dependente e os valores previstos, para calcularmos (sendo regressão linear simples ou múltipla) usaremos seguinte fórmula é: $\varepsilon_i = y_i - \hat{y}_i$ sendo:

- ε_i sendo o resíduo da i-ésima observação;
- y_i o valor observado da variável dependente para a i-ésima observação;

- $\hat{y}_i$ valor previsto pela regressão para a i-ésima observação.

Exemplo:

x = {1, 2, 3, 4 e 5}

y = {2.5, 3.5, 4, 5 e 5.5}

$\hat{y}_i = 1.85 + 0{,}75x_i$

Tabela 2 – Valores para Cálculo dos Resíduos

x_i	y_i	$\hat{y}_i$	Resíduos (ε_i)
1	2.5	3.60	-0.10
2	3.5	4.35	0.15
3	4	5.10	-0.10
4	5	5.85	0.15
5	5.5	6.60	-0.10

Fonte: Dados hipotéticos do autor (2024)

2.5 Pressupostos de uma regressão linear simples

De acordo com Chein (2019), os pressupostos da regressão múltipla servem como condições fundamentais que devem ser atendidas para garantir que os resultados obtidos a partir da análise sejam válidos e confiáveis, esses pressupostos são:

Linearidade: a relação entre as variáveis dependente e independentes deve ser linear;

Independência de erro: os erros de previsão devem ser independentes um do outro;

Homoscedasticidade: a variância dos erros deve ser constante em relação às variáveis independentes;

Normalidade dos Erros ($E[\varepsilon_i|X_1, \ldots, X_K)$: os erros de previsão devem ser distribuídos normalmente;

O valor esperado do erro é zero: os resíduos devem ter um valor médio de zero.

Iremos agora abordar "o porquê" de elas serem tão importantes

2.5.1 Linearidade

Por se tratar de um modelo em que queremos estabelecer ou a função de uma reta ou de um plano ou de um hiperplano é essencial supor que exista linearidade afim de quer os dados estejam o mais próximo possível deles

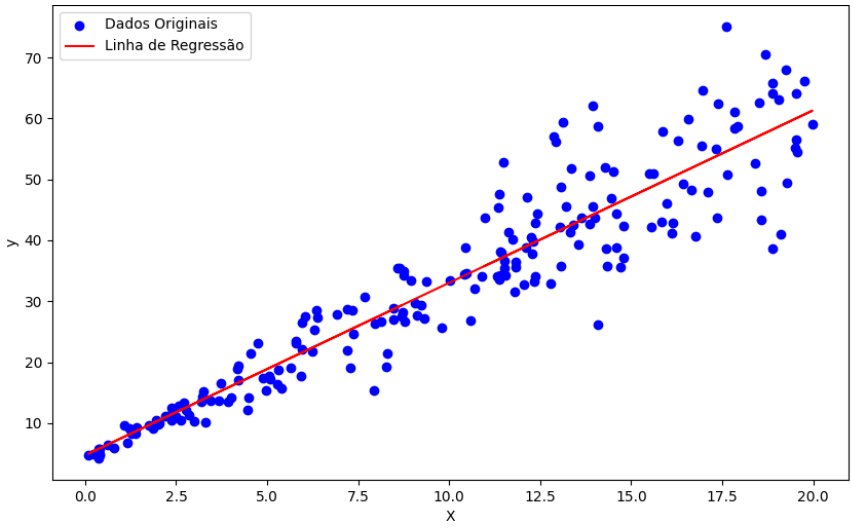
2.5.2 Independência de erro

Segundo Chein (2019), a independência do erro é uma suposição fundamental e implica que os erros de previsão (ou resíduos) não estão correlacionados entre si, logo, não há padrão perceptível nos resíduos quando são plotados contra os valores ajustados. Sua violação pode distorcer os resultados, pois os intervalos de confiança e os testes de significância podem ser inválidos, levando a conclusões errôneas sobre a relação entre as variáveis. Como essa violação afetá-las será explicada mais à frente.

2.5.3 Homoscedasticidade

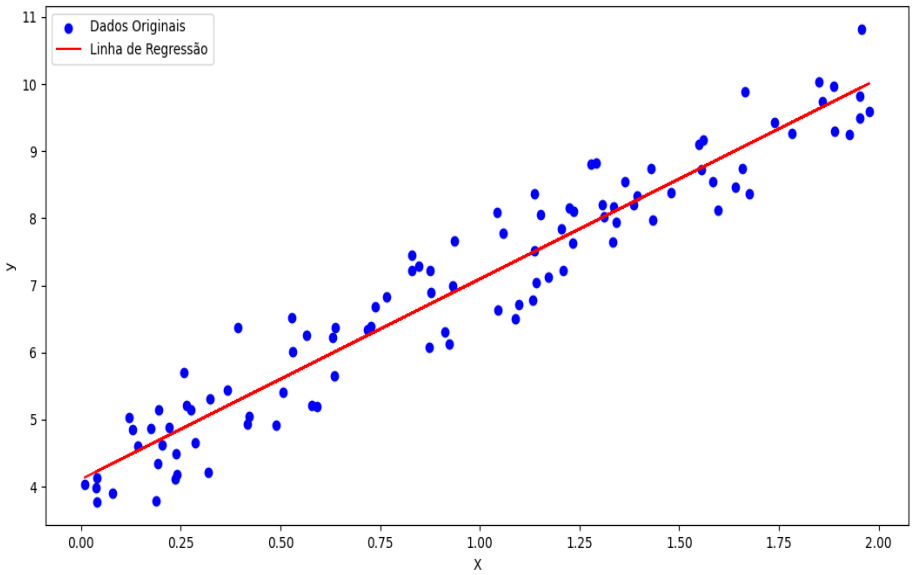
De acordo com Chein (2019), homoscedasticidade é de extrema importância, pois refere-se à consistência da variância dos erros ao longo de todas as observações em um modelo de regressão. Perceba que se tivermos uma heterocedasticidade nossos resíduos não terão consistência e como resíduo é a diferença do valor dado pelo valor previsto teremos quanto mais variáveis eles forem piores, pois implica baixa previsibilidade.

Gráfico 1 - Regressão com Heterocedasticidade:



Fonte: Dados hipotéticos do autor (2024)

Gráfico 2 - Regressão com Homoscedasticidade:



Fonte: Fonte: Dados hipotéticos do autor (2024)

2.5.4 Normalidade dos erros

De acordo com Chein (2019) em "Introdução aos Modelos de Regressão Linear", a normalidade dos erros na regressão linear é uma suposição fundamental para que os testes estatísticos e os intervalos de confiança associados à análise sejam válidos. Como ela afetara-las será explicada adiantamento

2.5.5 O Valor esperado do erro é zero ($E[\varepsilon_i|X_1, \ldots, X_K)$

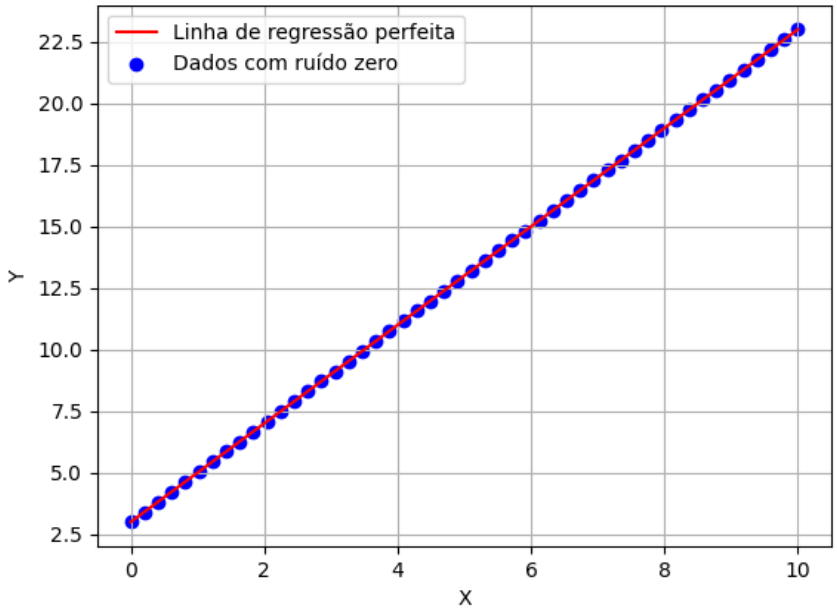
Conforme Chein (2019), Se o valor esperado (média) dos erros de previsão é zero, isso sugere que o modelo está capturando corretamente a relação entre as variáveis, isso é importante porque indica que o modelo não está sistematicamente superestimando ou subestimando os valores.

Para se calcular o valor esperado do erro na regressão linear é necessário entender que o valor esperado do erro é uma medida da precisão das previsões feitas pelo modelo de regressão e se calcula da seguinte forma: $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$

Se o valor esperado dos erros não for próximo de zero, isso indica que o modelo está sistematicamente errado em suas previsões. Por exemplo, se os resíduos têm uma média positiva, significa que o modelo está subestimando consistentemente os valores reais da variável dependente, semelhantemente, se os resíduos possuem uma média negativa, o

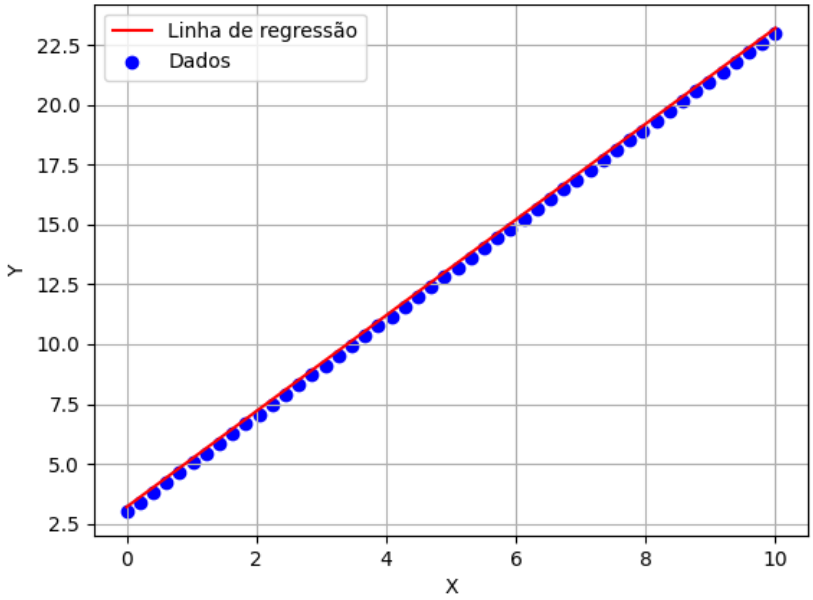
modelo está superestimando consistentemente os valores reais da variável dependente, isso é problemático porque significa que nossas previsões estão enviesadas. Se o modelo está consistentemente errado em uma direção, isso sugere que ele não está capturando adequadamente a relação entre as variáveis independentes e dependentes. Isso pode ser devido a uma série de razões, como a inclusão de variáveis irrelevantes, a exclusão de variáveis importantes ou a violação de pressupostos importantes da regressão linear. Perceba nos seguintes gráficos (sem ruido) o impacto quando ($E[\varepsilon_i|X_1, \dots, X_K) = 0$, ($E[\varepsilon_i|X_1, \dots, X_K) = 0{,}2$ e ($E[\varepsilon_i|X_1, \dots, X_K) = 5$. respectivamente. Vale mencionar que o valor esperado do erro ser 0 não implica em linearidade perfeita, nos exemplos a seguir haverá linearidade perfeita apenas para deixar mais claro a influência dele em uma regressão.

Gráfico 3 - Valor Esperado do Erro é 0:



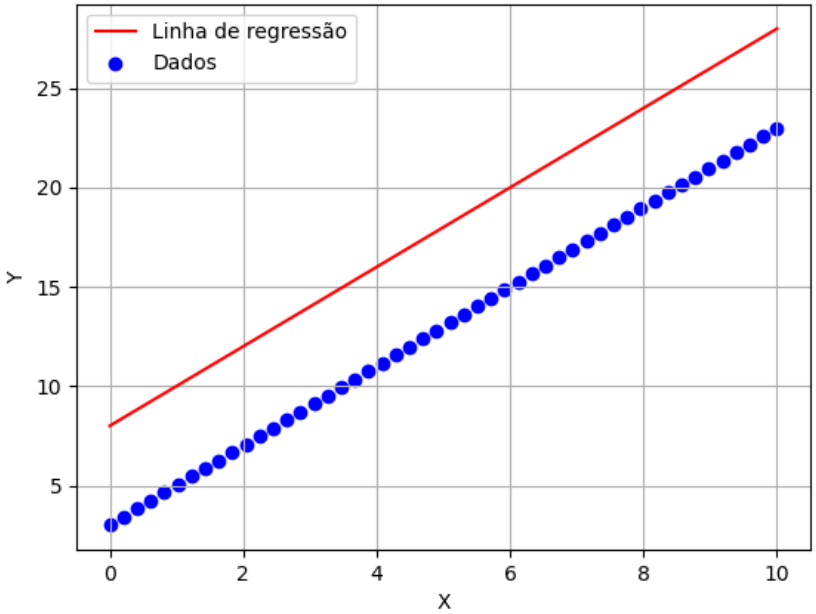
Fonte: Fonte: Dados hipotéticos do autor (2024)

Gráfico 4 - Valor Esperado do Erro é 0.2:



Fonte: Dados hipotéticos do autor (2024)

Gráfico 5 - Valor Esperado do Erro é 5:



Fonte: Fonte: Dados hipotéticos do autor (2024)

2.6 Coeficiente de correlação (Pearson (r))

Segundo Guimarães (2021), o coeficiente de correlação de Pearson (r) é uma medida adimensional que pode assumir valores no intervalo entre -1 e +1 que serve para calcular o grau de relação linear entre variáveis.

Sua fórmula é dada por: $p = \frac{\Sigma(x_i-\overline{y})(y_i-\overline{y})}{\sqrt{\Sigma(x_i-\overline{x})^2\Sigma(y_i-\overline{y})^2}} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$ com:

- x_i são os valores individuais da variável x
- $\overline{x}$média dos valores x
- y_i são os valores individuais da variável y
- $\overline{y}$ média dos valores y
- $cov(x,y)$ É a covariância de x e y. Consoante a Fávero e Belfiore (2017), o coeficiente de covariância varia de menos infinito a mais infinito, porém, ele é normalmente normalizado para estar entre -1 e 1, a fim de facilitar a interpretação, sendo -1 a covariância perfeita negativa os dois aumentam ou diminuem em direções opostas, na mesma proporção, 0 a ausência de covariância e 1 a covariância perfeita positiva, os dois aumentam ou diminuem juntas, na mesma proporção.
- σ_x é o desvio padrão de x
- σ_y é o desvio padrão de y

Com o coeficiente podemos definir se a correlação linear é perfeita (positiva ou negativa), muito forte, forte, moderada, fraca, muito fraca ou inexistente, veja a tabela:

Tabela 3 – Parâmetros do Coeficiente de Correlação

Valor do Coeficiente de Correlação	**Interpretação**
-1	Correlação negativa perfeita
]-1 a -0.90	Correlação negativa muito forte
-0.89 a -0.70	Correlação negativa forte
-0.69 a -0.40	Correlação negativa moderada
-0.39 a -0.20	Correlação negativa fraca
-0.19 a 0.00[	Correlação negativa muito fraca
0.00	Correlação nula ou inexistente
]0.00 a 0.19	Correlação positiva muito fraca
0.20 a 0.39	Correlação positiva fraca
0.40 a 0.69	Correlação positiva moderada
0.70 a 0.89	Correlação positiva forte
0.90 a +1[	Correlação positiva muito forte
+1	Correlação positiva perfeita

Fonte: Batista (2021)

Exemplo:

x = {1, 2, 3, 4 e 5}

y = {2.5, 3.5, 4, 5 e 5.5}

$\hat{y}_i = 1.85 + 0{,}75x_i$

Tabela 4 – Valores para o Cálculo do Coeficiente de Correlação

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(y_i - \overline{y})(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2$
1	2.5	-2	-1.6	3.2	4	2.56
2	3.5	-1	-0.6	0.6	1	0.36
3	4	0	-0.1	0	0	0.01
4	5	1	0.9	0.9	1	0.81
5	5.5	2	1.4	2.8	4	1.96

Fonte: Dados hipotéticos do autor (2024)

$$p = \frac{\Sigma(x_i - \overline{y})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \Sigma(y_i - \overline{y})^2}} = 0.992$$

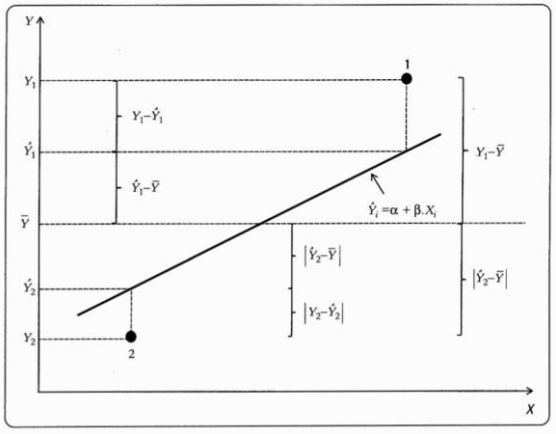
2.7 $Coeficiente de determinação$ (R^2)

Segundo França (2013), o coeficiente de determinação é uma medida estatística que é usado para avaliar a qualidade do ajuste de um modelo de regressão, ela é uma métrica que varia de 0 a 1. Quando R^2 é igual a 0, isso significa que o modelo não explica nada da variação dos dados, enquanto um R^2 igual a 1 indica que o modelo explica toda a variação dos dados. Em outras palavras, o R^2 representa a porcentagem da variação nos dados que é explicada pelo modelo. Para construirmos a fórmula necessárias para encontrar a expressão para R^2 usaremos a seguinte expressão: $SQT = SQR + SQU$ onde:

- SQT é a soma total dos quadrados, mostra a variação em Y em torno da própria média;
- SQR é a soma dos quadrados da regressão, oferece a variação de Y considerando as variáveis X utilizadas no modelo;
- SQU é a soma dos quadrados dos resíduos, apresenta a variação de Y que não é explicada pelo modelo elaborado
- y_i é o valor observado
- $\hat{y}_i$ é o valor ajustado
- $\overline{y}$ é a média dos valores observados

Veja a seguinte imagem:

Gráfico 6 – Desvio de y para duas observações



Fonte: Fávero, Belfiore (2017, p.522)

Perceba que:

- $y_i - \hat{y}_i$ é a parte não explicada
- $y_i - \overline{y}$ é o total
- $\hat{y}_i - \overline{y}$ é a parte explicada

Logo, podemos chegar à seguinte conclusão $SQT = SQR + SQU \Rightarrow \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

R^2 é a fração da variância da amostra de Y; explicada (ou prevista) pelas variáveis explicativas. O R^2 é obtido da seguinte forma: $R^2 = \frac{SQR}{SQR+SQU} = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ também podemos fazer da seguinte forma: $SQT = SQR + SQU \Rightarrow SQR = SQT - SQU \Rightarrow \frac{SQR}{SQT} = \frac{SQT}{SQT} - \frac{SQU}{SQT} \Rightarrow R^2 = 1 - \frac{SQU}{SQT}$

Exemplo:

x = {1, 2, 3, 4 e 5}

y = {2.5, 3.5, 4, 5 e 5.5}

$\hat{y}_i$ = 1.85 + 0,75x_i

Tabela 5 –Valores para Cálculo do Coeficiente de Determinação

x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \overline{y})$
1	2.5	3.60	0.01	-1.6
2	3.5	4.35	0.0225	0.6
3	4	5.10	0.01	0.01
4	5	5.85	0.0225	0.9
5	5.5	6.60	0.01	1.4

Fonte: Dados hipotéticos do autor (2024)

$$R^2 = 1 - \frac{SQU}{SQT} = 1 - 0.0132 = 0.9868$$

2.8 $Erro \approx Resíduos$

- Erro é uma medida da discrepância entre os valores observados e o valor previsto não captado pelo modelo;

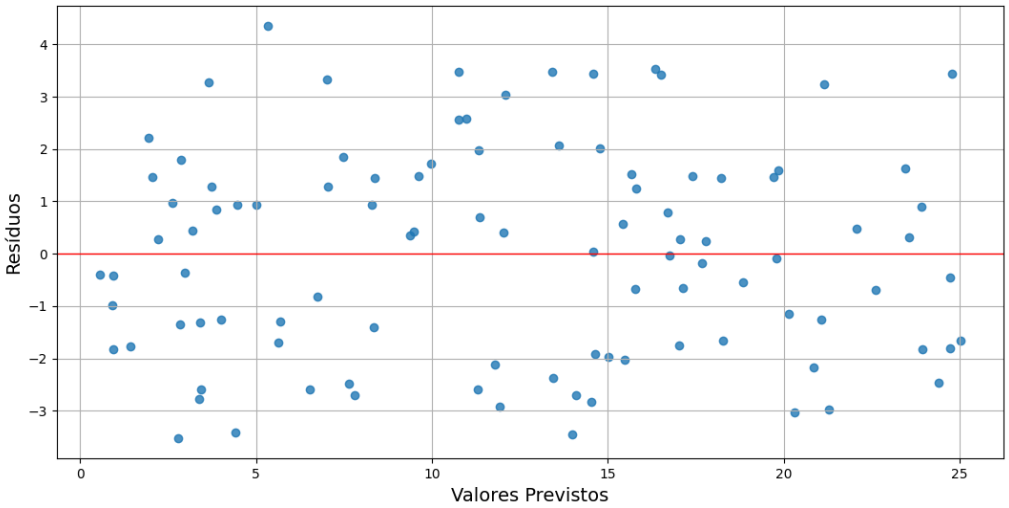
- Resíduo é diferença entre cada valor observado (y_i) e o valor previsto ($\hat{y}_i$).

Quando os resíduos são aproximadamente iguais aos erros para a maioria das observações, isso indica que o modelo está fazendo previsões precisas já que resíduo é a diferença entre valores observados e previstos se eles são constantes isso mostra que dados estão bem distribuídos em torno da reta e se são semelhantes ao erro ($Erro \approx Resíduos$) mostra que o erro explica esta distância dos pontos observados a os valores previstos, logo pode-se afirmar que o modelo está se ajustando bem aos dados. Vale mencionar que o "tamanho" da medida do erro pode mostrar que o modelo não está bem ajustado, logo o $Erro \approx Resíduos$ não implicará nele bem ajustado aos dados, necessitando primeiramente um bom ajuste no modelo para chegarmos à conclusão de que $Erro \approx Resíduos$ resulta em bom modelo. Vale mencionar que este ponto não se diferencia de uma regressão linear simples para múltipla.

2.9 Gráfico de Resíduos vs Valores Previstos

Segundo Perdoe, o gráfico de resíduos vs valores preditos é uma ferramenta comum na análise de regressão, ele mostra os resíduos em relação aos valores preditos. Sua construção tem como base a obtenção dos valores previstos e resíduos e na construção do gráfico cartesiano, com o eixo x sendo os valores previstos e o eixo y sendo os resíduos, e preenchendo de cada ponto de dados se dá colocando um ponto “seguindo o padrão (x , y)”.O gráfico é baseado no pressuposto da homogeneidade dos resíduos, que tem como base os princípios da independência dos resíduos, média zero dos resíduos e variância constante dos resíduos (homoscedasticidade), que afirma que os resíduos devem ser distribuídos aleatoriamente em torno de zero ao longo de todos os níveis dos valores preditos. Se os resíduos mostrarem um padrão isso sugere que o modelo de regressão não está capturando totalmente a relação entre as variáveis. Ele geralmente é usado para verificar a linearidade e homoscedasticidade, detectar outliers e observações influentes e avaliar a adequação do modelo. Veja o exemplo de um gráfico de resíduos vs valores previstos

Gráfico 7 – Resíduos vs Valores Previstos

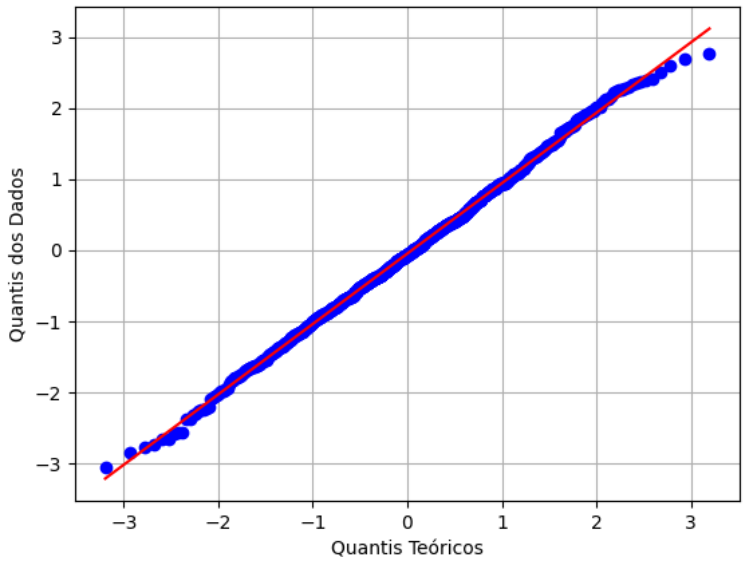


Fonte: Dados hipotéticos do autor (2024)

2.10 Gráfico Q-Q Plot

De acordo com Ford (2015), um gráfico Q-Q Plot é um gráfico de dispersão criado traçando dois conjuntos de quantis, quantis observados vs quantis esperados. Se ambos os conjuntos de quantos viessem da mesma distribuição, deveríamos ver os pontos formando uma linha aproximadamente reta, já que os seus dados são parecidos. Veja o exemplo e perceba que os pontos descrevem aproximadamente uma reta, logo eles provavelmente vêm de uma mesma distribuição

Gráfico 8 - Q-Q Plot



Fonte: Dados hipotéticos do autor (2024)

O gráfico Q-Q Plot é utilizado na regressão linear, principalmente para verificar a normalidade dos resíduos, lembrando que a normalidade dos resíduos é uma das suposições básicas da regressão linear.

2.11 Intervalo de confiança dos coeficientes para regressão linear

Intervalo de confiança é um conceito estatístico utilizado para estimar o valor verdadeiro de um parâmetro populacional com um certo nível de confiança, de modo que representa um intervalo de valores que provavelmente contém o verdadeiro parâmetro populacional

Segundo Bobbitt (2022) a formula que define o intervalo de confiança para os coeficientes de uma regressão linear simples é dada por: $\hat{\beta}_i \pm t_{n-2,\alpha/2} \cdot SE(\hat{\beta}_i)$ para o coeficiente $\hat{\beta}_1$ e para $\hat{\beta}_0$ temos: $\hat{\beta}_i \pm t_{n-2,\alpha/2} \cdot SE(\hat{\beta}_i)$, onde $SE(\hat{\beta}_1)$ é o erro padrão de $\hat{\beta}_1$ que dado por $SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}$ e quando estivermos calculando $\hat{\beta}_0$, $SE(\hat{\beta}_0)$ é o erro padrão de $\hat{\beta}_0$ que dado por $SE(\hat{\beta}_0) = \sqrt{\Theta^2(\frac{1}{n} + \frac{\hat{x}^2}{\sum(x_i - \bar{x})^2})}$, onde Θ^2 é a covariância do resíduos que é dado por $\frac{1}{n-2} \cdot \sum_{i=1}^{n} \varepsilon^2$, $x_i$valor da vaiável explicativa na posição i e o $\overline{x}$ a média dos valores de x já o $\pm t_{n-2,\alpha/2}$ é valor crítico da distribuição t de Student onde $n-2$ são os graus de liberdade (número de observações menos 2) e $\alpha/2$ é o nível de significância (por exemplo, 0,025 para um intervalo de confiança de 95%). Perceba que devemos multiplicar o erro padrão de $\hat{\beta}_1$com valor crítico da distribuição t de Student para limitarmos a o intervalo que desejamos e o "$\pm$" para estabelecer o limite superior e inferior.

Exemplo:

x = {1, 2, 3, 4 e 5}

y = {2.5, 3.5, 4, 5 e 5.5}

$\hat{y}_i$ = 1.85 + 0,75x_i

Número de observações = 5

Desvio padrão dos resíduos = $\frac{0.075}{5-2} = 0.025$

$$SE(\hat{\beta}_1) = \sqrt{\frac{\Theta^2}{\sum\sum(x_i - \bar{x})^2}} \approx 0.166$$

$SE(\hat{\beta}_0) = \sqrt{\Theta^2}$

$\alpha = 0.05 \rightarrow \frac{0.05}{2} = 0{,}025$

Graus de liberdade = 5 – 2 = 3

Valor crítico da distribuição t de Student = 3.182

Intervalo de confiança de $\hat{\beta}_1 = 0.75 \pm 3.182 \cdot 0.166$ o que implica em Intervalo de confiança de $\hat{\beta}_1$ sendo [0.591, 0.909]

Intervalo de confiança de $\hat{\beta}_0 = 1.85 \pm 3.182 \cdot 0.05$ o que implica em Intervalo de confiança de $\hat{\beta}_0$ sendo [1.322, 2.378]

2.12 Independência do erro e intervalo de confiança

Conforme Fávero e Belfiore (2017), o intervalo de confiança desempenha um papel crucial para análise de regressão linear, fornecendo informações sobre a incerteza associada às estimativas dos parâmetros do modelo.

Ainda segundo Fávero e Belfiore (2017), com a não independência do erro o intervalo de confiança se torna suscetível a influência do tamanho incorreto da amostra. Perceba que os intervalos de confiança são calculados com base na estimativa pontual dos coeficientes de regressão e sua variabilidade. Em uma regressão linear simples podemos calcular o intervalo de confiança da seguinte forma: $Intervalode\ Confiança = \beta_i \pm t_{\frac{\alpha}{2},df} \cdot SE(\beta_i)$ onde:

- β_i é o coeficiente da variável independente;
- $t_{\frac{\alpha}{2},df}$ é o valor crítico da distribuição t de Student para um dado nível de confiança (geralmente 95%) e um determinado grau de liberdade (df), onde α é o nível de significância;
- $SE(\beta_i)$ é o erro padrão do coeficienteβ_i.

A estimativa do erro padrão do coeficiente é calculada com base na variabilidade dos resíduos e no tamanho da amostra. Se os erros não são independentes, a variabilidade real dos resíduos pode ser subestimada ou superestimada. Isso pode levar a uma estimativa incorreta do erro padrão e, consequentemente, a intervalos de confiança incorretos.

2.13 Independência do erro e teste de significância

Segundo Fávero e Belfiore (2017), os testes de significância desempenham um papel fundamental para análise de regressão linear, pois ajudam a determinar a validade estatística das relações entre as variáveis independentes e dependentes.

Com a não independência do erro na regressão linear simples e múltipla, a falta de independência dos erros pode afetar a validade dos testes de significância podendo levar a testes de significância inválidos para os parâmetros do modelo, podendo ser excessivamente conservadores ou excessivamente liberais, dependendo da natureza da dependência dos erros.

O teste t para análise de regressão linear é usado para avaliar a significância estatística de um coeficiente de regressão estimado. Em outras palavras, ele determina se o efeito estimado de uma variável independente na variável dependente é estatisticamente diferente de zero.

Ele é usado principalmente na análise de variância (ANOVA) e na análise de regressão para avaliar a significância global do modelo. Enquanto o teste t é usado para avaliar a significância dos coeficientes individuais de regressão, o teste F é usado para avaliar se o conjunto de variáveis independentes, como um todo, tem um efeito significativo sobre a variável dependente.

Com a não independência do erro na regressão linear simples e múltipla o teste t e F não seguem as distribuições esperadas: Os testes de hipóteses na regressão linear dependem da suposição de que os erros têm distribuição normal e são independentes. Se os erros não forem independentes, as distribuições t e F associadas aos testes t e F podem não ser mais válidas. Isso ocorre porque as estatísticas t e F são derivadas da razão entre estimativas de efeito e estimativas de variabilidade (erro padrão), $Testet = \frac{\hat{\beta}_i - \beta_{i,0}}{SE(\hat{\beta}_i)}$ (com: $\hat{\beta}_i$ sendo o coeficiente estimado para a variável independente; $\beta_{i,0}$ sendo o valor sob a hipótese nula; $SE(\hat{\beta}_i)$ sendo o valor do erro padrão estimado do coeficiente $\hat{\beta}_i$) e $TesteF = \frac{SSR}{SSE/(N-K-1)}$ (com: SSR sendo a soma dos quadrados da regressão; SSE sendo a soma dos quadrados do erro; K sendo o número de variáveis independentes no modelo; N sendo o número de observações),e se a dependência entre os erros não for considerada, a variabilidade real pode ser sub ou superestimada já que influenciará diretamente os denominadores

2.14 *Box Plot*

Segundo Opencadd (2022), Box Plot, também conhecido como diagrama de caixa, é uma ferramenta gráfica utilizada para ilustrar um conjunto de dados. Ele permite a visualização da distribuição dos dados com base em cinco estatísticas principais:

- Mínimo;
- Primeiro quartil (Q1);
- Mediana;
- Terceiro quartil (Q3);
- Máximo.

Esses valores são frequentemente referidos como o resumo dos cinco números.

Além disso, o Box Plot pode destacar valores discrepantes (outliers), fornecendo uma medida adicional para compreender as características dos dados apresentados.

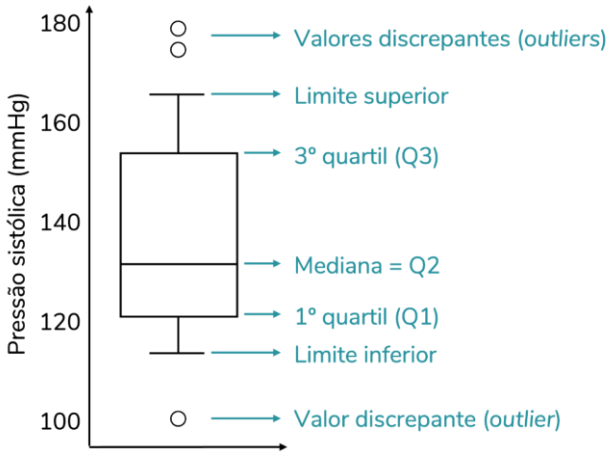
O Box Plot também ajuda a identificar a posição dos dados, sua simetria, dispersão, a extensão das caudas da distribuição e possíveis distorções.

No gráfico do Box Plot, a haste vertical é interpretada de baixo para cima: a parte inferior da haste representa o valor mínimo, enquanto a parte superior representa o valor máximo, ignorando quaisquer outliers.

O retângulo central na haste contém três linhas horizontais: a linha inferior do retângulo marca o primeiro quartil, a linha superior marca o terceiro quartil, e a linha interna representa a mediana, ou segundo quartil.

Valores discrepantes, outliers e extremos são geralmente representados por asteriscos ou pontos, destacando pontos atípicos no gráfico.

Gráfico 9 – Modelo de Box Plot



Fonte: Peres (2022)

Para construi-lo seguiremos os seguintes passos:

1. Organizar os dados em ordem crescente
2. Calcular mediada(Q2)
3. Calcular primeiro e terceiro quartil
4. Calcular o intervalo interquartil (AIQ)
5. Identificar outliers
6. Determinar o mínimo e máximo.
7. Desenhar gráfico

Exemplo:

Idade (anos): {24, 24, 25, 25, 27, 30, 32, 32, 35, 45}

Mediana = 28,5

Quartil 1 = 25

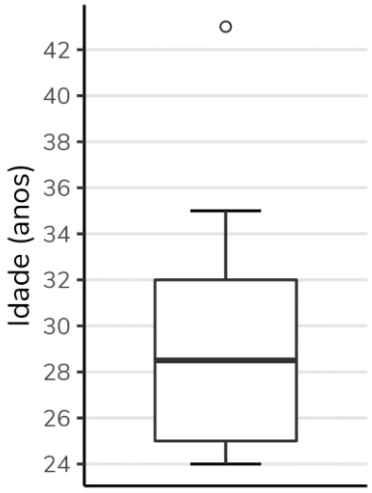
Quartil 3 = 32

Limite inferior teórico = Q1 − 1,5 × AIQ = 14,5

Limite superior teórico = Q3 + 1,5 × AIQ = 42,5

Observe que em nosso banco de dados não existem valores abaixo do nosso limite inferior teórico, portanto, o menor valor presente no banco é 24 será considerado nosso limite inferior verdadeiro e será utilizado no gráfico. Em contrapartida, há um valor que excede o limite superior teórico que é o 43, esse valor será tratado como um outlier. Assim, o limite superior verdadeiro, que será representado no gráfico, será o maior valor do banco excluindo o outlier, no caso esse valor é 35. O gráfico ficara assim:

Gráfico 10 – Box Plot Idade



Fonte: Peres (2022)

Cabe mencionar que a construção de um box plot é um procedimento descritivo que se baseia nas características da distribuição dos dados e não nos modelos de regressão, logo a sua construção não muda independente de que regressão seja.

2.15 O que é uma regressão linear múltipla

Segundo Triola (2017), a regressão linear múltipla é um modelo de análise que usamos quando modelamos a relação linear entre uma variável de dependente e múltiplas variáveis independente

2.16 Modelo de regressão linear múltipla

Como aponta Chein (2019), A Equação geral para uma regressão linear múltipla é: $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_i \cdot x_i + \varepsilon$

Onde:

- y é a variável dependente
- $x_1, x_2, \ldots, x_i$são as variáveis independentes
- $\beta_0, \beta_1, \ldots, \beta_i$são os coeficientes da regressão
- ε é o termo de erro que representando a variação não explicada pelo modelo

2.17 Cálculo dos coeficientes de uma regressão linear múltipla

Usando o método dos mínimos quadrados ordinários, que consiste na soma dos quadrados das diferenças entre os valores observados e os valores previstos pela linha seja mínima possível, chegamos nas seguintes fórmulas dos coeficientes β_0, $\beta_1, \beta_2, \beta_i$: $\beta = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$

com:

- X (da fórmula) está representando a matriz design de X(variável);
- Y o vetor que contém os valores da variável dependente;
- X^T é a matriz transposta de X;
- $(X^T \cdot X)^{-1}$ é a inversa da matriz $X^T \cdot X$;
- $X^T \cdot Y$ é o produto da transposta de *X* pelo vetor *Y*

Exemplo:

$x_1 = \{1,2,3,4,5\}$
$x_2 = \{3,4,2,5,1\}$

$y = \{10,20,25,30,35\}$

Matriz Design de $X = \begin{pmatrix} 1 & 1 & 3 \\ 1 & 2 & 4 \\ 1 & 3 & 2 \\ 1 & 4 & 5 \\ 1 & 5 & 1 \end{pmatrix}$

Vetor $Y = \begin{pmatrix} 10 \\ 20 \\ 25 \\ 30 \\ 35 \end{pmatrix}$

$X^T \cdot X = \begin{pmatrix} 5 & 15 & 15 \\ 15 & 55 & 15 \\ 15 & 15 & 39 \end{pmatrix}$

$X^T \cdot Y = \begin{pmatrix} 120 \\ 430 \\ 205 \end{pmatrix}$

$$(X^T \cdot X)^{-1} = \begin{pmatrix} 0{,}7 & -0{,}3 & 0{,}1 \\ -0{,}3 & 0{,}3 & -0{,}15 \\ 0{,}1 & 0{,}15 & 0{,}3 \end{pmatrix}$$

$$\beta = (X^T \cdot X)^{-1} \cdot X^T \cdot Y = \begin{pmatrix} 6 \\ 4 \\ 2 \end{pmatrix}$$

Logo, $\beta_0 = 6, \beta_1 = 4, \beta_2 = 2$

2.18 Cálculo de resíduos de uma regressão linear múltipla

Semelhante a regressão linear simples, os resíduos de uma regressão linear são as diferenças entre os valores observados da variável dependente e os valores previstos, para calcularmos (sendo regressão linear simples ou múltipla) usaremos seguinte fórmula é: $\varepsilon_i = y_i - \hat{y}_i$ sendo:

- ε_i sendo o resíduo da i-ésima observação;
- y_i o valor observado da variável dependente para a i-ésima observação;
- $\hat{y}_i$ valor previsto pela regressão para a i-ésima observação.

$x_1 = \{1,2,3,4,5\}$
$x_2 = \{3,4,2,5,1\}$

$y = \{10,20,25,30,35\}$

$\hat{y}_i = 6 + 4x_1 + 2x_2$

Tabela 6 –Valores para Cálculo do Resíduos na Regressão Linear Múltipla

x_1	x_2	y	$\hat{y}$	resíduo
1	3	10	16	-6
2	4	20	22	-2
3	2	25	22	3
4	5	30	32	-2
5	1	35	28	7

Fonte: Dados hipotéticos do autor (2024)

2.19 Pressupostos de uma regressão linear múltipla

De acordo com Chein (2019), os pressupostos da regressão múltipla servem como condições fundamentais que devem ser atendidas para garantir que os resultados obtidos a partir da análise sejam válidos e confiáveis, esses pressupostos são:

Linearidade: a relação entre as variáveis dependente e independentes deve ser linear;

Independência de erro: os erros de previsão devem ser independentes um do outro;

Independência das Variáveis Independentes: As variáveis independentes não devem estar altamente correlacionadas entre si, já pode dificultar a interpretação dos coeficientes e levar a estimativas imprecisas.

Homoscedasticidade: a variância dos erros deve ser constante em relação às variáveis independentes;

Normalidade dos Erros (E): os erros de previsão devem ser distribuídos normalmente;

O valor esperado do erro é zero: os resíduos devem ter um valor médio de zero.

Iremos agora abordar "o porquê" de elas serem tão importantes

2.19.1 Linearidade

Por se tratar de um modelo em que queremos estabelecer ou a função de uma reta ou de um plano ou de um hiperplano é essencial supor que exista linearidade afim de quer os dados estejam o mais próximo possível deles

2.19.2 Independência de erro

Conforme Chein (2019), de forma semelhante à regressão linear simples, na regressão linear múltipla os erros de previsão (ou resíduos) não estão correlacionados entre si, logo, não há padrão perceptível nos resíduos quando são plotados contra os valores ajustados. Suas violações também geram consequências também são semelhantes já que podem distorcer os resultados, pois os intervalos de confiança (que já foi explicada as consequências do não cumprimento na regressão linear múltipla em regressão linear simples) e os testes de

significância podem ser inválidos, levando a conclusões errôneas sobre a relação entre as variáveis (as consequências da violação desse princípio será explicado mais na frente).

2.19.3 Independência das variáveis independentes

De acordo com a perspectiva de Chein (2019), a independência das variáveis independentes é a ausência de qualquer relação entre as variáveis independentes em um modelo estatístico. Isso significa que cada variável independente traz informações únicas e não correlacionadas para o modelo.

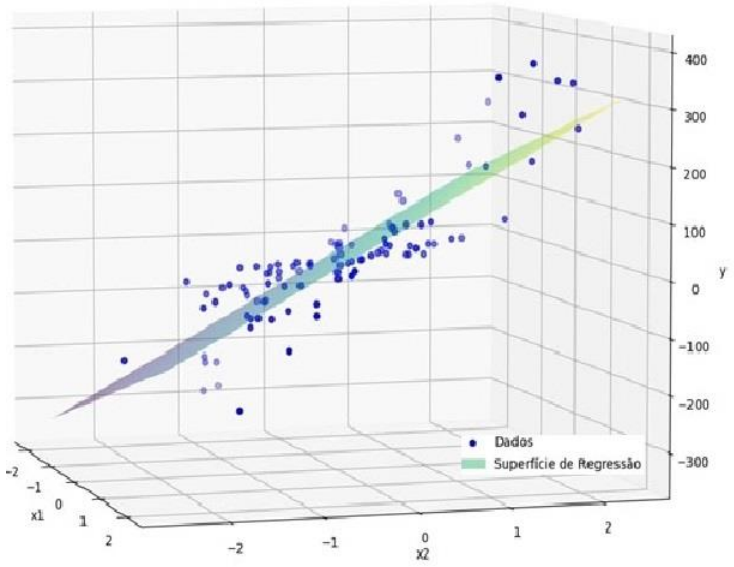
A Relação entre variáveis preditoras é extremamente problemática para a regressão múltipla já que ela atrapalha na interpretação dos coeficientes, pois torna-se difícil determinar o efeito único de cada variável sobre a variável dependente. e já que os coeficientes podem indicar que uma variável tem um grande efeito sobre a variável dependente pode haver uma confusão, pois o coeficiente pode indicar que um x_i é extremamente relevante quando, na realidade, esse efeito é devido à alta correlação com outras variáveis independentes. E por consequência disso teremos dificuldades na identificação do verdadeiro impacto das variáveis explicativas.

Resumindo, a dependência das variáveis independentes pode comprometer a interpretação, a estabilidade e o poder preditivo de um modelo de regressão o tornando menos confiável para fazer inferências sobre as relações entre as variáveis envolvidas.

2.19.4 Homoscedasticidade

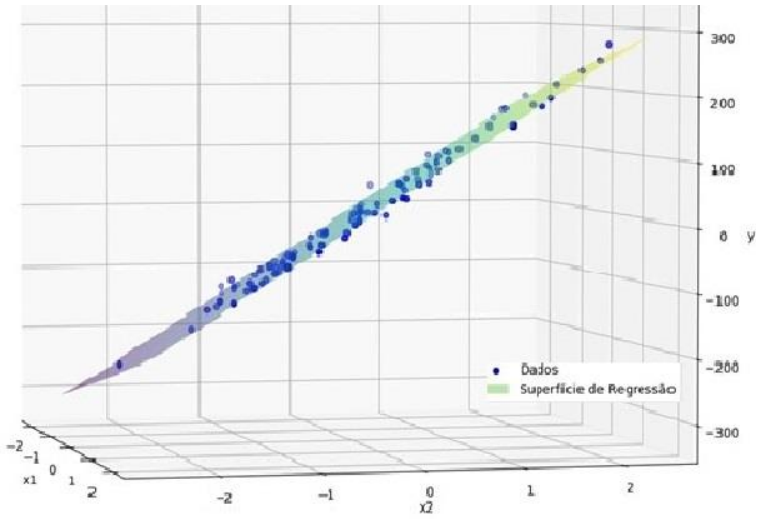
De acordo com Chein (2019), de maneira análoga à regressão linear simples, na regressão linear múltipla a Homoscedasticidade é de extrema importância, pois refere-se à consistência da variância dos resíduos ao longo de todas as observações em um modelo de regressão. Note que se tivermos heterocedasticidade nossos resíduos não terão consistência e como resíduo é a diferença do valor dado pelo valor previsto teremos quanto mais variáveis eles forem piores, pois implica baixa previsão. Veja os gráficos e perceba como ao calcular os resíduos em um gráfico heterocedástico elas terão uma grande viração nos resíduos.

Gráfico 11 - Regressão Linear Múltipla com Heterocedasticidade



Fonte: Dados hipotéticos do autor (2024)

Gráfico 12 - Regressão Linear Múltipla com Homoscedasticidade



Fonte: Dados hipotéticos do autor (2024)

2.19.5 Normalidade dos erros

Segundo Chein (2019) de forma parecida à regressão linear simples, na regressão linear múltipla, a normalidade dos erros na regressão linear é uma suposição fundamental para que os testes estatísticos e os intervalos de confiança associados à análise sejam válidos. Como ela afetara-as será explicada adiantamento

2.19.6 O valor esperado do erro é zero (E)

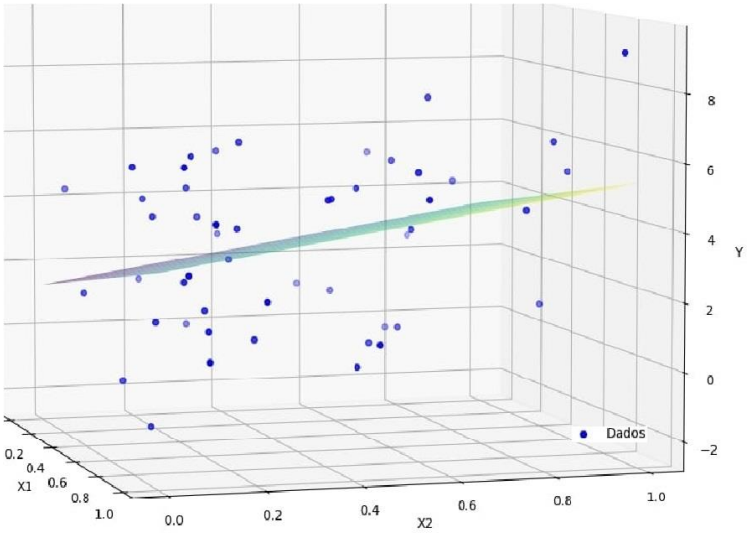
De acordo com Chein (2019), de maneira semelhante à regressão linear simples, na regressão linear múltipla o valor esperado refere-se à média condicional do erro, dado um conjunto de valores específicos das variáveis independentes

Se o valor esperado (média) dos erros de previsão é zero, isso sugere que o modelo está capturando corretamente a relação entre as variáveis, isso é importante porque indica que o modelo não está sistematicamente superestimando ou subestimando os valores.

Para se calcular o valor esperado do erro na regressão linear é necessário entender que o valor esperado do erro é uma medida da precisão das previsões feitas pelo modelo de regressão e se calcula da seguinte forma: $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$

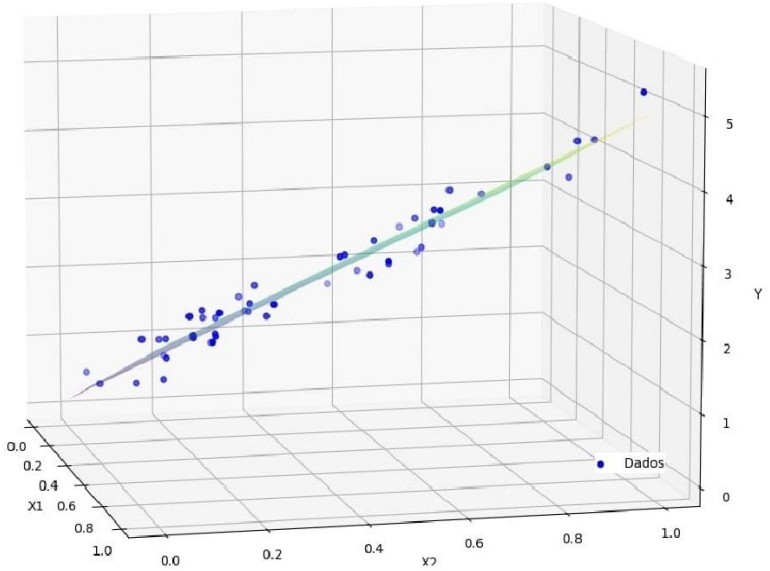
Se o valor esperado dos erros não for próximo de zero, isso indica que o modelo está sistematicamente errado em suas previsões. Por exemplo, se os resíduos têm uma média positiva, significa que o modelo está subestimando consistentemente os valores reais da variável dependente, semelhantemente, se os resíduos possuem uma média negativa, o modelo está superestimando consistentemente os valores reais da variável dependente, isso é problemático porque significa que nossas previsões estão enviesadas. Se o modelo está consistentemente errado em uma direção, isso sugere que ele não está capturando adequadamente a relação entre as variáveis independentes e dependentes. Isso pode ser devido a uma série de razões, como a inclusão de variáveis irrelevantes, a exclusão de variáveis importantes ou a violação de pressupostos importantes da regressão linear. Perceba nos seguintes gráficos (sem ruido) o impacto quando (E) = 0 e (E) = 5. respectivamente.

Gráfico 13 - Regressão Linear Multipla com Erro Esperando Sendo 5



Fonte: Dados hipotéticos do autor (2024)

Gráfico 14 - Regressão Linear Multipla com Erro Esperando Sendo 0



Fonte: Dados hipotéticos do autor (2024)

2.20 Coeficiente de correlação (Pearson (r))

Segundo MONTGOMERY, PECK, VINING (2012), o cálculo e a ideia de correlação são semelhantes, só que diferentemente de uma regressão linear simples onde temos apenas uma variável explicativa(independente) na regressão linear múltipla temos várias variáveis explicativas, portanto y já considera todas as variáveis independentes simultaneamente, logo não é muito adequado usar o coeficiente de Pearson (já que ele mede a

correlação linear simples entre duas variáveis) para verificar a linearidade do modelo mais ainda podemos calcular a correção da variável dependente com cada variável independente, só que como foi mencionado anteriormente o y leva consideração "tudo". Para determinar a linearidade devemos usar o $R^2{}_{ajustado}$ que será explicado adiante. Contudo, partido da ideia de que não possuímos a equação do modelo, logo sem a possibilidade de calcular o $R^2{}_{ajustado}$, ainda se pode usar o coeficiente de Pearson entre a variável dependente e cada variável explicativa para verificar a linearidade do modelo, mas vale lembrar que o $R^2{}_{ajustado}$ é mais confiável

2.21 *Coeficiente de determinação* ($R^2{}_{ajustado}$)

Segundo Chein (2019), assim como R^2 ele mede o quanto o modelo explica a variação dos dados. Só que com a presença de inúmeras variáveis, por se tratar de uma regressão linear múltipla, mesmo que tenham muito pouco poder explicativo sobre a variável dependente, aumentarão o valor de R^2, logo para combater esta tendência deve haver penalização para inclusão de variáveis independentes adicionais que não contribuem significativamente assim teremos: $R^2{}_{ajustado} = 1 - \frac{(n-1)}{n-(k+1)} \cdot (1 - R^2)$ com:

- n sendo o número de observações;
- R^2 sendo o coeficiente de determinação;
- (k + 1) representa o número de variáveis explicativas mais a constante

Exemplo:

$x_1 = 1,2,3,4,5$

$x_2 = 3,4,2,5,1$

$y = \{10,20,25,30,35\}$

$\hat{y}_i = 6 + 4x_1 + 2x_2$

$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Soma dos quadrados dos resíduos: 102

Soma dos quadrados totais: 370

$R^2 = 1$ - (102 - 370) $= 1$ - 0.2757 $=$ 07243

$$R^2{}_{ajustado} = 1 - \frac{(n-1)}{n-(k+1)} \cdot (1 - R^2)$$

Número de observações $= 5$

Número de variáveis explicativas $= 2$

$R^2{}_{ajustado}$ = 1 - ((1 - 0.7243) - (5 -2 -1)) x (5 - 1) = 0.4486

2.22 Gráfico de Resíduos vs Valores Previstos

Independente de que tipo de regressão seja, resíduos é resíduo e valor preditivo é valor preditivo, logo os princípios de criação são semelhantes, é só criar um gráfico com o eixo x sendo os valores preditos e y sendo resíduos e preencher com os dados normalmente

2.23 Gráfico Q-Q Plot

Independente de que tipo de regressão seja podemos fazer o Q-Q, já que se tratar de um gráfico quantil por quantil com objetivo de verificar se duas distribuições de dados são semelhantes entre si, logo não há distinção em seu uso e em sua "montagem" de uma regressão linear simples e múltipla.

2.24 Intervalo de confiança dos coeficientes para regressão linear múltipla

De acordo com Forjan (2019) de maneira análoga a regressão linear simples na regressão linear múltipla o intervalo de confiança é um conceito estatístico utilizado para estimar o valor verdadeiro de um parâmetro populacional com um certo nível de confiança, de modo que representa um intervalo de valores que provavelmente contém o verdadeiro parâmetro populacional

A formula que define o intervalo de confiança para os coeficientes de uma regressão linear múltipla é dada por: $\beta_i \pm t_{\frac{\alpha}{2},n-k-1} \cdot SE(\beta_i)$ para o coeficiente β_i onde $SE(\beta_i)$é o erro padrão de β_i que dado por $SE(\beta_i) = \sqrt{VAR(\beta_i)}$ onde $VAR(\beta_i)$ é a matriz de variância-covariância dos estimadores, $VAR(\beta_i) = \sigma^2(x^t x)^{-1}$ com σ^2 sendo $\frac{1}{n-k-1} \cdot \sum_{i=1}^{n} \varepsilon^2$, $t_{\frac{\alpha}{2},n-k-1}$ é o valor crítico da distribuição t de Student onde n–k -1 são os graus de liberdade sendo n o número de observações, k o número de variáveis explicativas e $\alpha/2$ é o nível de significância (por exemplo, 0,025 para um intervalo de confiança de 95%). Perceba

que devemos multiplicar o erro padrão de $\beta_i$com valor crítico da distribuição t de Student para limitarmos a o intervalo que desejamos e o "$\pm$" para estabelecer o limite superior e invoce agferior.

Exemplo:

$x_1 = \{1,2,3,4,5\}$

$x_2 = \{3,4,2,5,1\}$

$y = \{10,20,25,30,35\}$

$\hat{y}_i = 6 + 4x_1 + 2x_2$

resíduos = {−6, −2, 3, −2 e 7}

Soma dos Quadrados dos Resíduos = 102

teta quadrado: 102 / (5 - 2 - 1) = 102 / 2 = 51

$(X^T \cdot X)_{00}{}^{-1} = 2$

$(X^T \cdot X)_{11}{}^{-1} = 0.26$

$(X^T \cdot X)_{22}{}^{-1} = 0.26$

$SE_{\beta_0} = \sqrt{(X^T \cdot X)_{00}{}^{-1}\Theta^2} = 10.1$

$SE_{\beta_1} = \sqrt{(X^T \cdot X)_{11}{}^{-1}\Theta^2} = 3.64$

$SE_{\beta_2} = \sqrt{(X^T \cdot X)_{22}{}^{-1}\Theta^2} = 3.64$

Número de observações = 5

Número de variáveis explicativas = 2

Alpha = 0.05

Graus de liberdade = 5 - 2 - 1 = 2

Valor crítico t de Student = 0.05 / 2 = 0.025

Com 2 graus de liberdade e valor crítico de 0,025 na tabela de distribuição t de Student teremos 4.303

Logo, intervalo de confiança β_0 = [-37.46, 49.46] β_1 = [-11.66, 19.66] β_2 = [-13.66, 17.66]

Esses intervalos de confiança indicam a faixa dentro da qual os coeficientes verdadeiros da população se encontram com 95% de certeza

2.25 Independência do erro e intervalo de confiança

Segundo Forjan (2019), de forma semelhante a regressão linear simples o intervalo de confiança desempenha um papel crucial para análise de regressão linear, fornecendo informações sobre a incerteza associada às estimativas dos parâmetros do modelo.

Com a não independência do erro o intervalo de confiança se torna suscetível a influência do tamanho incorreto da amostra. Perceba que os intervalos de confiança são calculados com base na estimativa pontual dos coeficientes de regressão e sua variabilidade. Na regressão linear múltipla podemos calcular desta maneira:

$IntervalodeConfiança = \beta_i \pm t_{\frac{\alpha}{2},df} \cdot SE(\beta_i)$ onde:

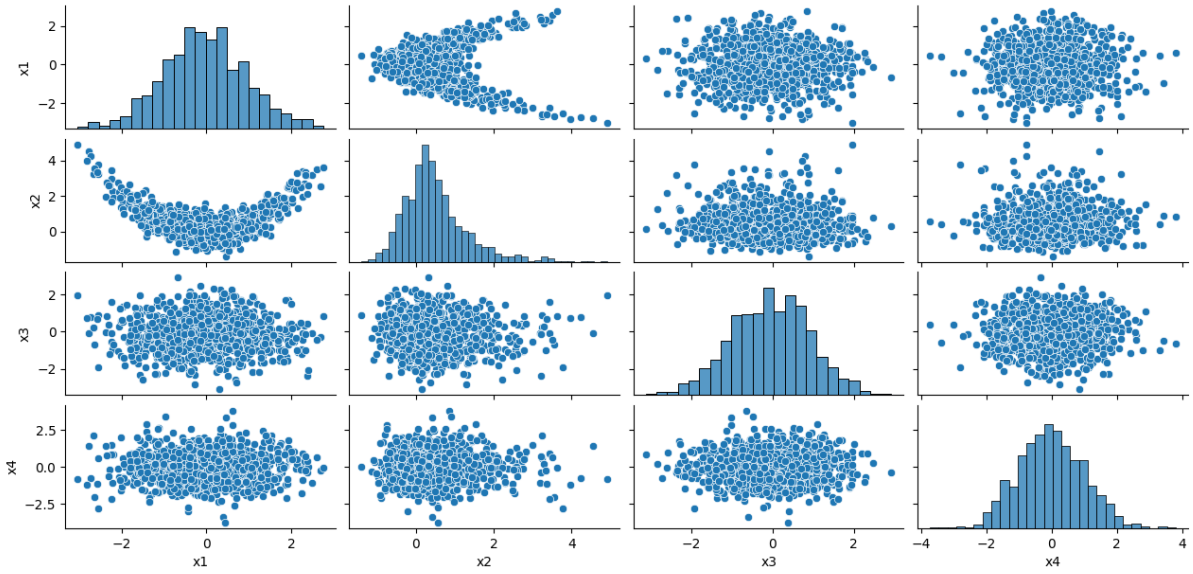
- β_i é o coeficiente da variável independente;
- $t_{\frac{\alpha}{2},df}$ é o valor crítico da distribuição t de Student para um dado nível de confiança (geralmente 95%) e um determinado grau de liberdade (df), onde α é o nível de significância;
- $SE(\beta_i)$ é o erro padrão do coeficiente β_i.

A estimativa do erro padrão do coeficiente é calculada com base na variabilidade dos resíduos e no tamanho da amostra. Se os erros não são independentes, a variabilidade real dos resíduos pode ser subestimada ou superestimada. Isso pode levar a uma estimativa incorreta do erro padrão e, consequentemente, a intervalos de confiança incorretos.

2.26 Matriz de dispersão

Segundo MONTGOMERY, RUNGER (2019), uma matriz de dispersão é uma representação gráfica de dados que visa mostrar a distribuição conjunta de dois ou mais conjuntos de valores, ela consiste em um grupo de gráficos de dispersão, onde cada célula da grade representa a relação entre duas variáveis, é usada para visualizar padrões, tendências e correlações entre múltiplas variáveis de forma simultânea. Na diagonal de uma matriz de dispersão, normalmente se encontra histogramas ou gráficos de densidade para cada variável individual de modo que cada célula diagonal representa a distribuição uni-variada de uma variável específica, por exemplo, se em uma matriz de dispersão das variáveis X1, X2, X3 e X4 a diagonal principal terá quatro subgráficos. Cada subgráfico mostrará a distribuição dos valores de uma variável individualmente, como um histograma ou um gráfico de densidade. Veja o exemplo as seguir dessa matriz de dispersão:

Gráfico 15 – Matriz de Dispersão com Variáveis x1, x2, x3 e x4



Fonte: Dados hipotéticos do autor (2024)

Veja outro exemplo de uma matriz de dispersão com os seguintes valores

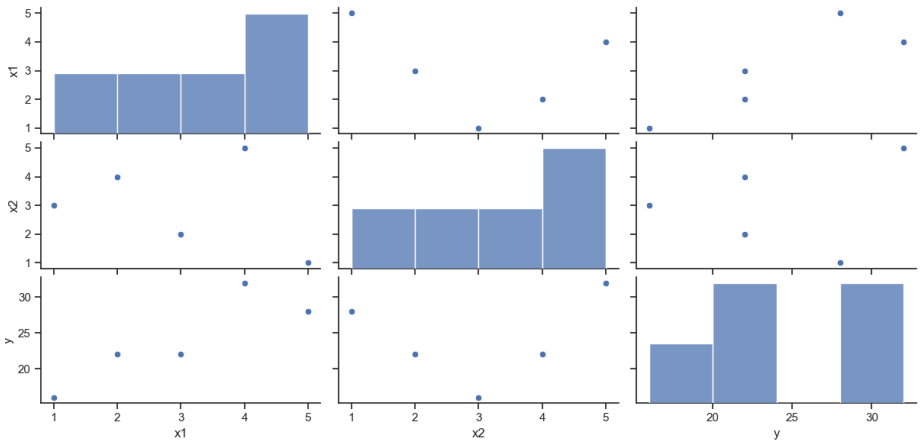
$x_1 = \{1,2,3,4,5\}$

$x_2 = \{3,4,2,5,1\}$

$\hat{y}_i = 6 + 4x_1 + 2x_2$

Teremos:

Gráfico 16 – Matriz de Dispersão com Variáveis Dadas



Fonte: Dados hipotéticos do autor (2024)

2.27 Matriz de correlação

Uma matriz de correlação é uma tabela que apresenta os coeficientes de correlação entre diferentes variáveis. Ela mostra a correlação entre todos os pares possíveis de valores em um conjunto de dados. Essa matriz é uma ferramenta poderosa para resumir informações de um grande conjunto de dados e para identificar padrões visuais nos dados fornecidos.

Cada célula da matriz representa o coeficiente de correlação entre duas variáveis específicas. As linhas e colunas da matriz correspondem às variáveis estudadas, e os valores dentro das células indicam a força e a direção da relação linear entre as variáveis correspondentes.

Além disso, a matriz de correlação é comumente utilizada em conjunto com outras análises estatísticas. Por exemplo, ela desempenha um papel crucial na análise de modelos de regressão linear múltipla ao revelar os coeficientes de correlação entre as variáveis independentes do modelo. Isso ajuda os analistas a entenderem como as variáveis estão inter-relacionadas e quais podem ter impacto significativo nos resultados do modelo. Geralmente se usa uma matriz de calor para representar a matriz de correlação já que se usa cores diferentes para indicar diferentes níveis de correlação, geralmente, se usa um esquema de cores com o vermelho (para correlações positivas mais fortes) e azul (para correlações negativas mais fortes) é utilizado para destacar visualmente essas relações.

Veja o exemplo:

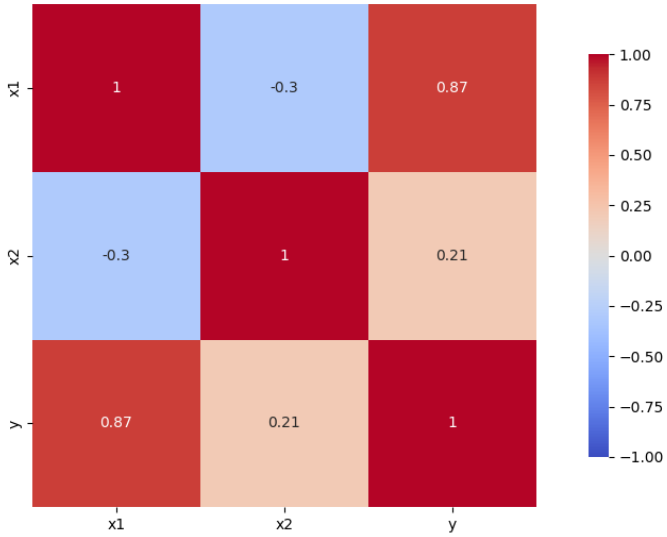
$x_1 = \{1,2,3,4,5\}$

$x_2 = \{3,4,2,5,1\}$

$\hat{y}_i = 6 + 4x_1 + 2x_2$

Gera a seguinte matriz de correlação:

Gráfico 17 - Matriz de Correlação



Fonte: Dados hipotéticos do autor (2024)

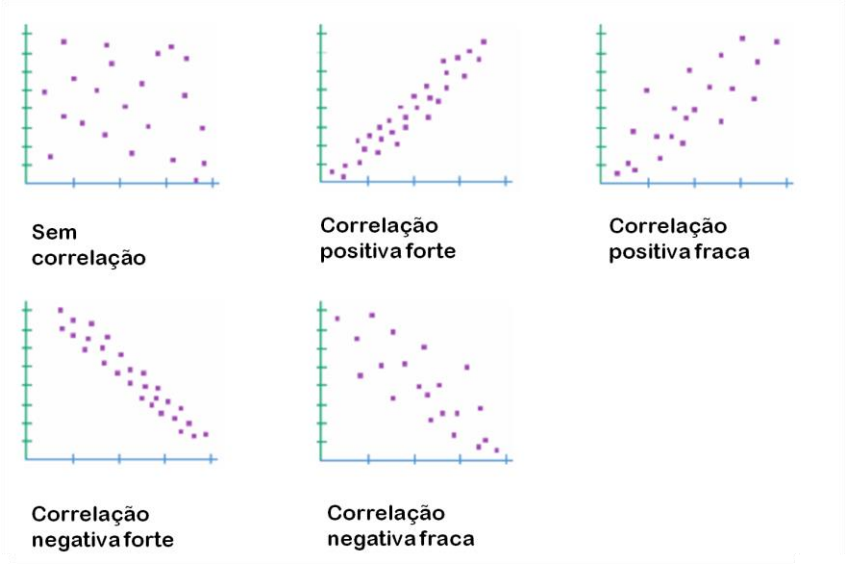
3 TÉCNICAS UTILIZADAS PARA OBTEÇÃO DE DADOS UTEIS PARA ANÁLISE DE DADOS.

Neste tópico iremos abordar como a regressão linear, seja ela simples ou múltiplas, nos fornece informações uteis para tomada de decisões.

3. 1 Identificação de relações lineares:

Conforme Chein (2019), sabemos que o coeficiente de correlação nos permite inferir se existe uma relação linear entre duas variáveis, ela pode se estabelecer das seguintes formas:

Gráfico 18 – Tipos de Correlação Linear



Fonte: Ross (2022)

Também sabemos que o coeficiente de determinação (R^2) é um percentual de ajuste da reta de regressão em relação aos dados amostrais y_i ou seja quanto maior o seu valor mais os dados estão distribuídos próximo da reta de regressão implicando em uma relação linear entre variáveis explicativas e dependentes. com seu valor variando de 0 a 1.

Com relação a ambos coeficientes podemos estabelecer a relação: $r^2(quadradodocoeficinetedepearson) = R^2(coeficinetededeterminação)$, logo através do R^2 podemos inferir se a relação entre as variáveis é linear ou não. Vale mencionar que essa relação é válida apenas para a regressão linear simples, ela não se aplica à regressão linear múltipla devido à complexidade adicional introduzida pela quantidade de variáveis independentes e suas interações.

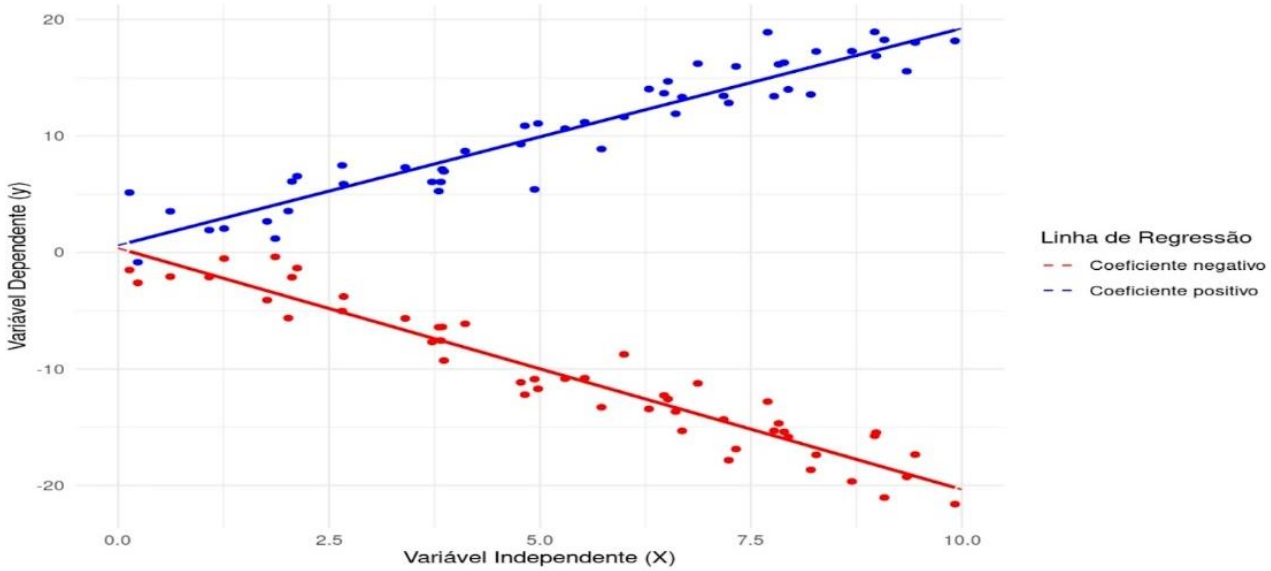
3.2 Significância do coeficiente estimados

Os coeficientes estimados em um modelo de regressão podem influenciar a distribuição da variável dependente de várias maneiras. Através dos coeficientes estimados pela Regressão Linear, é possível entender a direção e magnitude da influência das variáveis independentes na variável dependente, fornecendo insights para as tomadas de decisões

3.2.1 Coeficiente estimados e direção dos dados

A direção(inclinação) do modelo de regressão linear é definida pelos coeficientes estimados, tendo em vista que na regressão linear simples β_1 Indica quanto $\hat{y}_i$ varia a cada unidade de x_i , ele indica se vai para cima ou para baixo dependendo do seu sinal, positivo ou negativo; O mesmo ocorre na regressão linear múltipla, contudo como há uma maior quantidade de variáveis e coeficientes é necessária uma maior análise para chegar as conclusões desejadas. Desta maneira, através deles podemos inferir se os valores das variáveis independentes futuras irão ou crescer ou descrever. Veja um exemplo em uma regressão linear simples.

Gráfico 19 - Regressão linear com os Coeficientes Estimados sendo Positivos e Negativos

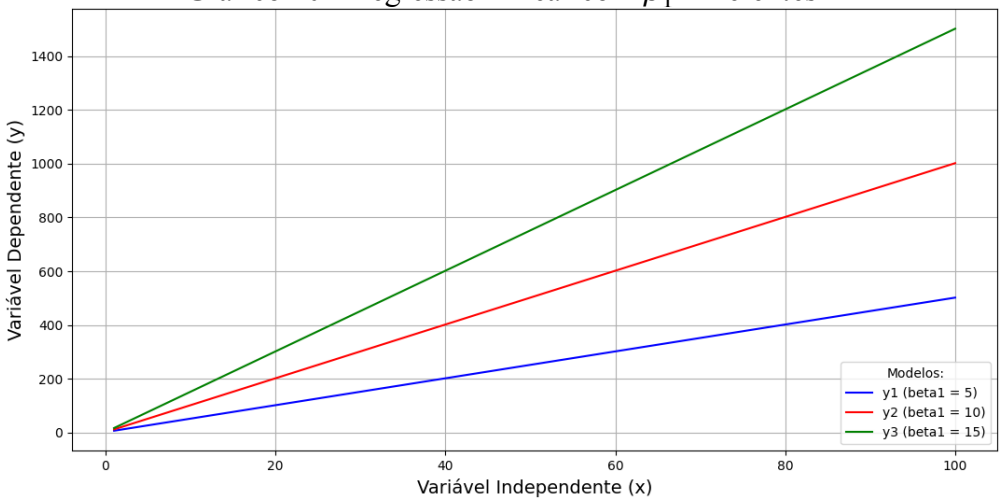


Fonte: Dados hipotéticos do autor (2024)

3.2.2 Coeficiente estimados e a magnitude do impacto

Perceba como a magnitude do impacto dos coeficientes estimados também é importante tendo em vista que um coeficiente maior indica um efeito mais forte da variável independente sobre a variável dependente. Por exemplo se β_1 é grande, então uma pequena mudança em x_i pode resultar em uma grande mudança na variável dependente. Desta maneira podemos ter a noção se pequenas mudanças podem ou não acarretar grandes mudanças e dependendo da aplicação isso é de extrema importância, como por exemplo na área farmacêutica. Na regressão linear múltipla ainda podemos inferir mais, dependendo do "tamanho" dos coeficientes estimados podemos deduzir quais variáveis explicativas têm maior importância (interfere mais nas variáveis dependentes). Veja os dois exemplos seguir que mostram como o β_1 influência na magnitude dos valores previstos na regressão simples e múltipla:

Gráfico 20 - Regressão Linear com β_1 Diferentes

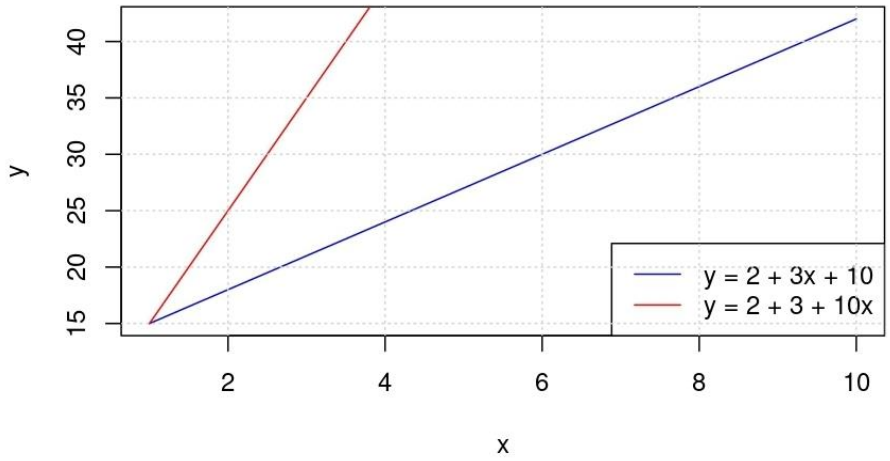


Fonte: Dados hipotéticos do autor (2024)

Note que sabendo β_1 podemos prever se uma pequena ou grande variação de X será impactante. Em resumo, se tratando de regressão linear simples, podemos fazer uma análise com uma função de primeiro grau, $y = a \cdot x + b$, com $a$sendo o coeficiente angular que "manda" se a reta irá subir ou descer e sua amplitude.

Além disso, na regressão múltipla o tamanho do coeficiente aponta qual variável tem maior impacto no modelo. Veja o exemplo da regressão múltipla $y = 2 + 3x_1 + 10x_2$:

Gráfico 21 - Influência dos Coeficientes na Regressão Múltipla



Fonte: Dados hipotéticos do autor (2024)

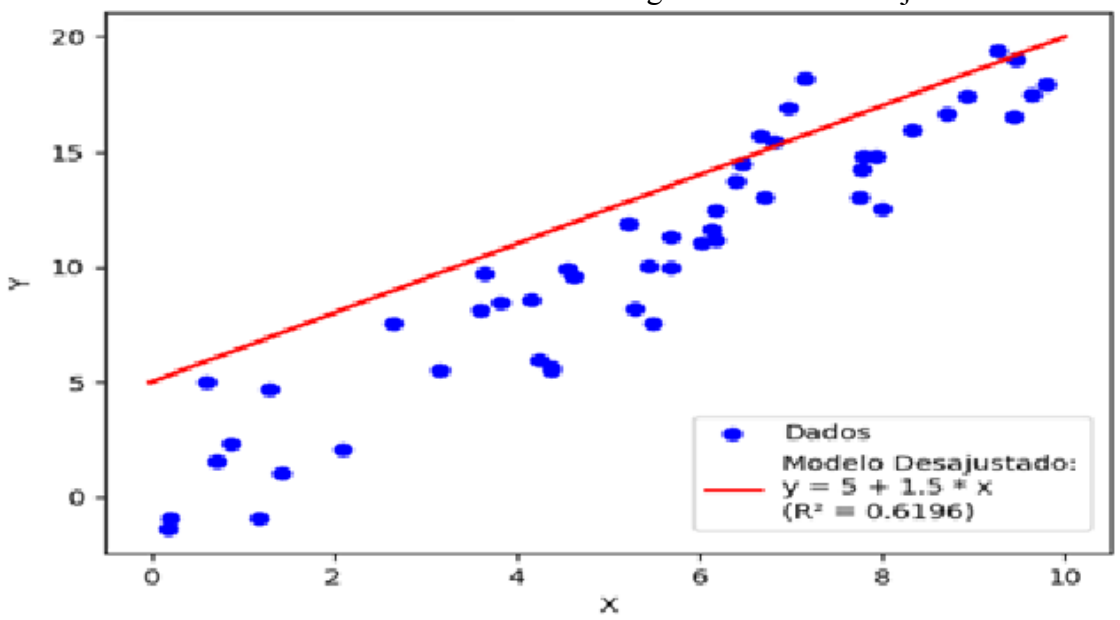
Perceba que na equação da reta azul x_2 é constante e igual a 1 e reta vermelha x_1é constante e igual a 1 a inclinação da reta vermelha é bem maior logo no modelo de regressão ela terá mais impacto nos resultados previstos

3.2.3 Coeficiente estimados e função corretiva

Os coeficientes estimados também possuem uma forte relação corretiva em modelos de regressão simples e múltipla. Isso acontece pois são usados para ajustar o modelo aos dados observados, já que o modelo tenta descrever a distribuição da variável dependente(aleatória) em relação às variáveis independentes(explicativas) com base nos coeficientes estimados. Um bom ajuste do modelo (boa escolha de coeficientes estimados) resulta em um modelo que se alinha bem com a distribuição observada da variável dependente. Uma boa escolha deles é feita através dos Mínimos Quadrados Ordinários (tanto para linear simples e múltipla) e logo após utilizar R^2 para escolher o melhor modelo disponivel. Os mínimos quadrados ordinários serve para encontrar os valores dos coeficientes que minimizam a soma dos quadrados dos resíduos (erros) e para encontrar o melhor intercepto, logo podemos estabelecer os melhores coeficientes para criarmos o melhor modelo possível descartando eventuais coeficientes escolhidos anteriormente, a fim de que possa explicar o melhor possível os dados, contudo nem sempre este método ir lhe dar os melhores resultados pois o mesmos é bastante afetado por: violação das suposições(linearidade, independência dos erros, homoscedasticidade e normalidade dos erros); presença de multicolinearidade (no caso da regressão múltipla); outliers ou se porventura o modelo de

regressão linear não capturar a complexidade dos dados. Veja o exemplo de como os coeficientes podem nos ajudar a corrigir um modelo regressão simples desajustado:

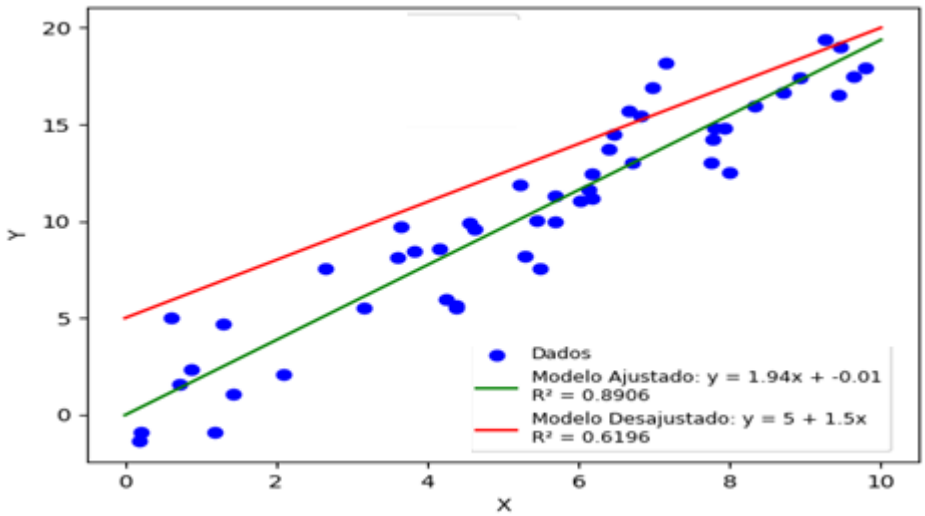
Gráfico 22 – Modelo de Regressão Linear Desajustado



Fonte: Dados hipotéticos do autor (2024)

Só visivelmente podemos visualizar que o modelo dado por $y = 5 + 1.5 \cdot x$ está desajustado, logo previsões feitas por ela serão muito imprecisas, agora veja que com ajustes nos coeficientes β_0 e β_1 podemos criar um modelo mais ajustado as informações que possuímos e por consequência podemos fazer previsões mais precisas.

Gráfico 23 – Modelos de Regressão Linear Ajustado e Desajustado



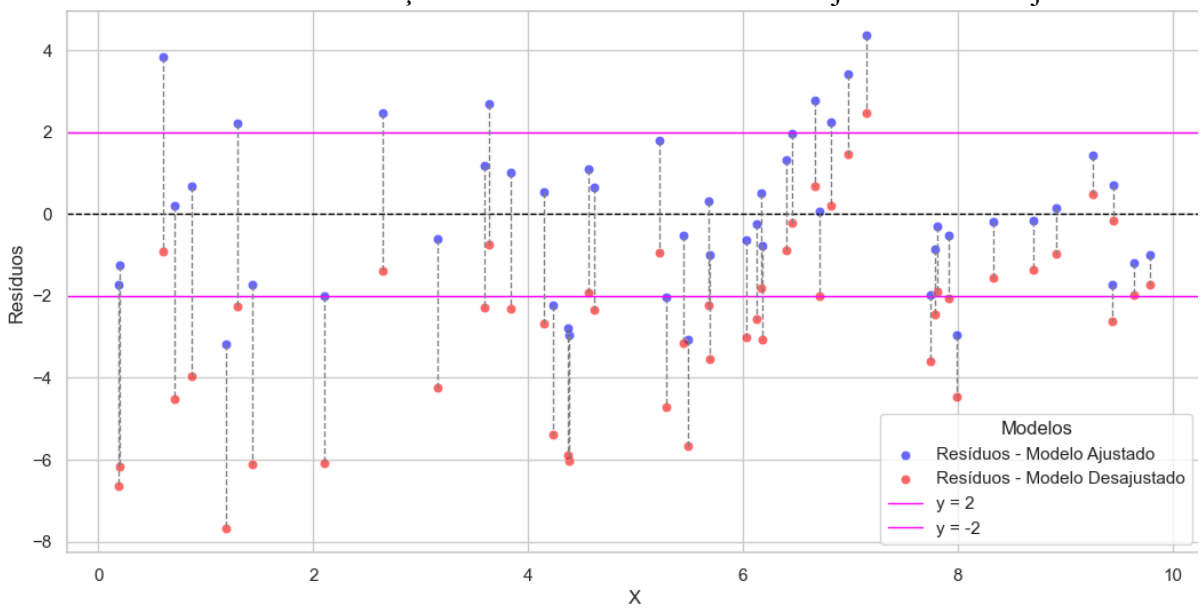
Fonte: Dados hipotéticos do autor (2024)

Perceba que os coeficientes foram alterados, e essas alterações fizeram o modelo se encachar “melhor” a nossos dados.

Podemos visualizar pelo R^2 que a reta de regressão verde se ajusta bem melhor ao modelo. Note que sempre após o ajuste deve ser feito uma validação a fim de confirmar se realmente houve uma melhora no modelo, isso geralmente envolve verificar a adequação do modelo por meio de métricas como o coeficiente de determinação R^2. Dessa maneira, podemos ter um modelo bem mais preciso para realizações de inferência estatística. Vale ressaltar que tal interpretação não diferencia para regressão simples ou múltipla

Sabemos, então que bons coeficientes estimados fazem com que a reta de regressão se adapte melhor aos elementos de estudo e por consequência esses coeficientes ajustados são essenciais para fazer previsões da variável dependente para diferentes valores das variáveis independentes. Se os coeficientes são precisos e representam adequadamente a relação subjacente entre as variáveis, as previsões do modelo serão mais confiáveis e úteis. Veja o exemplo do gráfico a seguir que mostra a diferença de resíduos ajustados e não-ajustados:

Gráfico 24 - Diferença entre Resíduos dos Modelos Ajustados e Desajustados



Fonte: Dados hipotéticos do autor (2024)

Perceba que quanto mais próximo e em maior quantidade os resíduos estiverem do 0 melhor e mais fidedigna descreve a situação. Note que a quantidade de resíduos entre y = 2 e y = - 2 são 35 ajustados e 19 não-ajustados, logo podemos afirmar que quando melhor for

ajuste nos coeficientes dos modelos, o modelo descreverá mais precisamente a situação estudada

Em resumo, os coeficientes estimados em um modelo de regressão não apenas quantificam a relação entre variáveis, mas também desempenham um papel fundamental na determinação da distribuição da variável dependente, conforme influenciada pelas variáveis independentes incluídas no modelo. Eles ajudam a caracterizar e explicar como a variável dependente varia em relação às variáveis independentes, facilitando a interpretação e a análise dos dados.

3.3 Análise de resíduos

Primeiro deixemos claro a diferença de erro e resíduo. Os erros são os valores reais não observados que representam a variação não explicada nos dados já os resíduos são as estimativas dos erros calculadas como diferenças entre os valores observados e os valores previstos pelo modelo. Contudo em modelo bem estruturado seus valores são extremamente parecidos, tanto que as propriedades dos erros também se aplicam nos resíduos, logo se a propriedade do resíduo não condizer com a dos erros significa que o modelo não está bem construído.

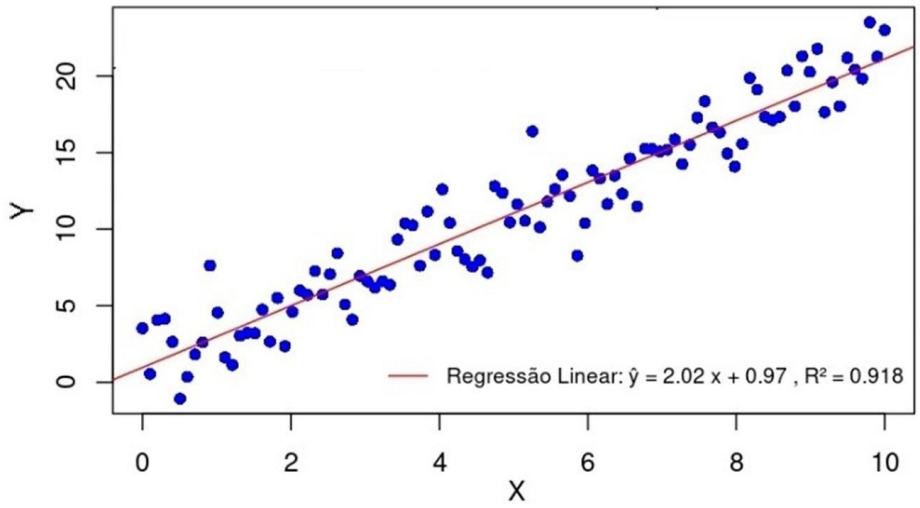
Uma análise detalhada de resíduos é uma etapa essencial no processo de modelagem estatística, especialmente em análises de regressão.

3.3.1 Gráfico Resíduos vs Valores Preditos (Teste de homoscedasticidade)

O gráfico de resíduos vs valores preditos é uma ferramenta importante para diagnosticar a qualidade de um modelo de regressão linear. Ele ajuda a identificar padrões nos resíduos e avaliar se a suposição de homoscedasticidade foi cumprida. A homoscedasticidade é uma importante propriedade dos resíduos em um modelo de regressão e é fundamental para que as estimativas dos parâmetros sejam eficientes e não enviesadas. O gráfico deve ser aleatório, sem mostrar nenhum padrão específico (como tendência crescente ou decrescente) e com o eixo “x” sendo os valores preditos e o “y” os resíduos.

Veja o seguinte exemplo de como os resíduos podem nos ajudar a analisar os dados para a identificação de não heterocedasticidade de um modelo:

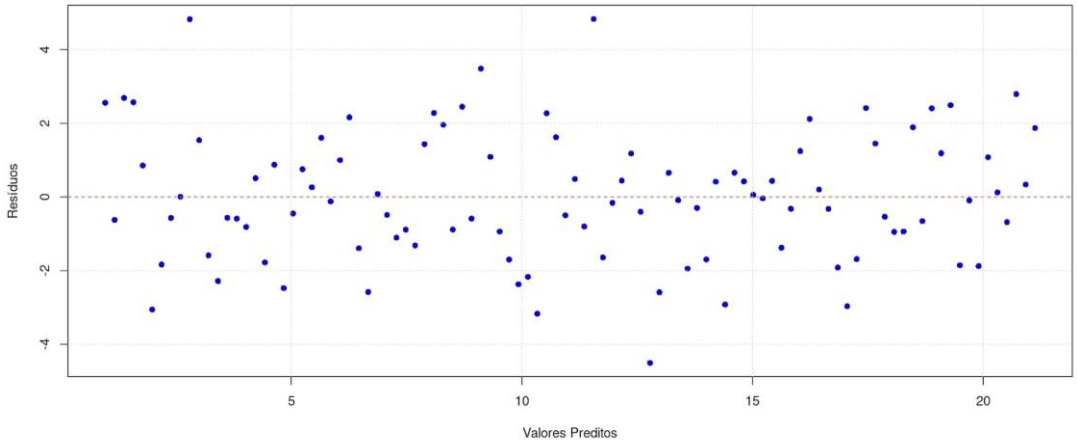
Gráfico 25 - Regressão Linear Homocedastica



Fonte: Dados hipotéticos do autor (2024)

Primeiramente verifiquemos pelo gráfico de resíduos vs valores preditos se há homoscedasticidade ou não.

Gráfico 26 - Resíduos vs Valores Previstos (homoscedasticidade)



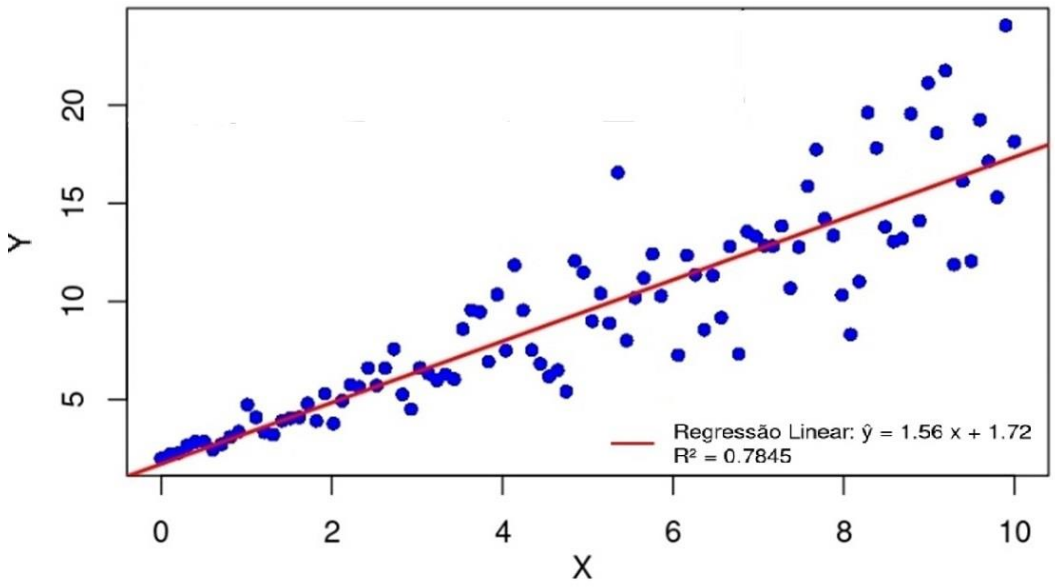
Fonte: Dados hipotéticos do autor (2024)

Observe que não há padrão da distribuição dos pontos no gráfico assim podemos afirmar que existe homoscedasticidade no modelo.

Note que nesse caso em específico como já tínhamos R^2 e seu valor era muito alto, já poderíamos ter afirmado sem usar o gráfico que o modelo possuía homoscedasticidade pois o próprio coeficiente determinação “declara” que ele se adaptou bem as informações.

Agora usaremos o caso seguinte para mostrar como verificar se há heterocedasticidade em um modelo. Veja o modelo a seguir:

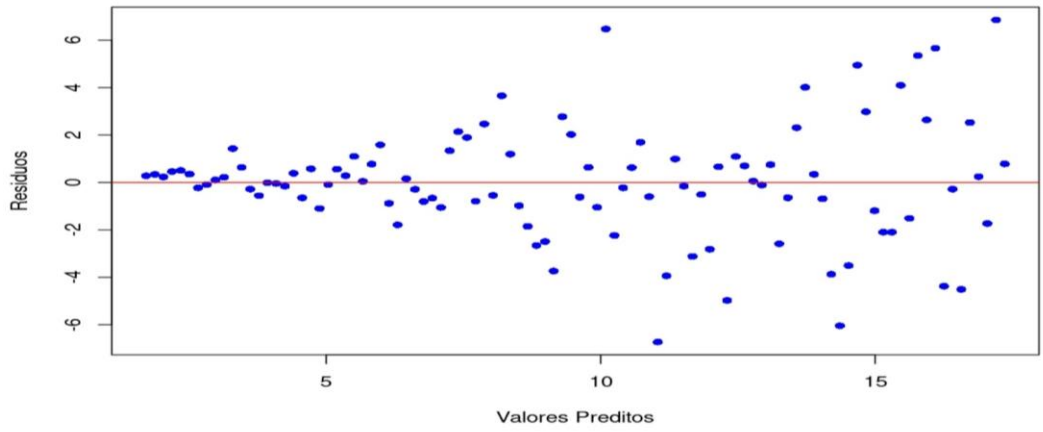
Gráfico 27 - Regressão Linear Heterocedastica



Fonte: Dados hipotéticos do autor (2024)

Vamos agora verificar pelo gráfico de resíduos vs valores preditos se há heterocedasticidade ou não.

Gráfico 28 - Resíduos vs Valores Previstos (Heterocedasticidade)



Fonte: Dados hipotéticos do autor (2024)

Observe que há um padrão (de funil) na distribuição dos pontos no gráfico, assim podemos afirmar que existe heteroscedasticidade no modelo.

Note que nesse caso em específico devido a termos muitos pontos no início do modelo próximo a reta de regressão elevou muito o R^2, mas o existe heteroscedasticidade no modelo conforme identificado, o acarretara dados imprecisos

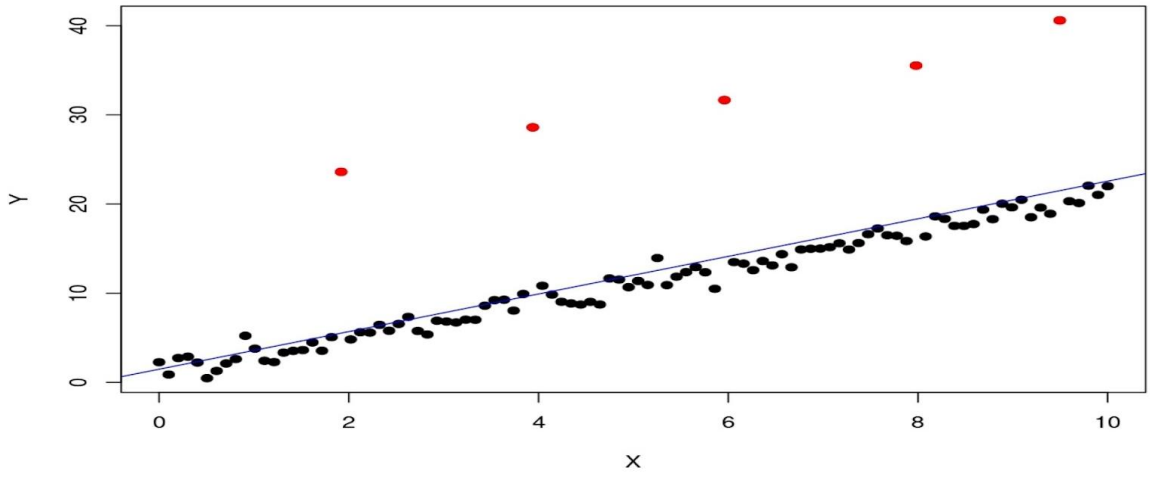
Problemas com a heterocedasticidade pode gerar problemas graves como:

Estimativas de coeficientes enviesada já que os mínimos quadrados podem não ser mais o melhores estimadores não enviesados e o erro padrão da estimativa podem não ser mais preciso; Inferências incorretas como os coeficientes não são mais confiáveis os intervalos de confiança e os testes de hipóteses podem ser distorcidos, levando a conclusões erradas sobre a importância das variáveis independentes; Dificuldades na interpretação já que modelos afetados pela heterocedasticidade podem ser mais difíceis de interpretar, pois a variância dos erros varia de forma não uniforme ao longo do intervalo das variáveis independentes; Problemas de previsão já que os intervalos de confianças podem estar muito amplos. Vale ressaltar que tal interpretação não diferencia para regressão simples ou múltipla

3.3.2 Uso do erro padrão para identificação de outliers

Os resíduos são diferenças entre os valores observados e os valores previstos pelo modelo de regressão, com o resíduo padrão sendo a relação do resíduo pelo desvio padrão de todos os resíduos. Os resíduos padronizados são úteis para identificar observações que estão significativamente distantes da média dos resíduos, o que pode indicar a presença de outliers ou padrões não esperados nos dados já que valores absolutos dos resíduos padronizados maiores que 2 (ou 3, dependendo do contexto) podem indicar a presença de observações atípicas. Veja a influência do outliers em uma regressão linear simples no gráfico as seguir

Gráfico 29 - Regressão Linear Simples (Considerando Outliers)

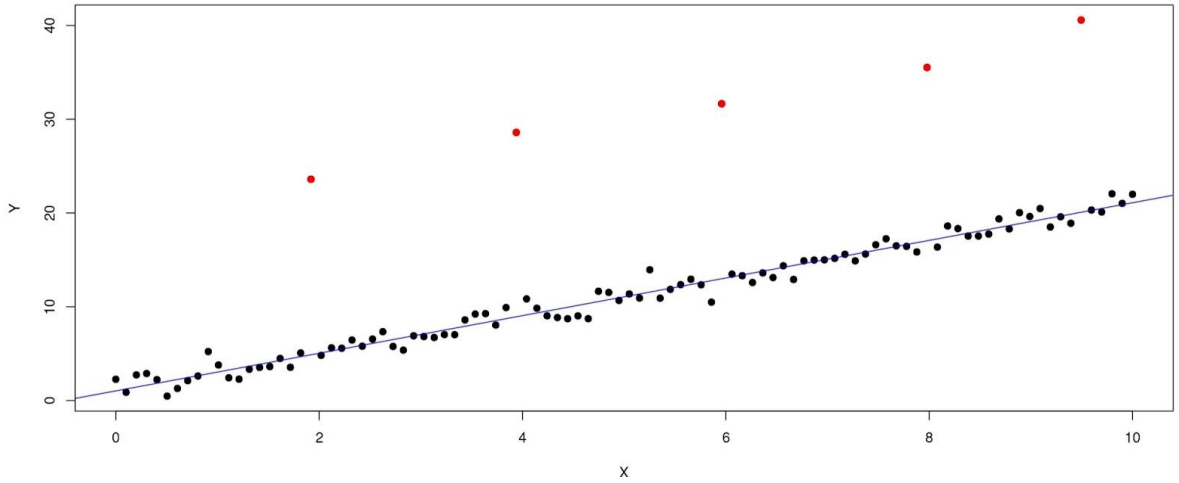


Fonte: Dados hipotéticos do autor (2024)

Veja que eles “puxam” a reta de regressão para cima, logo os valores previstos $\hat{y}$ estarão mais distantes dos valores das amostras coletadas, de maneira que se a levarmos em

consideração enviesaremos o resultado. Perceba o que aconteceria se desconsiderássemos esses Outliers:

Gráfico 30 Regressão Linear Simples (Desconsiderando Outliers)



Fonte: Dados hipotéticos do autor (2024)

Note que a regressão descreveria melhor os dados

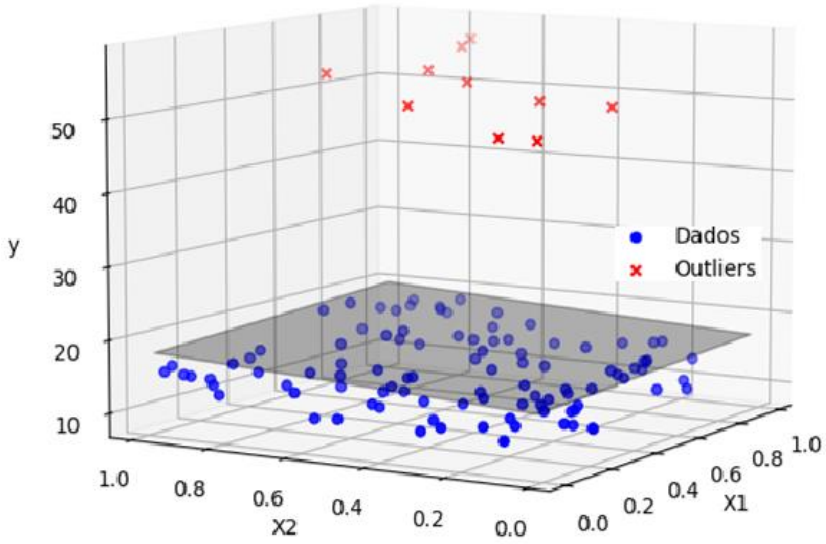
Usaremos o erro padrão para a identificação dos outliers, erro padrão é a medida de dispersão dos resíduos.

Perceba que os únicos valores (ver Apêndice A) que possuem os únicos valores que possuem desvio padrão acima ou abaixo de 2 são os valores na linha 21, 41, 61, 81 e 96 logo eles são outliers

O uso do erro padrão para o encontro outliers não se limita apenas a regressão linear simples o mesmo acontece na regressão múltipla

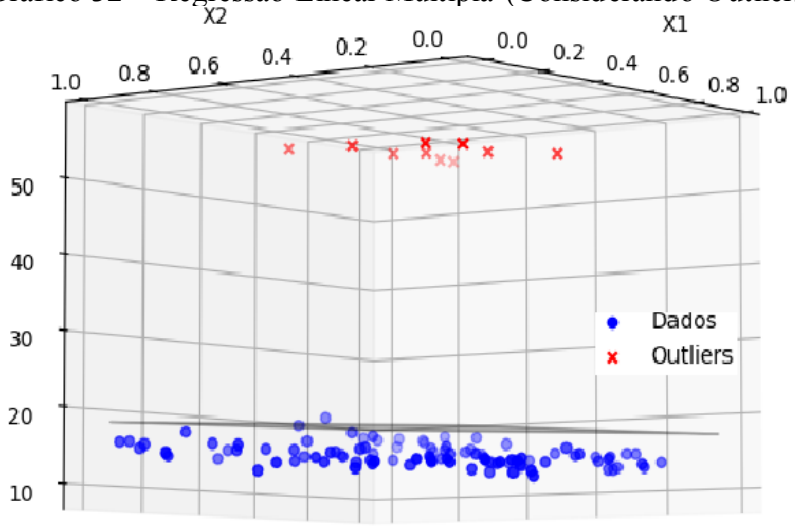
Veja a influência do outliers em uma regressão linear múltipla no gráfico as seguir

Gráfico 31 – Regressão Linear Múltipla (Considerando Outliers)



Fonte: Dados hipotéticos do autor (2024)

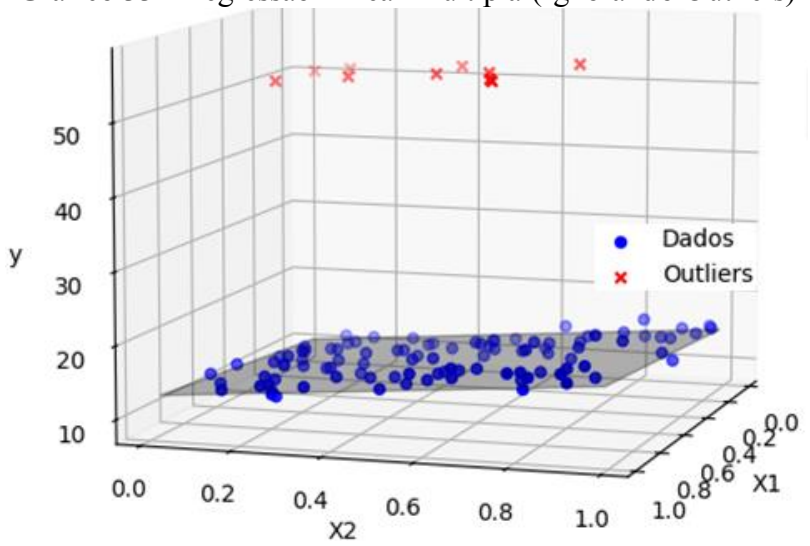
Gráfico 32 – Regressão Linear Múltipla (Considerando Outliers)



Fonte: Dados hipotéticos do autor (2024)

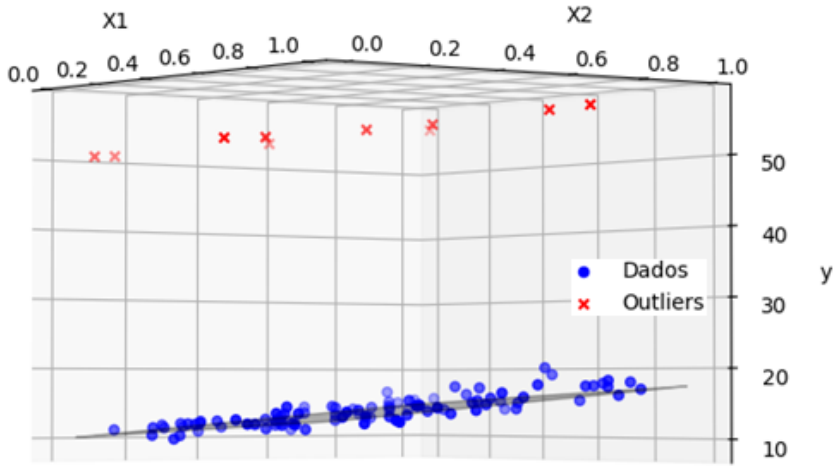
Veja que eles “puxam” o hiperplano de regressão para cima, logo os valores previstos $\hat{y}$ estarão mais distantes dos valores das amostras coletadas, de maneira que se a levarmos em consideração enviesaremos o resultado. Perceba o que aconteceria se ignorássemos esses Outliers:

Gráfico 33 - Regressão Linear Múltipla (Ignorando Outliers)



Fonte: Fonte: Dados hipotéticos do autor (2024)

Gráfico 34 - Regressão Linear Múltipla (Ignorando Outliers)



Fonte: Dados hipotéticos do autor (2024)

Note que a regressão descreveria melhor os dados

Usaremos o erro padrão para a identificação dos outliers. Erro padrão é a medida de dispersão dos resíduos.

Perceba que os únicos valores (ver Apêndice B) que possuem os únicos valores que possuem desvio padrão acima ou abaixo de 2 são os valores na linha 102, 103, 104, 105, 106, 107, 108, 109, 110 e 101, logo eles são outliers.

A presença de outliers no modelo pode ter várias consequências como:

viés nos coeficientes, inflação do erro padrão, baixa eficiência do estimador, modelo não representativo, instabilidade do modelo e violação dos pressupostos fundamentais da regressão

Vale mencionar que se devemos levar em consideração ou não os outliers depende do contexto da análise de dados e do objetivo da modelagem estatística.

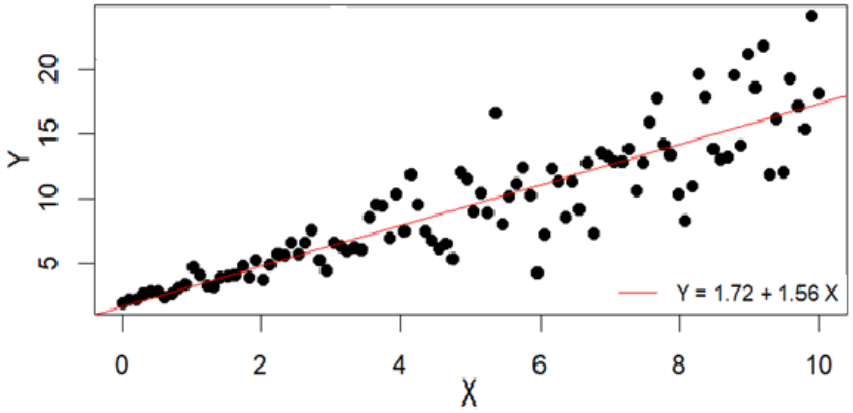
Veja que em uma análise exploratória de dados, durante a análise inicial dos dados, é importante identificar outliers para entender se eles representam erros nos dados ou padrões genuínos; em uma análise de distribuição, ao investigar a distribuição dos dados, outliers podem indicar a presença de caudas pesadas ou comportamentos extremos; em uma modelagem estatística robusta, em certos contextos, como na estimativa de parâmetros usando métodos robustos, é essencial levar em consideração outliers para garantir que o modelo seja resistente a pontos extremos; quando tivermos poucos dados.

3.3.3 Teste Q-Q Plot (Teste de normalidade)

Há a suposição de normalidade dos erros em uma regressão linear, para que as informações obtidas pelo modelo de regressão possam expressar melhor o objeto de estudo, logo se não a normalidade nos erros a inferência com base nos dados será bastante prejudicada. Com base nisso o teste de normalidade consiste em determinar se os resíduos seguem uma distribuição linear.

O teste Q-Q (Quantile-Quantile) é uma ferramenta utilizada para avaliar se uma determinada amostra de dados segue ou não uma distribuição específica, como a distribuição normal. Veja um exemplo a seguir de como identificar a normalidade dos resíduos pelo método Q-Q:

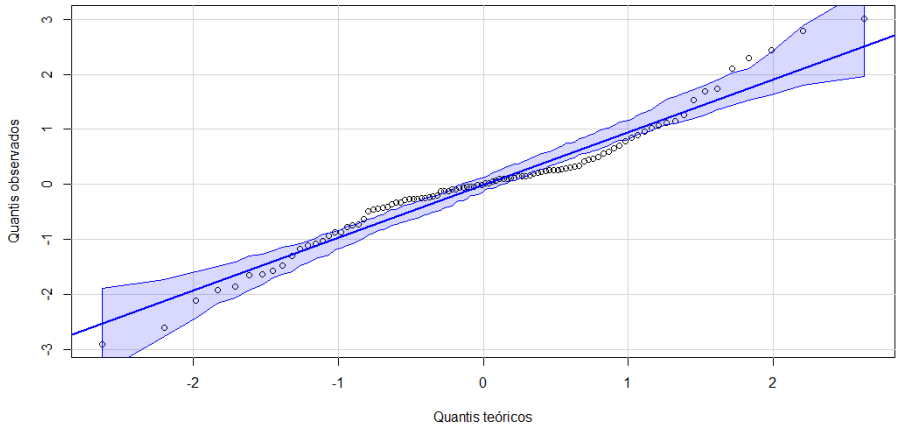
Gráfico 35 - Regressão Linear com Erros Não Normais



Fonte: Dados hipotéticos do autor (2024)

Agora vamos usar o Q-Q para verificar se existe normalidade nos resíduos

Gráfico 36 - Q-Q Plot, com Bada de Confiança, dos Resíduos da Regressão Linear



Fonte: Dados hipotéticos do autor (2024)

Pela quantidade de pontos fora do gráfico Q-Q podemos ver que os resíduos não seguem uma distribuição normal. E por consequência ao analisar os dados obtidos por esse modelo deve-se ter bastante cuidado nas inferências tendo em vista que a violações de suposições elementares da regressão torna os dados, mas suscetíveis a erro.

Vale mencionar que a falta da normalidade dos resíduos ocasiona vários problemas estatísticos como:

- Inferências inválidas: Muitas técnicas de inferência estatística, como intervalos de confiança e testes de hipóteses, pressupõem que os resíduos do modelo são normalmente distribuídos, se essa suposição não for atendida as conclusões tiradas a partir dessas análises podem ser inválidas ou imprecisas;
    - Vieses nos estimadores: A distribuição não normal dos resíduos pode resultar em estimativas enviesadas dos parâmetros do modelo, isso significa que os coeficientes de regressão estimados podem não representar corretamente a relação entre as variáveis independentes e dependentes na população;
    - Diagnóstico de problemas: Os resíduos normalmente são utilizados para diagnosticar problemas no modelo, como, autocorrelação ou influência de outliers se os resíduos não seguem uma distribuição normal, esses diagnósticos podem ser menos confiáveis ou imprecisos;

3.4 Uso da regressão linear para previsão de dados

A regressão linear escabele-se uma relação linear entre variáveis, então podemos usar o modelo obtido para fazer previsões de eventos na situação estudada.

3.4.1 Previsão de dados de uma regressão linear simples

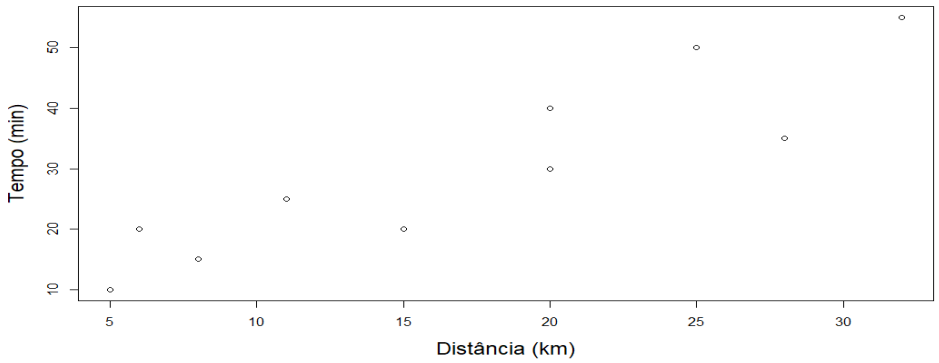
Temos dois conjuntos de dados, tempo (em minutos) e distância percorrida (em km).

Distância(km) = {8, 6,15, 20, 25, 11, 5, 32, 28, 20}

Tempo(min) = {15, 20, 20, 40, 50, 25, 10, 55, 35, 30}

Criaremos um gráfico de dispersão (distância pelo tempo) para verificar a distribuição desses dados no gráfico cartesiano

Gráfico 37 - Gráfico de Dispersão



Fonte: Dados hipotéticos do autor (2024)

Aparentemente, pelo gráfico, pode-se estabelecer uma relação linear entre x e y. Vamos conferir se realmente existe essa relação entre tempo e distância com o coeficiente de Peterson. Calculando o coeficiente obteremos o seguinte resultado:

Tabela 7 – Valores do Coeficientes de Pearson e Valores de x e y da Regressão Simples

Coeficiente de Pearson	0.905221
x	8, 6,15, 20, 25, 11, 5, 32, 28, 20
y	15, 20, 20, 40, 50, 25, 10, 55, 35, 30

Fonte: Dados hipotéticos do autor (2024)

Podemos ver que o valor está bem próximo de um (correlação perfeita) que mostra que existe uma relação linear forte entre tempo e distância. Agora vamos "encontrar" os coeficientes, calculando teremos:

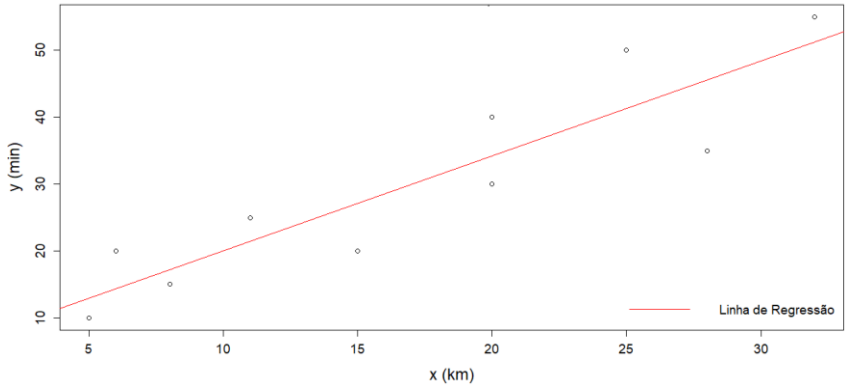
Tabela 8 – Valores β_0 e β_1 e Valores de x e y da Regressão Linear Simples

β_0	5.88
β_1	1.42
x	8, 6,15, 20, 25, 11, 5, 32, 28, 20
y	15, 20, 20, 40, 50, 25, 10, 55, 35, 30

Fonte: Dados hipotéticos do autor (2024)

Logo a função linear é:
$y = 1{,}4189x + 5{,}8784$
Podemos então construir o gráfico contendo a reta de regressão, teremos assim:

Gráfico 38 - Regressão Linear Simples



Fonte: Dados hipotéticos do autor (2024)

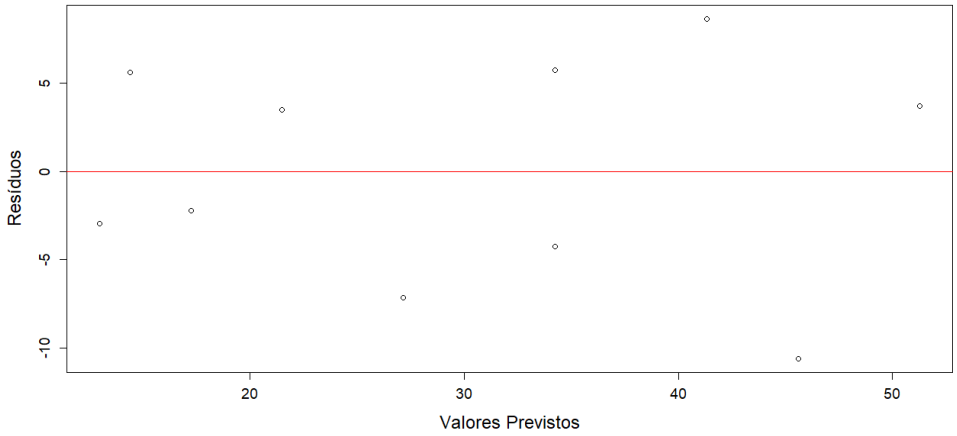
Vamos encontrar o R^2 para ver o quanto variabilidade no tempo pode ser explicada pela distância no modelo ajustado

$Calculando$ o coeficiente de determinação para o modelo de regressão linear simples acharemos 0.8194. Logo, cerca de 81.94% da variabilidade no tempo pode ser explicada pela distância no modelo ajustado.

Agora vamos conferir os pressupostos da regressão linear para que possamos ter a certeza que nossas previsões serão precisas.

Vamos construir o gráfico de resíduos versus valores previstos para verificar a homoscedasticidade.

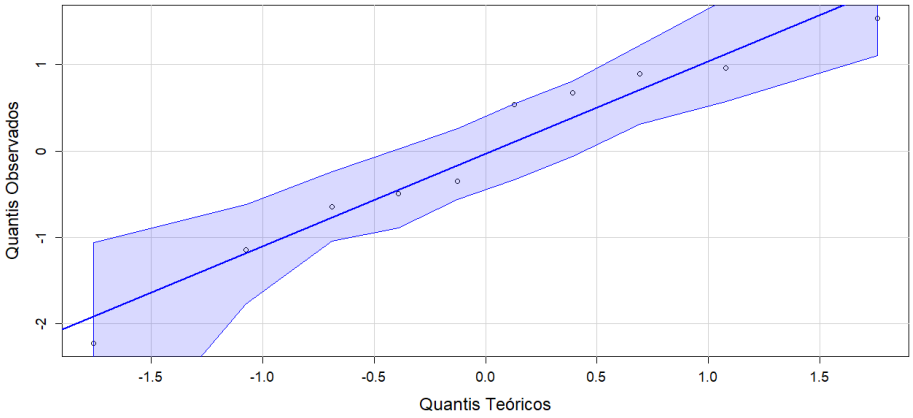
Gráfico 39 - Resíduos vs Valores Previstos



Fonte: Dados hipotéticos do autor (2024)

Como não há padrão entre os pontos podemos inferir que há homoscedasticidade. Vamos agora conferir se a normalidade dos resíduos com o gráfico Q-Q

Gráfico 40 - Q-Q Plot com Banda de Confiança



Fonte: Dados hipotéticos do autor (2024)

Vemos que os pontos estão dentro do intervalo de confiança e perto da reta de referência da distribuição normal, logo os resíduos seguem uma distribuição normal.

Tendo em vista a baixa quantidade de dados, a busca e eliminação de outliers se torna não recomendável, logo não será feita.

Agora vamos inserir um intervalo de confiança para termos uma visão da variabilidade dos dados e da confiabilidade das conclusões tendo em vista que nem todos têm o mesmo percentual de confiabilidade. Calculando o intervalo de confiança de cada coeficiente teremos a seguinte tabela:

Tabela 9 – Valores do Intervalo de Confiança dos Coeficientes β_0 e β_1 dá Regressão Linear Simples

Coeficiente	Intervalo de Confiança
β_0	[- 4,5731; 16,3299]
β_1	[0,8758; 1,9619]

Fonte: Dados hipotéticos do autor (2024)

Agora além de obtermos os coeficientes, podemos concluir que os parâmetros β_0 e β_1 que assumem intervalos de [- 4,5731; 16,3299] e [0,8758; 1,9619], respectivamente, ao nível de confiança de 95%. Sendo assim, as equações que determinam os valores mínimo e máximo do tempo de percurso para este nível de são:

- $TempoMinimo = -4{,}5731 + 0{,}8758 \cdot x$
- $TempoMáximo = 16{,}3299 + 1{,}9619 \cdot x$

Veja o exemplo:

Tempo mínimo a uma distância de 17 km:

$TempoMinimo = -4{,}5731 + 0{,}8758 \cdot x$ = - 4,5731 + 0,8758 ·(17) = 10,3155 minutos

Tempo máximo a uma distância de 17 km:

$TempoMáximo = 16{,}3299 + 1{,}9619 \cdot x$ = 16,3299 + 1,9619 ·(17) = 49,6822 minutos

Logo, podemos dizer que há 95% de confiança de que uma pessoa que percorre 17 quilômetros para chegar ao local desejado leve entre 10,3155 minutos e 49,6822 minutos, com tempo médio estimado de 29,9997 minutos.

Com tudo obtido e verificado vamos fazer previsões do tempo em minutos que leva para percorrer uma distância qualquer em quilômetros.

Vamos fazer a previsão de quanto tempo leva para as seguintes distancia em quilômetros: 111, 83, 123, 92, 118, 103, 129, 98, 134, 101, 114, 125, 122, 129, 81, 132, 86, 130, 94 e 112. Essa previsão nos leva aos seguintes valores:

Tabela 10 – Valores da Previsão da Regressão Linear Simples

x	Val_Regressão	Min_Regressão	Max_Regressão	Médio_Regressão
111	163.3763	92.6407	234.1008	163.37075
83	123.6471	68.1183	179.1676	123.64295
123	180.4031	103.1503	257.6436	180.39695
92	136.4172	76.0005	196.8247	136.4126
118	173.3086	98.7713	247.8341	173.3027
103	152.0251	85.6343	218.4056	152.01995
129	188.9165	108.4051	269.415	188.91005
98	144.9306	81.2553	208.5961	144.9257
134	196.011	112.7841	279.2245	196.0043
101	149.1873	83.8827	214.4818	149.18225
114	167.633	95.2681	239.9865	167.6273
125	183.2409	104.9019	261.5674	183.23465
122	178.9842	102.2745	255.6817	178.9781
129	188.9165	108.4051	269.415	188.91005
81	120.8093	66.3667	175.2438	120.80525
132	193.1732	111.0325	275.3007	193.1666
86	127.9038	70.7457	185.0533	127.8995
130	190.3354	109.2809	271.3769	190.3289
94	139.255	77.7521	200.7485	139.2503
112	164.7952	93.5165	236.0627	164.7896

Fonte: Dados hipotéticos do autor (2024)

Veja que podemos fazer previsões de quanto tempo podemos levaremos para andarmos uma dada distância e ainda deduzir o tempo máximo e mínimo para tal bem como o seu tempo médio.

3.4.2 Previsão de dados de uma regressão linear múltipla:

Temos três conjuntos de dados eles são distância (quilometro), semáforo e tempo (minutos).

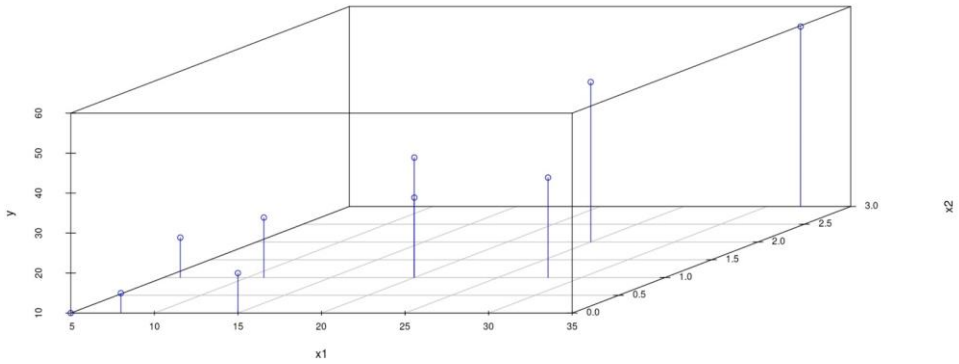
Distância(km) = {8, 6, 15, 20, 25, 11, 5, 32, 28, 20}

Semáforo = {0, 1, 0, 1, 2, 1, 0, 3, 1, 1}

Tempo(minutos) = {15, 20, 20,40, 50, 25, 10, 55, 35, 30}

Criemos primeiramente um gráfico com os dados acima, com x1 sendo a distância; x2 sendo a quantidade de semáforos e y sendo o tempo

Gráfico 41 - Gráfico de Dispersão 3D



Fonte: Dados hipotéticos do autor (2024)

Por se tratar de uma regressão múltipla não usaremos o coeficiente de Pearson para conferir a linearidade e sim usaremos o $R^2{}_{ajustado}$, só que mais adiante.

Agora vamos construir o hiperplano da regressão, para isso vamos calcular os coeficientes

Após calcular os coeficientes teremos os seguintes valores:

Tabela 11 – Valores de β_0, β_1e β_2 e Valores de x_1, x_2 e y da Regressão Linear Múltipla

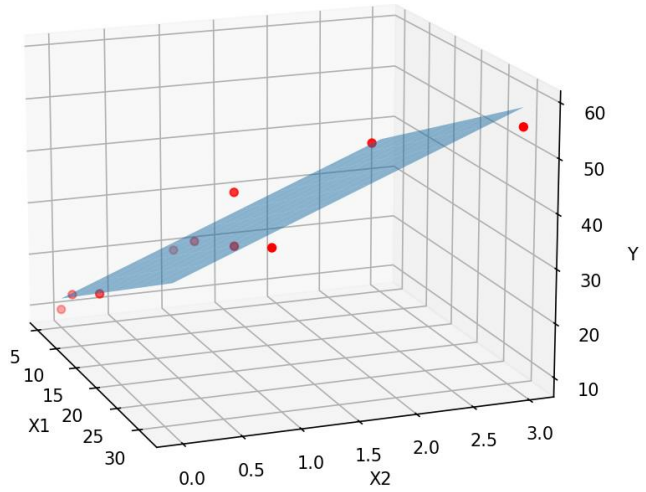
β_0	8.15
β_1	0.797
β_2	8.3
x_1	8, 6, 15, 20, 25, 11, 5, 32, 28, 20
x_2	0, 1, 0, 1, 2, 1, 0, 3, 1, 1
y	15, 20, 20,40, 50, 25, 10, 55, 35, 30

Fonte: Dados hipotéticos do autor (2024)

Logo, a reta de regressão é $y = 8{,}17 + 0.797x_1 + 8.3x_2$

Assim podemos então afirmar que o gráfico com o hiperplano é o seguinte:

Gráfico 42 – Regressão Linear Múltipla



Fonte: Dados hipotéticos do autor (2024)

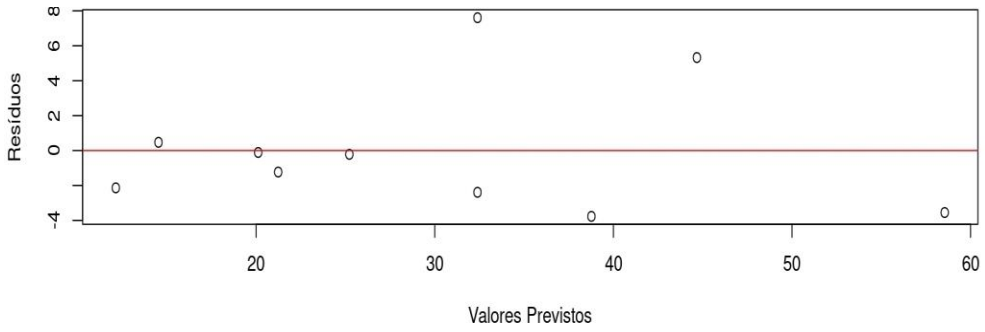
Aparentemente, pelo gráfico, pode-se estabelecer uma relação linear entre x_1, x_2 e y. Vamos conferir se realmente existe essa relação entre tempo, quantidade de semáforos e distância com o $R^2{}_{ajustado}$.

O valor do $R^2{}_{ajustado}$ para o modelo de regressão dado é aproximadamente 0.9195. Isso indica que o modelo explica cerca de 91.95% da variabilidade dos dados ajustados, levando em conta o número de preditores no modelo, logo podemos afirmar que de fato há relação linear entre as variáveis independentes e a dependente

Agora vamos conferir os pressupostos da regressão linear para que possamos ter a certeza de que nossas previsões serão precisas.

Vamos construir o gráfico de resíduos versus valores previstos para verificar há homoscedasticidade.

Gráfico 43 - Gráfico de Resíduos vs Valores Previstos

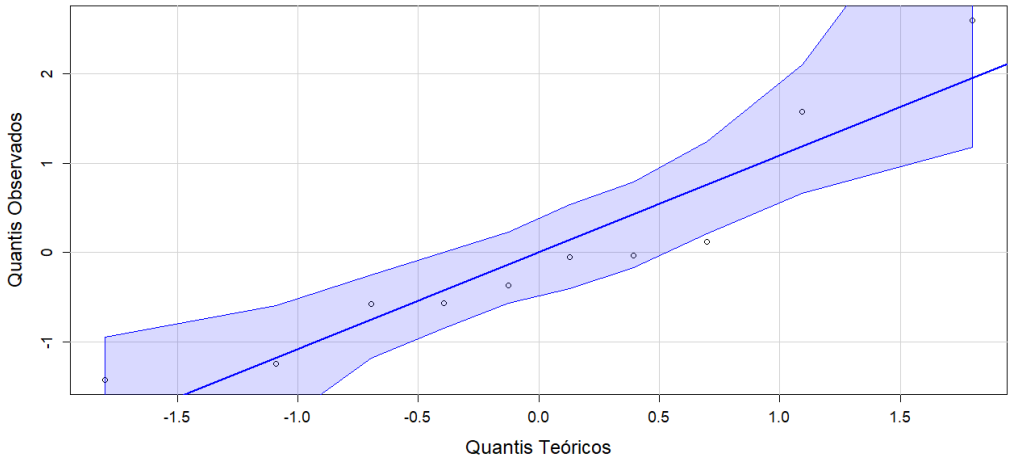


Fonte: Dados hipotéticos do autor (2024)

Como não há padrão entre os pontos podemos inferir que há homoscedasticidade.

Vamos agora conferir se há normalidade dos resíduos com o gráfico Q-Q

Gráfico 44 - Gráfico Q-Q Plot com Banda de Confiança



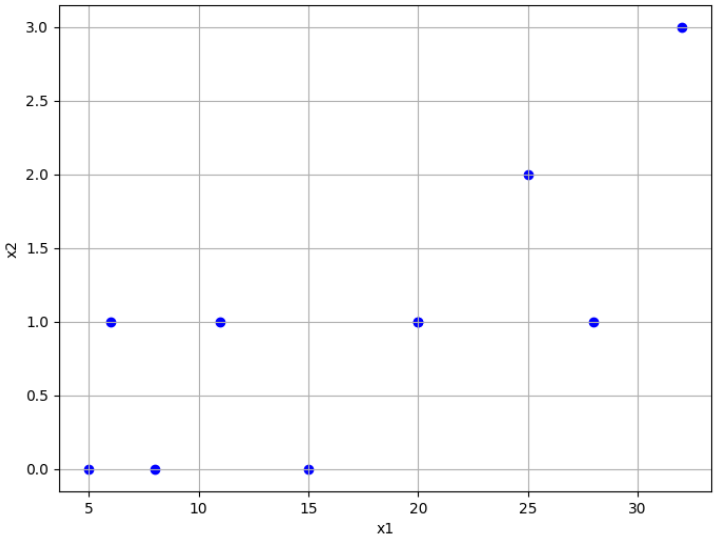
Fonte: Dados hipotéticos do autor (2024)

Vemos que os pontos estão dentro do intervalo de confiança e perto da reta de referência da distribuição normal, logo os resíduos seguem uma distribuição normal.

Devido a quantidade pequena de dados se torna inviável a eliminação de outliers.

Para verificar se existe relação entre as variáveis explicativa normalmente se usaria a matriz de dispersão ou a matriz de covariância, mas como só temos duas variáveis explicativas e há poucos dados, assim será usada apenas um gráfico de dispersão de x_1 pôr x_2

Gráfico 45 - Gráfico de Dispersão de x1 por x2



Fonte: Dados hipotéticos do autor (2024)

Podemos Ver que não existe relação entre as variáveis independentes

Agora vamos inserir um intervalo de confiança para termos uma visão da variabilidade dos dados e da confiabilidade das conclusões tendo em vista que nem todos têm o mesmo percentual de confiabilidade.

Calculando o intervalo de confiança teremos de cada coeficiente teremos a seguinte tabela:

Tabela 12 – Valores do Intervalo de Confiança dos Coeficientes da Regressão Linear Múltipla

Coeficiente	Intervalo de Confiança
β_0	[1,2462920; 15,056109]
β_1	[0,2619049; 1,332506]
β_2	[2,8966699; 13.695949],

Fonte: Dados hipotéticos do autor (2024)

Agora além de obtermos os coeficientes, podemos concluir que os parâmetros β_0, β_1 e β_2 que assumem intervalos de [1,2462920; 15,056109], [0,2619049; 1,332506] e [2,8966699; 13.695949], respectivamente, ao nível de confiança de 95%. Sendo assim, as equações que determinam os valores mínimo e máximo do tempo de percurso para este nível de são:

- $TempoMinimo = 1{,}2462920 + 0{,}2619049 \cdot x_1 + 2{,}8966699 \cdot x_2$
- $TempoMáximo = 15{,}056109 + 1{,}332506 \cdot x_1 + 13.695949 \cdot x_2$

Veja o exemplo:

Tempo Mínimo a uma distância de 17 km com 1 semáforo:

$TempoMinimo = 1{,}2462920 + 0{,}2619049 \cdot x_1 + 2{,}8966699 \cdot x_2 = 1{,}2462920 + 0{,}2619049 \cdot 17 + 2{,}8966699 \cdot 1 =$ 8,5972869 minutos

Tempo Máximo a uma distância de 17 km com 1 semáforo:

$TempoMáximo = 15{,}056109 + 1{,}332506 \cdot x_1 + 13.695949 \cdot x_2 = 15{,}056109 + 1{,}332506 \cdot 17 + 13.695949 \cdot 1 = 51{,}405660 minutos$

Logo, podemos dizer que há 95% de confiança de que uma pessoa que percorre 17 quilômetros para chegar ao local desejado leve entre 8,5972869 minutos e $51{,}405660 minutos$, com tempo médio estimado de 29,50147395 minutos.

Com tudo obtido e verificado vamos fazer previsões do tempo em minutos que leva para percorrer uma distância qualquer em quilômetros com uma determinada quantidade de semáforo.

Vamos fazer a previsão de quanto tempo leva para as seguintes distancia em quilômetros: 101, 89, 92, 122, 140, 123, 108, 83, 105, 85, 115, 121, 139, 130, 87, 88, 143, 136, 101 e 129. com seguinte quantidade de semáforos: 2, 1, 0, 3, 4, 0, 3, 1, 0, 1, 2, 1, 2, 4, 3, 4, 1, 3, 2 e 0. Essa previsão nos leva aos seguintes valores:

Tabela 13 – Valores das Predições da Regressão Linear Múltipla

x_1	x_2	Val_Regressão	Min_Regressão	Max_Regressão	Médio_Regressão
101	2	105.2615553	33.4920267	177.031113	105.26156985
89	1	87.3987817	27.452498	147.345092	87.398795
92	0	81.4940876	25.3415428	137.646661	81.4941019
122	3	130.2991766	41.8886995	218.709688	130.29919375
140	4	152.945182	49.4996576	256.390745	152.9452013
123	0	106.2074519	33.4605947	178.954347	106.20747085
108	3	119.1383024	38.2220309	200.054604	119.13831745
83	1	82.6155499	25.8810686	139.350056	82.6155623
105	0	91.8577565	28.7463065	154.969239	91.85777275
85	1	84.2099605	26.4048784	142.015068	84.2099732
115	2	116.4224295	37.1586953	195.686197	116.42244615
121	1	112.9093513	35.8334548	189.985284	112.9093694
139	2	135.5553567	43.4444129	227.666341	135.55537695
130	4	144.973129	46.8806086	243.065685	144.9731468
87	3	102.3969911	32.722028	172.071978	102.397003
88	4	111.4905064	35.8806028	187.100433	111.4905179
143	1	130.4478679	41.5953626	219.300416	130.4478893
136	3	141.4600508	45.5553681	237.364772	141.46007005
101	2	105.2615553	33.4920267	177.031113	105.26156985
129	0	110.9906837	35.0320241	186.949383	110.99070355

Fonte: Dados hipotéticos do autor (2024)

Veja que podemos fazer previsões de quanto tempo podemos levaremos para andarmos uma dada distância e ainda deduzir o tempo máximo e mínimo para tal bem como o seu tempo médio.

3.5 Aplicação da regressão linear em uma análise de caso real:

Caso e baco de dados abordados obtido de Kumar (2018)

Motivação:

Uma empresa automobilística chinesa Geely Auto aspira entrar no mercado dos

EUA instalando sua unidade de fabricação lá e produzindo carros localmente para dar concorrência aos seus equivalentes dos EUA e da Europa.

Eles contrataram uma empresa de consultoria automotiva para entender os fatores dos quais depende o preço dos carros. Especificamente, eles querem entender os fatores que afetam o preço dos carros no mercado americano, uma vez que podem ser muito diferentes do mercado chinês. A empresa quer saber:

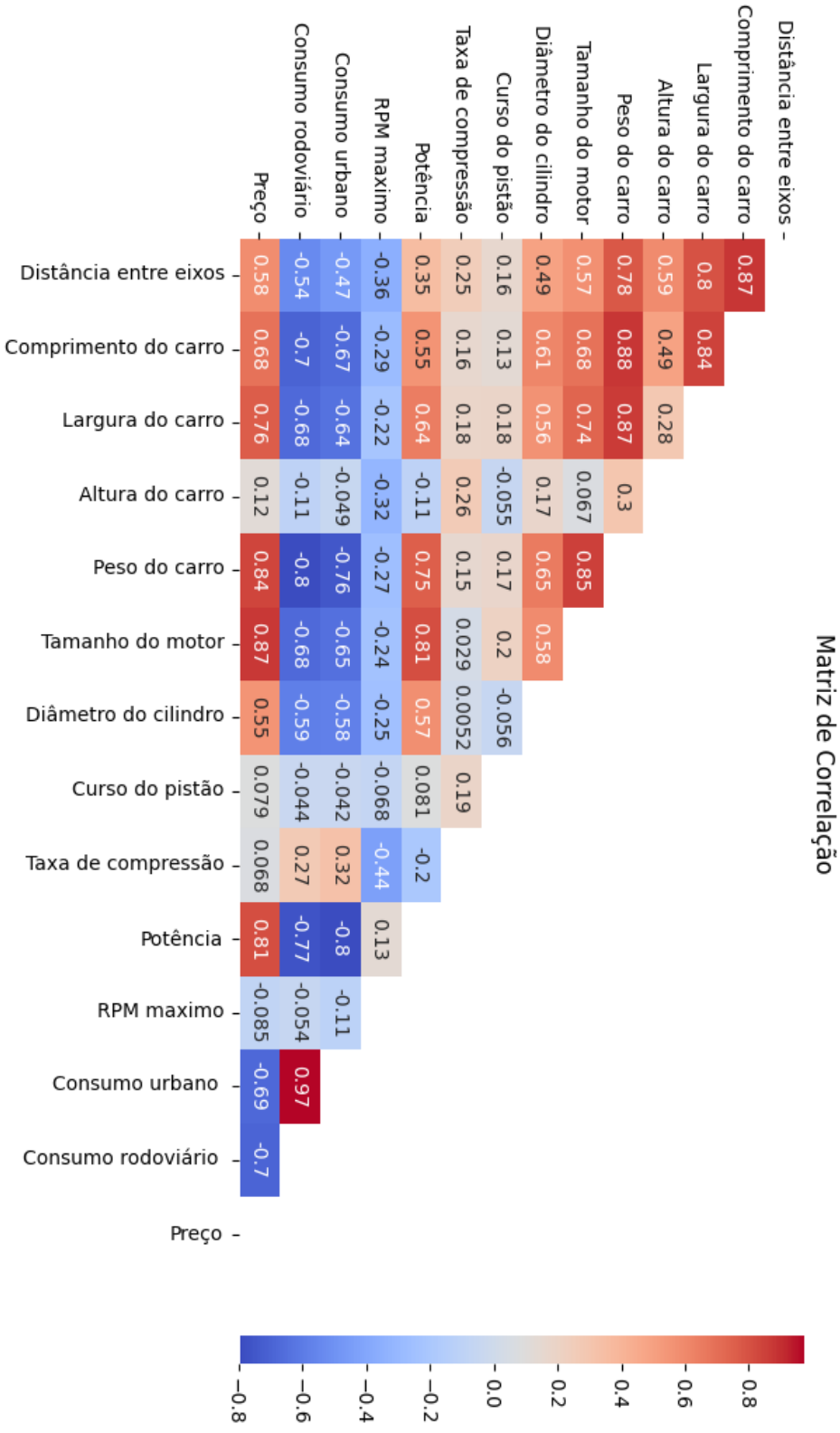
- Quais variáveis são significativas na previsão do preço de um carro
- Quão bem essas variáveis descrevem o preço de um carro

Com base em várias pesquisas de mercado, a empresa de consultoria reuniu um grande conjunto de dados de diferentes tipos de carros no mercado americano.

Objetivo de negócios:

Somos obrigados a modelar o preço dos carros com as variáveis independentes disponíveis. Ele será usado pela gerência para entender exatamente como os preços variam com as variáveis independentes. Eles podem, portanto, manipular o design dos carros, a estratégia de negócios etc. para atender a certos níveis de preço. Além disso, o modelo será uma boa maneira para a gerência entender a dinâmica de preços de um novo mercado.

Gráfico 46 – Matriz de Correlação



Fonte: Dados hipotéticos do autor (2024)

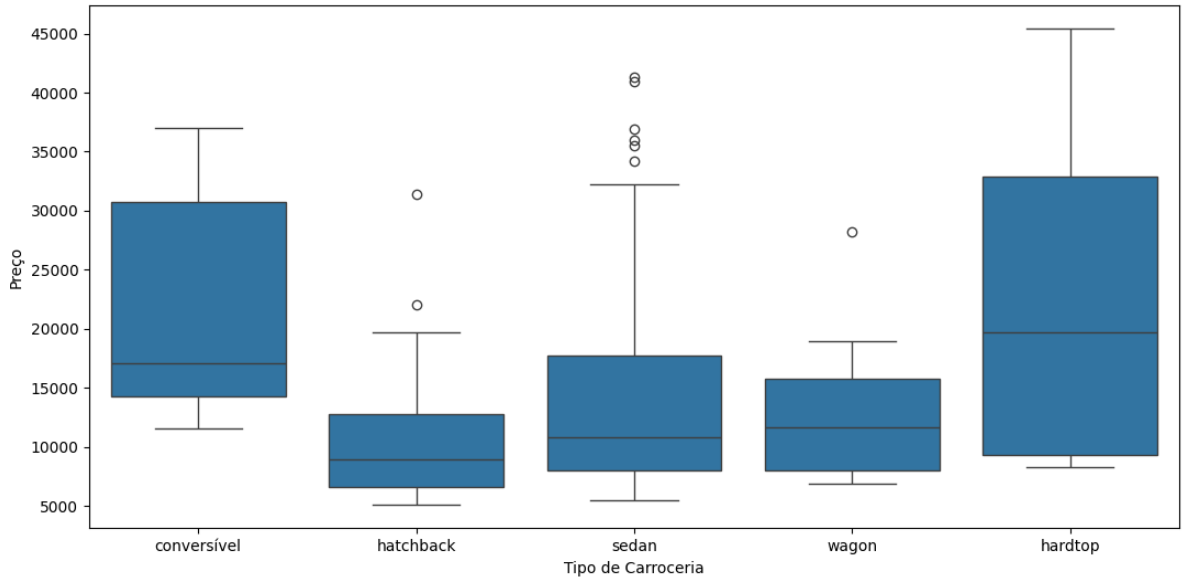
A largura, o comprimento, o peso bruto, o tamanho do motor e a potência do carro apresentam uma ‘boa” correlação positiva com o preço. Em outras palavras, quanto maiores esses atributos, mais caro será o carro. Por outro lado, o consumo rodoviário e o consumo urbano têm uma “boa” correlação negativa, ou seja, quanto maior o consumo na rodovia, mais baixo será o preço do carro. De acordo com a regressão linear, essas são as variáveis que influenciam significativamente o preço final.

Observe que o consumo rodoviário e urbano, bem como a largura e o comprimento do carro, apresenta uma alta correlação entre si. Inclui-las na regressão linear pode inflacionar os valores desnecessariamente. Portanto, vamos excluir o comprimento do carro e o consumo urbano da previsão dos preços. Optamos por manter as outras duas variáveis devido as suas correlações serem mais forte com o preço.

Perceba que variáveis como curso do pistão, taxa de compressão, altura do carro, distância entre eixos, rpm máximo e diâmetro do cilindro têm uma correlação baixa com o preço, assim na regressão linear elas terão pouca influência no valor final, desta forma podemos descartá-las da nossa análise.

Agora vamos analisar as variáveis categóricas em relação aos preços usando Box Plot, com o ele podemos ter uma boa visualização entre a relação dos elementos abordados e o preço, além de que se torna bem claro a presença de outliers

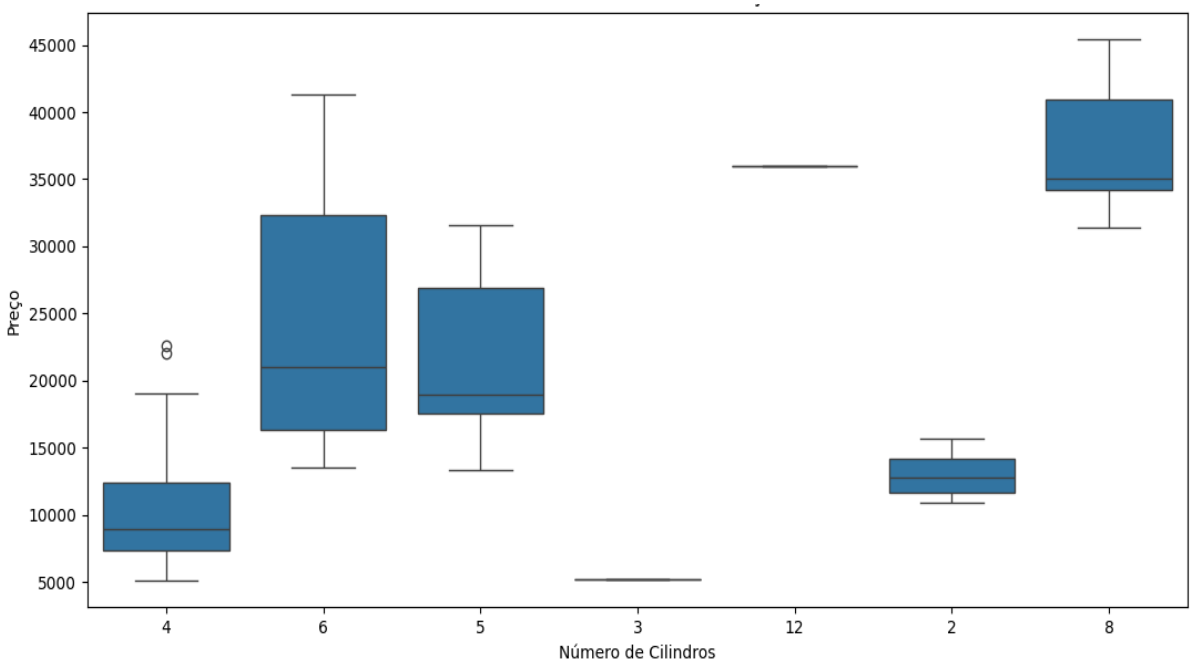
Gráfico 47 – Tipo de Carroceria vs Preço



Fonte: Autor (2024)

Todos os tipos de carroceria são relativamente mais baratos em comparação com carrocerias conversíveis e hardtop, contudo são aquelas com menos quantidades de carros, além de quê, a categoria sedan é tipo de carroceria com mais carros e tem uma grande quantidade de outliers. Assim podemos afirmar que sua inclusão no modelo será prejudicial já que a regressão linear é muito sensível outliers e uma quantidade desbalanceada de dados gera instabilidade nos coeficientes, logo pela regressão ela não é deve ser usada na previsão do preço

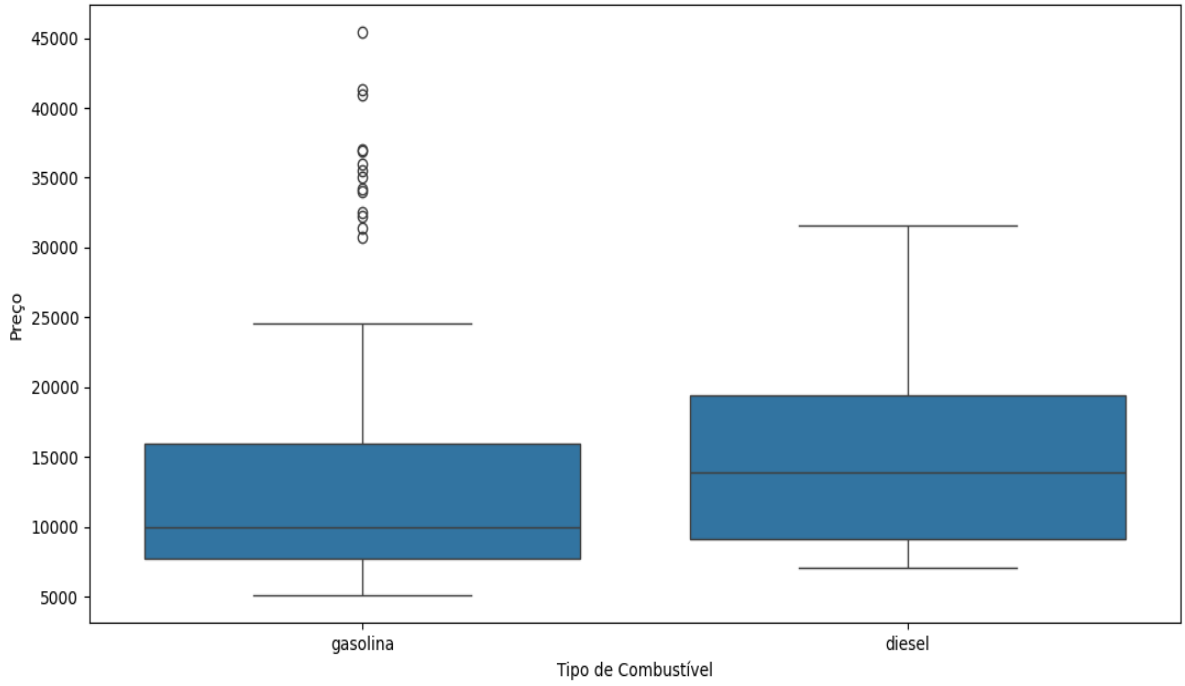
Gráfico 48 – Número Cilindros vs Preço



Fonte: Autor (2024)

Quanto mais cilindros mais prováveis é de o preço ser maior do que o de cilindragem menor, contudo a grande maioria dos carros são 4 cilindros ao considerar as outras variáveis a regressão estarão comprometidas, pois o preço com os outros cilindros será influenciado pelos de 4 cilindros, logo o número de cilindros não deve ser considerado ao calcular os preços dos carros

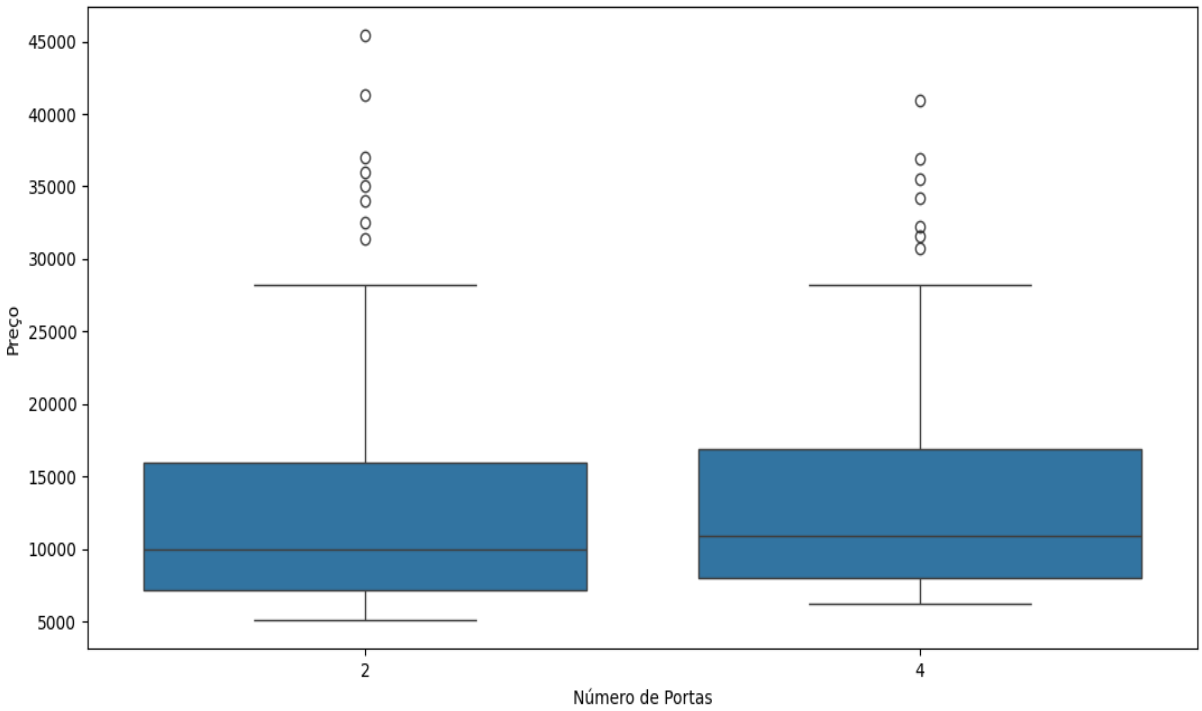
Gráfico 49 – Tipo de Combustível vs Preço



Fonte: Autor (2024)

Carros com combustível a diesel são comparativamente mais caros do que carros com a gasolina, contudo quantidade de carros à diesel é bem pequena compara a aos movidos a gasolina também pode-se notar a quantidade de outliers de veículos gasolina é bem grande. Assim podemos afirmar que sua inclusão no modelo será prejudicial já que a regressão linear é muito sensível outliers e uma quantidade desbalanceada de dados gera instabilidade nos coeficientes, logo pela regressão ela não é deve ser usada na previsão do preço

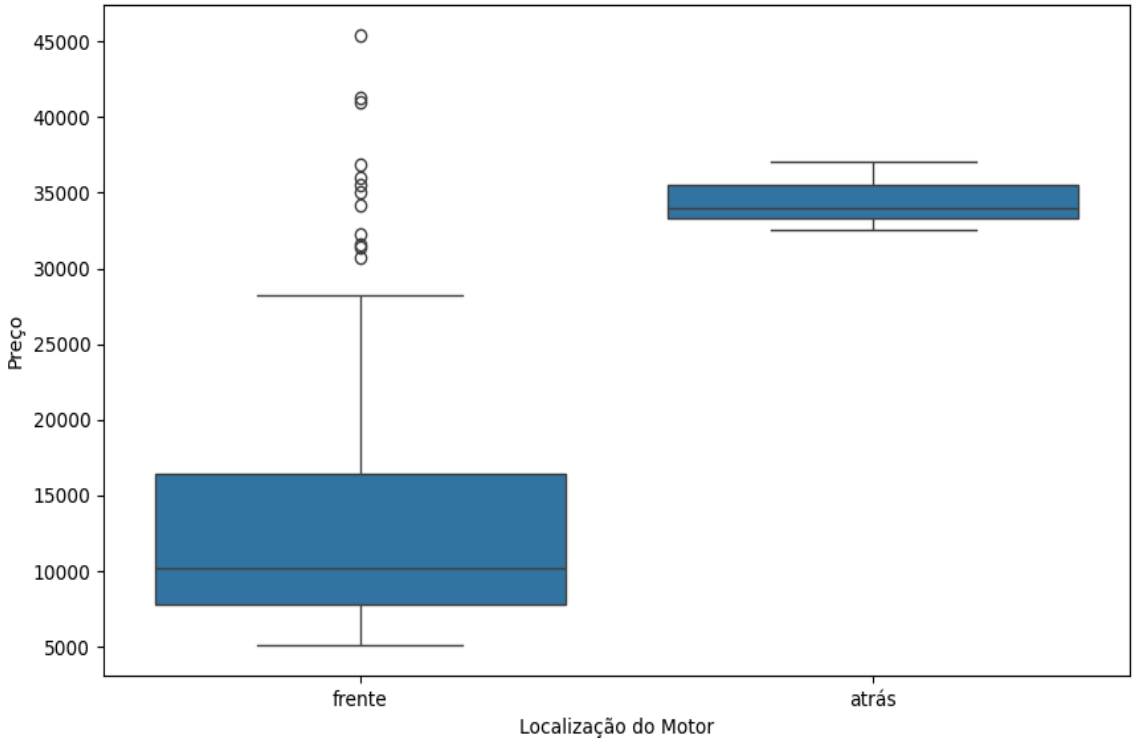
Gráfico 50 - Número de Portas vs Preço



Fonte: Autor (2024)

O preço do carro com 4 portas é um pouco mais elevado. Embora tenham quantidade semelhante de veículos com 2 e 4 portas a presença de outliers e a pouca diferença no preço inviabiliza seu uso na regressão linear, logo não deve ser considerado no cálculo do preço dos veículos

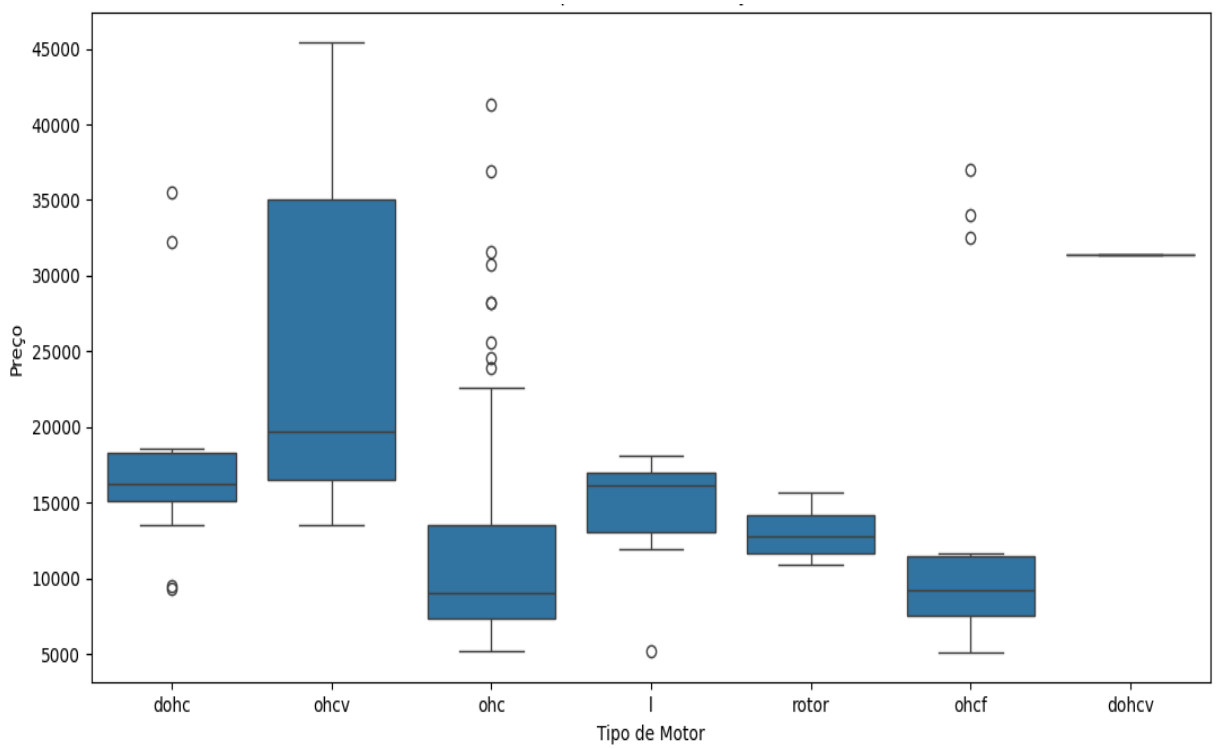
Gráfico 51 – Localização do Motor vs Preço



Fonte: Autor (2024)

Os carros com motor traseiro são muito mais caros do que carros com motor dianteiro, contundo a quantidade de carros com motor traseiro é muito pequena e quando se trata-se de carro com motores na frente a quantidade de outliers é muito grande. Assim podemos afirmar que sua inclusão no modelo será prejudicial já que a regressão linear é muito sensível outliers e uma quantidade desbalanceada de dados gera instabilidade nos coeficientes, logo pela regressão ela não é deve ser usada na previsão do preço.

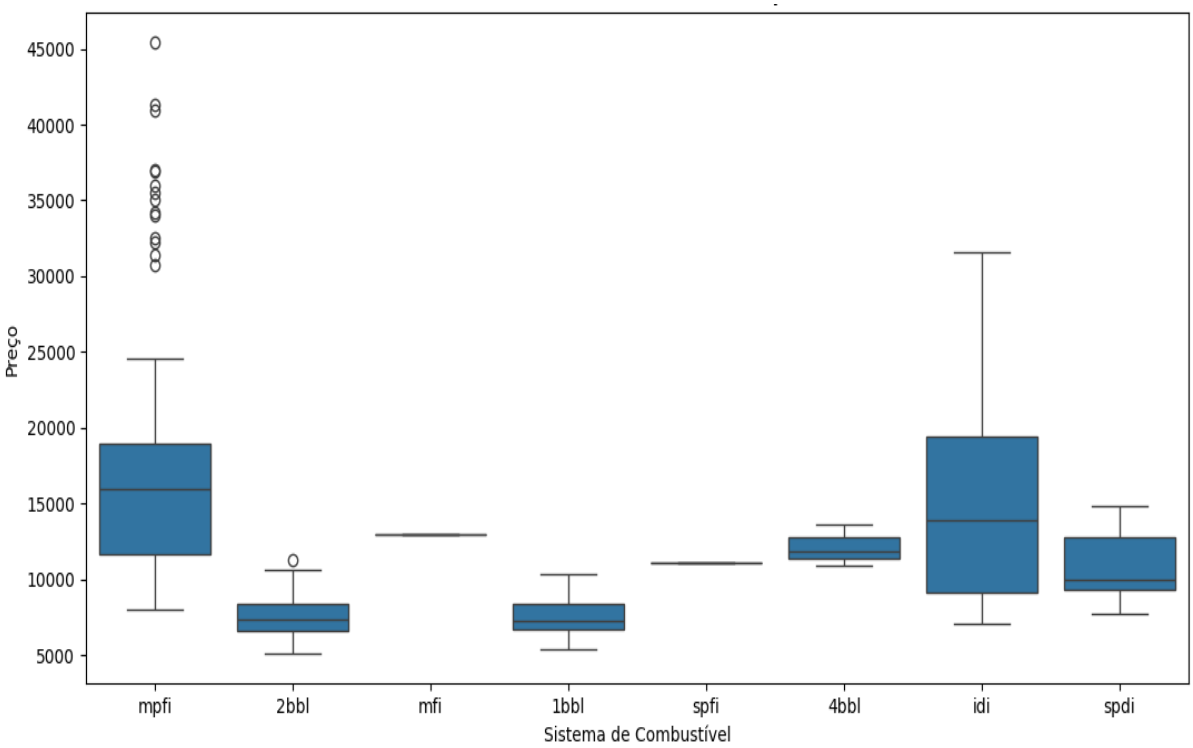
Gráfico 52 – Tipo de Motor vs Preço



Fonte: Dados hipotéticos do autor (2024)

O tipo de motor ohcv entra em carros de faixa de preço mais alta, contudo a quantidade de veículos está bem distante do motor que equipa a maioria dos carros que é o ohc, além de que que o moto ohc possui muitos outliers. Assim podemos afirmar que sua inclusão no modelo será prejudicial já que a regressão linear é muito sensível outliers e uma quantidade desbalanceada de dados gera instabilidade nos coeficientes, logo pela regressão ela não é deve ser usada na previsão do preço.

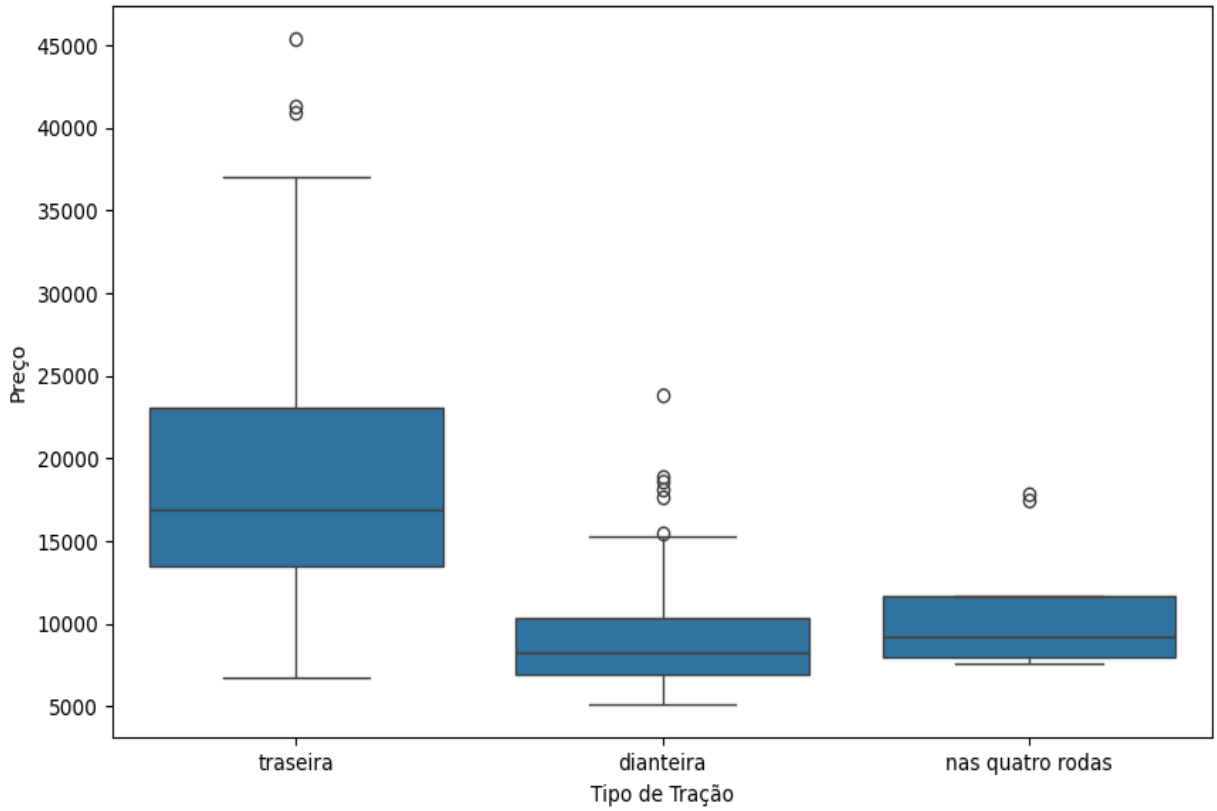
Gráfico 53 – Fonte: Dados hipotéticos do autor (2024)



Sistema de Combustível vs Preço

Os carros com sistemas de combustíveis mpfi e idi são os possuem os preços mais elevados, contudo os sistemas de combustíveis mais usadas são 2bbl e mpfi, além de que o sistema mpfi tem uma grande quantidade de outliers. Assim podemos afirmar que sua inclusão no modelo será prejudicial já que a regressão linear é muito sensível outliers e uma quantidade desbalanceada de dados gera instabilidade nos coeficientes, logo pela regressão ela não é deve ser usada na previsão do preço.

Gráfico 54 – Tipo de Tração vs Preço



Fonte: Dados hipotéticos do autor (2024)

Carro com tração traseira são mais os caros, no entanto eles não são os mais numerosos logo os valores das outras trações seriam influenciados por eles e ainda a grande quantidade de outliers. Assim podemos afirmar que sua inclusão no modelo será prejudicial já que a regressão linear é muito sensível outliers e uma quantidade desbalanceada de dados gera instabilidade nos coeficientes, logo pela regressão ela não é deve ser usada na previsão do preço.

Por meio da análise, podemos verificar que, embora as variáveis categóricas influenciem o preço, elas não são os principais determinantes no valor do carro. Utilizando a regressão linear identificamos que as seguintes variáveis têm uma influência significativa sobre o preço do carro: peso bruto, tamanho do motor, potência e consumo na estrada. Ao calcular o coeficiente de determinação ajustado, obtivemos um valor aproximado de 0,9. Isso indica que essas variáveis explicam aproximadamente 90% do comportamento do preço do carro

4 CONCLUSÃO

Conforme foi discutido no tópico 3.1 a regressão pode "mostrar" a existência de uma relação entre um conjunto de variáveis independentes e uma variável dependente já no tópico 3.2 observamos que, através dos coeficientes da regressão, é possível determinar se as variáveis irão crescer ou diminuir. Além disso, no caso da regressão múltipla, é possível identificar qual variável independente tem maior impacto sobre a variável dependente.

A regressão linear também permite a identificação de outliers, a verificação da homoscedasticidade e se a análise dos resíduos segue uma distribuição normal, além de possibilitar fazer previsões de eventos futuros. Portanto, podemos afirmar que a regressão linear é uma ferramenta essencial na análise de dados, permitindo modelar e entender as relações entre as variáveis

A simplicidade e eficiência computacional da regressão linear a tornam acessível, embora seja importante considerar suas limitações como a sensibilidade a outliers, limitação a linearidade, não dar a certeza de que a correlação implica em causalidade e cuidado ao lidar com variáveis não quantitativas contínuas. Além disso, a regressão linear serve de base para métodos mais avançados e pode ser integrada com outras técnicas de análise de dados para abordar questões complexas.

A constante evolução das ferramentas de análise de dados oferece muitas oportunidades para expandir seu uso em novos contextos. Em resumo, a regressão linear é uma ferramenta indispensável para a exploração e compreensão das relações entre variáveis, sendo essencial para qualquer analista de dados.

REFERÊNCIAS

BATISTA, Ivanildo. **Testes de Correlação:** Breve explicação sobre correlação e testes estatísticos em R e Python. Medium, 2021. Disponível em: https://ivanildo-batista13.medium.com/testes-de-correla%C3%A7%C3%A3o-3cb0a37e0f2. Acesso em: 27 / 07 / 2024.

BOBBITT, Zach. **How to Calculate Confidence Interval for Regression Slope.** Statology, 2022. Disponível em: https://www.statology.org/confidence-interval-for-regression-slope/. Acesso em: 27 / 07 / 2024.

CHEIN, Flávia. **Introdução aos Modelos de Regressão Linear**. Brasília-DF: ENAP, 2019.

FORJAN, James. **Hypothesis Test and Confidence Intervals in Multiple Regression.** ANALYST PREP, 2019. Disponível em: https://analystprep.com/study-notes/frm/part-1/quantitative-analysis/hypothesis-tests-and-confidence-intervals-in-multiple-regression/. Acesso em: 27 / 07 / 2024.

FRANÇA, Alex. **Entenda o que é o Coeficiente de Determinação na Regressão Linear**. Psicometria Online, 2023. Disponível em: https://www.blog.psicometriaonline.com.br/entenda-o-que-e-o-coeficiente-de-determinacao-na-regressao-linear/. Acesso em: 27 / 07 / 2024.

FAVERO, Luiz; BELFIORE, Patrícia. **Manual de Análise de Dados:** Estatística e Análise Multivariada com Excel, SPSS e Stata. Rio de Janeiro: LTC, 2017.

GUIMARAES, Amanda**. Estatista:** Análise de Correlação usando Python e R. Medium, 2021. Disponível em: https://medium.com/omixdata/estat%C3%ADstica-an%C3%A1lise-de-correla%C3%A7%C3%A3o-usando-python-e-r-d68611511b5a#:~:text=O%20coeficiente%20de%20Pearson%2C%20tamb%C3%A9m,valores%20entre%20%2D1%20e%201. Acesso em: 27 / 07 / 2024.

KUMAR**,** Manish. **CarPrice_Assignment**. 2018. Kaggle. Disponível em: https://www.kaggle.com/datasets/hellbuoy/car-price-prediction. Acesso em: 27 / 07 / 2024.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis.** 5th ed. Hoboken: John Wiley & Sons, 2012

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. 6. ed. Rio de Janeiro: LTC, 2016.

OPENCADD. **O que é Box Plot?** OPENCADD, 2022. Disponível em: https://www.opencadd.com.br/box-plot-o-que-e-e-para-que-serve/. Acesso em: 27 / 07 / 2024.

PARDOE, Ian. **Residuals vs Predictor Plot.** Eberly College of Science**.** Disponível em: https://online.stat.psu.edu/stat462/node/247/. Acesso em: 27 / 07 / 2024.

PERES, Fernanda Fiel. **Como Interpretar (e construir) um gráfico Box Plot?** Fernanda Peres Estatística Aplicada À Vida Real, 2022. Disponível em: https://fernandafperes.com.br/blog/interpretacao-boxplot/. Acesso em: 27 / 07 / 2024.

REBEKKA, Topp; GÓMEZ, Guadalupe. **Residual Analysis in Linear Regression Models With an Interval-censored Covariente,** National Library of Medicine 2004. Disponível em: https://pubmed.ncbi.nlm.nih.gov/15490424/. Acesso em: 27 / 06 / 2024

REBEKKA, Topp; GÓMEZ, Guadalupe**. Residual Analysis in Linear Regression Models With an Interval-censored Covariate. National** Library of Medicine, 2004. Disponível em: https://pubmed.ncbi.nlm.nih.gov/15490424/. Acesso em: 27 / 07 / 2024.

TRIOLA, Mario F; **Introdução a Estatística**. 4 ed. Rio de Janeiro: LTC, 2017.

YIN, Robert K**. Estudo de caso:** planejamento e métodos. 2. ed. Porto Alegre: Bookman, 2001.

APÊNDICE A – TABELA COM RESIDUOS PADRONIZADOS DA REGRESSÃO LINEAR SIMPLES

Tabela 14 - Valores para Identificações de Outliers numa Regressão Linear Simples

	x	y	Residuos	Residuos.Padronizados
1	0.0000000	2.2629543	0.8429136	0.18995601
2	0.1010101	0.8757868	-0.7584483	-0.17081667
3	0.2020202	2.7338397	0.8854100	0.19929148
4	0.3030303	2.8784899	0.8158658	0.18353072
5	0.4040404	2.2227222	-0.0540964	-0.01216214
6	0.5050505	0.4701510	-2.0208622	-0.45408307
7	0.6060606	1.2835542	-1.4216535	-0.31926782
8	0.7070707	2.1194210	-0.7999812	-0.17956000
9	0.8080808	2.6103944	-0.5232022	-0.11737442
10	0.9090909	5.2228352	1.8750441	0.42043127
11	1.0101010	3.7837955	0.2218098	0.04971066
12	1.1111111	2.4232130	-1.3529672	-0.30307273
13	1.2121212	2.2765854	-1.7137892	-0.38371907
14	1.3131313	3.3368011	-0.8677681	-0.19420552
15	1.4141414	3.5290677	-0.8896959	-0.19902469
16	1.5151515	3.6187922	-1.0141660	-0.22677090
17	1.6161616	4.4845467	-0.3626060	-0.08104602
18	1.7171717	3.5424223	-1.5189249	-0.33935713
19	1.8181818	5.0720469	-0.2034947	-0.04544679
20	1.9191919	3.6008454	-1.8888907	-0.42168855
21	2.0202020	24.8161362	19.1122055	4.26516754
22	2.1212121	5.6198199	-0.2983053	-0.06654754
23	2.2222222	5.5777808	-0.5545389	-0.12366711
24	2.3232323	6.4506542	0.1041400	0.02321648
25	2.4242424	5.7913781	-0.7693306	-0.17145643
26	2.5252525	6.5541130	-0.2207901	-0.04919125
27	2.6262626	7.3382946	0.3491969	0.07777703
28	2.7272727	5.7635916	-1.4397006	-0.32057616
29	2.8282828	5.3719663	-2.0455204	-0.45535106
30	2.9292929	6.9053120	-0.7263691	-0.16165486
31	3.0303030	6.8248995	-1.0209762	-0.22716488
32	3.1313131	6.7197380	-1.3403322	-0.29815188
33	3.2323232	7.0313361	-1.2429285	-0.27642424
34	3.3333333	7.0171950	-1.4712642	-0.32713786
35	3.4343434	8.5954376	-0.1072161	-0.02383502
36	3.5353535	9.2226188	0.3057706	0.06796296
37	3.6363636	9.2648876	0.1338450	0.02974436
38	3.7373737	8.0452344	-1.3000028	-0.28885395
39	3.8383838	9.9150718	0.3556401	0.07900989
40	3.9393939	8.5994416	-1.1741846	-0.26082493
41	4.0404040	30.8387112	20.8508905	4.63110495
42	4.1414141	9.8435744	-0.3584408	-0.07960311
43	4.2424242	9.0320645	-1.3841452	-0.30736335
44	4.3434343	8.8548254	-1.7755788	-0.39425164
45	4.4444444	8.7223183	-2.1222804	-0.47119938
46	4.5454545	9.0253185	-2.0334747	-0.45145493
47	4.6464646	8.7291472	-2.5438405	-0.56473465
48	4.7474747	11.6514865	0.1643043	0.03647436
49	4.8484848	11.5290168	-0.1723599	-0.03826171
50	4.9494949	10.6716612	-1.2439100	-0.27612894
51	5.0505051	11.3671475	-0.7626182	-0.16928955
52	5.1515152	10.9263276	-1.4176326	-0.31469652
53	5.2525253	13.9464151	1.3882604	0.30818374
54	5.3535354	10.9117316	-1.8606176	-0.41305862
55	5.4545455	11.8542134	-1.1323303	-0.25139043
56	5.5555556	12.3612524	-0.8394858	-0.18638686

57	5.6565657	12.9313746	-0.4835581	-0.10736982
58	5.7575758	12.3425280	-1.2865992	-0.28570228
59	5.8585859	10.4932714	-3.3500503	-0.74398454
60	5.9595960	11.6555775	-2.4019387	-0.53348465
61	6.0606061	33.4799410	19.2082303	4.26677833
62	6.1616162	13.3121868	-1.1737184	-0.26075619
63	6.2626263	12.5846034	-2.1154964	-0.47005242
64	6.3636364	13.6114474	-1.3028468	-0.28953158
65	6.4646465	13.1143242	-2.0141645	-0.44768383
66	6.5656566	14.3735766	-0.9691066	-0.21544043
67	6.6666667	12.9082349	-2.6486428	-0.58892982
68	6.7676768	14.9012947	-0.8697776	-0.19343639
69	6.8686869	14.9857864	-0.9994803	-0.22233066
70	6.9696970	15.0046821	-1.1947791	-0.26583564
71	7.0707071	15.1605705	-1.2530852	-0.27887654
72	7.1717172	15.6007727	-1.0270775	-0.22863660
73	7.2727273	14.8964445	-1.9456003	-0.43322416
74	7.3737374	15.6283060	-1.4279332	-0.31804518
75	7.4747475	16.6136306	-0.6568031	-0.14633336
76	7.5757576	17.2524843	-0.2321440	-0.05173664
77	7.6767677	16.4973068	-1.2015159	-0.26786028
78	7.7777778	16.4378020	-1.4752153	-0.32898615
79	7.8787879	15.8455074	-2.2817043	-0.50901480
80	7.9797980	15.5220097	-2.8193965	-0.62918947
81	8.0808081	36.3645266	17.8089259	3.97578325
82	8.1818182	18.6177195	-0.1520758	-0.03396332
83	8.2828283	18.3377988	-0.6461910	-0.14437154
84	8.3838384	17.5481611	-1.6500231	-0.36879649
85	8.4848485	17.5448867	-1.8674921	-0.41757747
86	8.5858586	17.7527371	-1.8738362	-0.41917653
87	8.6868687	19.3707242	-0.4700435	-0.10519521
88	8.7878788	18.2999795	-1.7549827	-0.39294233
89	8.8888889	20.0337966	-0.2353602	-0.05272208
90	8.9898990	19.6264724	-0.8568789	-0.19203846
91	9.0909091	20.4811305	-0.2164153	-0.04852566
92	9.1919192	18.5105763	-2.4011640	-0.53867361
93	9.2929293	19.5942295	-1.5317052	-0.34379932
94	9.3939394	18.9070071	-2.4331222	-0.54641841
95	9.4949495	20.5861580	-0.9681657	-0.21754461
96	9.5959596	40.3116368	18.5431186	4.16892676
97	9.6969697	20.1117655	-1.8709472	-0.42087350
98	9.7979798	22.0519480	-0.1449593	-0.03262799
99	9.8989899	21.0269994	-1.3841024	-0.31172560
100	10.0000000	21.9965439	-0.6287523	-0.14169338

Fonte: Dados hipotéticos do autor (2024)

APÊNDICE B - TABELA COM RESIDUOS PADRONIZADOS DA REGRESSÃO LINEAR MÚLTIPLA

Tabela 15 - Valores para Identificações de Outliers numa Regressão Linear Multipla

	X1	X2	y	Resíduo	Resíduo Padronizado
2	0,37454	0,95071	15,1972	-0,51444	-0,044727225
3	0,73199	0,59866	15,4215	0,41943	0,036466914
4	0,15602	0,15599	11,5411	0,22777	0,019803001
5	0,05808	0,86618	13,7908	-0,65328	-0,056798916
6	0,60112	0,70807	17,2095	2,04085	0,177439262
7	0,02058	0,96991	15,3851	0,54704	0,047561583
8	0,83244	0,21234	12,3677	-1,06001	-0,092161298
9	0,18182	0,1834	12,1191	0,60419	0,052531035
10	0,30424	0,52476	12,5618	-0,92008	-0,079995429
11	0,43195	0,29123	13,5391	0,82617	0,071830753
12	0,61185	0,13949	13,6916	1,21404	0,105553172
13	0,29214	0,36636	11,8876	-0,80352	-0,069861398
14	0,45607	0,78518	16,2575	1,11551	0,096987259
15	0,19967	0,51423	13,583	0,4366	0,037959711
16	0,59241	0,04645	12,8316	0,85214	0,074088851
17	0,60754	0,17052	14,572	1,95775	0,170214403
18	0,06505	0,94889	14,6942	-0,16459	-0,014310115
19	0,96563	0,8084	16,1851	-0,45764	-0,039788956
20	0,30461	0,09767	10,5127	-0,92687	-0,080585809
21	0,68423	0,44015	13,4377	-0,67582	-0,058758876
22	0,12204	0,49518	12,7649	-0,07856	-0,006830461
23	0,03439	0,90932	14,9909	0,40506	0,035217546
24	0,25878	0,66252	14,3656	0,34859	0,030307896
25	0,31171	0,52007	14,3627	0,88281	0,076755124
26	0,54671	0,18485	12,5774	0,06046	0,005256313
27	0,96958	0,77513	18,238	1,74354	0,151590246
28	0,9395	0,89483	17,028	0,04294	0,003733748
29	0,5979	0,92187	19,1232	2,94045	0,255654371
30	0,08849	0,19598	11,8711	0,55057	0,047868572
31	0,04523	0,32533	10,9052	-0,91617	-0,079655832
32	0,38868	0,27135	11,4519	-1,04789	-0,091107611
33	0,82874	0,35675	14,7525	0,64389	0,055982088
34	0,28093	0,5427	13,3328	-0,17135	-0,014897852

35	0,14092	0,8022	15,1478	0,78387	0,068152746
36	0,07455	0,98689	15,6313	0,56482	0,049107476
37	0,77224	0,19872	13,2375	0,03911	0,003400807
38	0,00552	0,81546	13,2471	-0,81099	-0,070510602
39	0,70686	0,72901	14,2508	-1,30643	-0,113586129
40	0,77127	0,07404	12,2375	-0,36171	-0,031448508
41	0,35847	0,11587	12,5111	0,83765	0,07282867
42	0,8631	0,6233	15,9199	0,44233	0,038458259
43	0,3309	0,06356	10,0647	-1,28328	-0,11157353
44	0,31098	0,32518	12,732	0,1866	0,016223943
45	0,72961	0,63756	15,7619	0,58023	0,050447275
46	0,88721	0,47221	14,1389	-0,68161	-0,059261842
47	0,11959	0,71324	14,0787	0,1986	0,017267402
48	0,76079	0,56128	15,1469	0,24518	0,021316795
49	0,77097	0,4938	13,6389	-0,96777	-0,084141773
50	0,52273	0,42754	14,0637	0,45101	0,039213037
51	0,02542	0,10789	11,1765	0,4495	0,039081076
52	0,03143	0,63641	14,3594	1,08733	0,094536477
53	0,31436	0,50857	14,5397	1,10767	0,096305506
54	0,90757	0,24929	12,5915	-1,21792	-0,105891044
55	0,41038	0,75555	14,0711	-0,80453	-0,069949164
56	0,2288	0,07698	11,5863	0,45254	0,039346053
57	0,28975	0,16122	12,1891	0,48608	0,042261338
58	0,9297	0,80812	17,3447	0,80129	0,069667026
59	0,6334	0,87146	20,1102	4,07183	0,354020826
60	0,80367	0,18657	13,9148	0,68879	0,059885837
61	0,89256	0,53934	16,51	1,35374	0,117699658
62	0,80744	0,89609	17,8568	1,22586	0,106581506
63	0,318	0,11005	12,1557	0,62036	0,053936157
64	0,22794	0,42711	12,5041	-0,30252	-0,026302669
65	0,81801	0,86073	17,5167	1,02609	0,089212616
66	0,00695	0,51075	11,8018	-0,80231	-0,069756215
67	0,41741	0,22211	12,126	-0,21659	-0,018831472
68	0,11987	0,33762	11,5623	-0,52138	-0,045330751

69	0,94291	0,3232	14,5266	0,26719	0,023230613
70	0,51879	0,70302	17,3861	2,46619	0,214420498
71	0,36363	0,97178	14,0825	-1,70011	-0,147814505
72	0,96245	0,25178	14,8325	0,86151	0,074903078
73	0,49725	0,30088	11,3834	-1,55374	-0,13508838
74	0,28484	0,03689	10,567	-0,52778	-0,045887488
75	0,60956	0,50268	15,431	1,22205	0,106250138
76	0,05148	0,27865	11,6119	-0,00309	-0,00026905
77	0,90827	0,23956	12,8449	-0,9199	-0,079980175
78	0,14489	0,48945	12,1666	-0,71176	-0,061883662
79	0,98565	0,24206	14,8468	0,85908	0,074691648
80	0,67214	0,76162	15,0941	-0,52439	-0,04559228
81	0,23764	0,72822	14,5705	0,29675	0,025800971
82	0,36778	0,63231	14,3105	0,14068	0,012231559
83	0,63353	0,53577	13,9279	-0,50483	-0,043891881
84	0,09029	0,8353	16,5913	2,20714	0,191897696
85	0,32078	0,18652	12,5289	0,62012	0,053915939
86	0,04078	0,59089	11,0516	-2,02813	-0,176333778
87	0,67756	0,01659	12,3021	0,23332	0,0202859
88	0,51209	0,2265	12,007	-0,6148	-0,053452826
89	0,64517	0,17437	13,6598	0,92448	0,080377729
90	0,69094	0,38674	13,214	-0,66222	-0,057576116
91	0,93673	0,13752	13,3831	0,02887	0,00251013
92	0,34107	0,11347	12,0956	0,48098	0,041818132
93	0,92469	0,87734	18,0265	1,16555	0,101337528
94	0,25794	0,65998	12,8734	-1,12917	-0,098174722
95	0,81722	0,5552	14,8932	-0,13345	-0,011602788
96	0,52965	0,24185	12,3233	-0,41986	-0,036504042
97	0,0931	0,89722	14,1121	-0,57602	-0,050081423
98	0,90042	0,6331	17,6322	2,00598	0,174408159
99	0,33903	0,34921	13,1681	0,43123	0,037492883
100	0,72596	0,89711	15,4025	-1,01102	-0,087902332
101	0,88709	0,77988	17,4785	1,18639	0,103149704
102	0,95487	0,7379	56,5541	40,278	3,50192728
103	0,55435	0,61172	54,7217	40,1416	3,4900669
104	0,4196	0,24773	52,4975	40,0263	3,480050502
105	0,35597	0,75785	54,8571	40,119	3,488102007
106	0,01439	0,11607	50,6235	39,8875	3,46797606
107	0,046	0,04073	50,3417	39,8798	3,467313465
108	0,85546	0,70366	56,0847	40,2435	3,498928822
109	0,47417	0,09783	51,9117	40,0089	3,478535181
110	0,49162	0,47347	53,8422	40,0946	3,485988523
111	0,1732	0,43385	52,6889	39,9993	3,477696729

Fonte: Dados hipotéticos do autor (2024)